UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DISEÑO Y CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN DE DEPRESIÓN EN ADULTOS MAYORES EN CHILE

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

MAURICIO ANDRÉS GONZÁLEZ GÓMEZ

PROFESOR GUÍA:
JUAN VELASQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
ASTRID CONTRERAS FUENTES
VICTOR HERNANDEZ MARTÍNEZ

SANTIAGO DE CHILE
2019

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: MAURICIO ANDRÉS GONZÁLEZ GÓMEZ
FECHA: 2019
PROF. GUÍA: JUAN VELASQUEZ SILVA

DISEÑO Y CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN DE DEPRESIÓN EN ADULTOS MAYORES EN CHILE

El presente trabajo de título tiene por objetivo construir un modelo predictivo de riesgo de depresión en adulto mayores, con ello elaborar un indicador que muestre cuan propenso es un adulto mayor a sufrir este tipo de enfermedades. En en el mundo, la depresión afecta al 4,4 % de la población mundial lo que significa que 300 millones de personas sufren de este trastorno, lo que implica un costo social alrededor 1 billón de dolares del PIB mundial debido a la depresión.

En Chile, se estima que una 1 de cada 5 personas presenta síntomas depresivos y la depresión afecta al 5 % de la población, lo que coloca a Chile por sobre el promedio mundial y uno de los más afectados por este trastorno a nivel latinoamericano. La depresión se define como aquel estado de animo triste que persiste pese a haberse disipado la causa externa, en algunos casos inclusive no tienen causa externa precipitante. El paciente con depresión pierde el interés en sus actividades, trabajo, familia e inclusive las ganas de vivir. En el caso de los adultos mayores la depresión puede traer consigo múltiples consecuencias como la falta de apetito, problemas de insomnio, etc. Incluso en casos severos de depresión pueden derivar en otras patologías como la demencia o en el suicidio. Actualmente en Chile, los adultos mayores lideran la tasa de suicidios en la población chilena llegando a los 17,7 suicidios por cada 100.000 habitantes.

Para la construcción de este modelo predictivo de riesgo de depresión, se entiende como un problema de datos complejos en el que se pretende utilizar herramientas de Data Science y los algoritmos de Machine Learning, cuya principal ventaja radica en encontrar patrones para descubrir asociaciones presentes en registros y profundizar mas allá de la evidencia de un caso puntual. Mediante un algoritmo de clasificación y los datos de la Encuesta Nacional de Salud junto con la metodología "Knowledge Discovery in Databases"(KDD) y el apoyo de software como SPSS y Python. Se toma el desafío de realizar un clasificador y un indicador de riesgo de depresión para los adultos mayores.

Los resultados del experimento después de la investigación, selección y balance de datos arrojó que si existe la posibilidad de formular un predictor de riesgo de depresión para adultos mayores en base a las preguntas contenidas en la Encuesta Nacional de Salud, el modelo con mejores resultados se realizó mediante un estudio de variables de riesgo de depresión junto con el método de clasificación de Random Forest que alcanzo un Recall de 0.57 y un AUC aproximado del 0.81, lo que lo convierte en un modelo con una alta capacidad predictiva.

La principal conclusión es la capacidad de formular un predictor de riesgo depresión dentro de la población chilena con un gran rendimiento. Sin embargo, no se obtuvieron mejoras significativas al acotar el experimento solo a adultos mayores, ya que los resultados al probar el modelo con muestras acotadas y no acotadas tuvieron resultados similares.

A mi familia que los amo profundamente.

Agradecimientos

Mis más profundos agradecimientos a toda mi familia, en especial a mis padres que con su amor y su crianza me dieron el impulso y la fortaleza para completar mis estudios, a mis hermanas y hermano que entre risas y juegos me hacían gozar de los momentos y a mi abuelo que con su apoyo incondicional me dio ánimo para desarrollar mi memoria. También estoy profundamente agradecido de mi polola, que con su cariño tuvo la fuerza para sacarme una sonrisa en los momentos más difíciles y llenarme de felicidad. Doy las gracias a todos mis amigos que hicieron de mi experiencia universitaria inolvidable, llena de aventuras y risas.

Agradecer al WIC y a su equipo que me brindaron su apoyo y experiencia para completar este desafió, además de brindar una buena compañía todos los días.

Por último, dar las gracias Dios por ayudarme en los momentos más oscuros y darme la sabiduría necesaria para llegar a este momento.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Antecedentes generales

Esta memoria se desarrolla en el centro de investigación Web Intelligence Centre (WIC) de la Universidad de Chile, como un proyecto complementario al proyecto Koreisha que tiene por objetivo caracterizar y cuantificar los trastornos cognitivos en adultos mayores, con el fin de elaborar un instrumento de medición (índice) que permita determinar el nivel de dependencia y las necesidades de salud de las personas con trastornos cognitivos. Además de crear un prototipo en adultos mayores con enfermedades cognitivas en la ciudad de Aysén. Este proyecto está a cargo del profesor Juan Velásquez Silva y un equipo multidisciplinario del WIC, médicos especialistas y ejecutivos del Fondef.

1.1.1. Web Intelligence Centre

El Web Intelligence Centre es un centro de investigación que nace hace aproximadamente 10 años bajo la iniciativa del profesor Juan Velázquez en el marco del departamento de ingeniería Civil Industrial de la Universidad de Chile. El WIC se compromete con 3 principales objetivos que son:

- Publicar investigación en las principales revistas y conferencias relacionadas la Web Intelligence.
- Proveer a servicios profesionales de excelencia y rápida para sus clientes.
- Dictar cursos de orientación práctica hacer de las tecnologías de la información y su aplicación en los negocios.

WIC es un centro experto en Web intelligence dedicada a la investigación científica e industrial. La misión, visión y objetivos del WIC se detallan en su página web.

Mision: ¡Creando soluciones ingenieriles inteligentes!

"Creamos tecnologías de información usando data science para apoyar la toma de decisiones en organizaciones que se interesan en innovar. Creemos que esta disciplina puede generar un gran impacto en la sociedad y eso nos apasiona."

Visión: ¡Impactar en el mundo!

"Queremos ser un actor relevante en el área del data science y sus aplicaciones en la salud. Para ello contaremos al 2022 con al menos 3 proyectos transferidos en distintos centros de salud."

Los principales proyectos del WIC son:

- **AKORI:** Plataforma enfocada en el análisis de las estructuras de las páginas web, entregando información acerca de las formas en que ciertos perfiles de usuarios, definidos por el cliente, visualizan dichas páginas.
- **Delirium:** Es financiado por Fondef en su concurso para iniciativas tecnológicas innovadoras dirigidas a adultos mayores, población que proporcionalmente día a día crece más en el país, este proyecto busca mediante un software, aliviar la confusión mental de adultos mayores en estado crítico.
- **DOCODE:** Plataforma de detección de plagio que permite automatizar el proceso de análisis de documentos digitales, permitiendo encontrar coincidencias que podrían ser potencialmente consideradas como plagio, con el propósito de entregar información objetiva para el tomador de decisiones.
- **Koreisha:** Tiene por objetivo realizar predicciones sobre el potencial de cobertura de salud que pudiese tener una persona según sus características médicas y psicosociales.
- **OpinionZoom:** Analiza minuto a minuto las opiniones que los usuarios chilenos de Twitter. Esto permite obtener información valiosa para cualquier organización que tenga por bien conocer a sus clientes y escuchar sus inquietudes para ofrecerles los mejores productos o un mejor servicio.
- **Sonama:** Plataforma que permite realizar seguimiento o monitorización en línea de la opinión sobre el consumo de marihuana y alcohol, y dependencia de este último, en los usuarios chilenos de Twitter.

1.1.2. Proyecto Koreisha

El proyecto Koreisha tiene el objetivo el caracterizar y cuantificar los trastornos cognitivos en adultos mayores, con el fin de elaborar un instrumento de medición (índice) que permita determinar el nivel de dependencia y las necesidades de salud de las personas con trastornos cognitivos y que apoye la asignación de intervenciones de salud y socio-sanitaria de acuerdo a cada consultante para así aportar a mejorar la calidad de vida y la de su entorno, además de optimizar la gestión de recursos[80].

Actualmente Chile es un país con un envejecimiento creciente, donde la población sobre los 60 años ha aumentado su representación de 15 % en 2009 a representar un 19,3 % de la población nacional para el 2017. Existen problemas de salud de alta prevalencia en adultos

mayores que son factores de desigualdad en la población debido a los altos costos monetarios y sociales que involucran. Por lo que es relevante contar con métodos eficientes que permitan monitorear como responde el sistema publico de salud ante las necesidades del adulto mayor. Actualmente para monitorear la efectividad y comportamiento de la salud publica en Chile es a través de la Encuesta Nacional de Salud pero esta metodología es lenta y costosa y no permite un monitoreo dinámico del sistema de salud.

El proyecto Koreisha tiene por principal objetivo realizar predicciones sobre el potencial de cobertura de salud que pudiera tener una persona según algunas de sus características médicas y psicosociales, entendiendo por cobertura hasta que punto del proceso de entrega de un servicio de salud este satisface a sus pacientes (Ver figura 1.1).

Dentro de este objetivo el proyecto primero pretende diseñar una encuesta especializada que logre capturar datos que permitan caracterizar la cobertura de salud y características médicas y psicosociales, segundo tiene la misión de aplicar dicha encuesta en una población seleccionada de adultos mayores en la región de Aysén. Tercero pretende diseñar un índice que permita estimar la cobertura de salud en función de un problema de salud en particular, utilizando herramientas de Data Mining y Machine Learning. Por último, realizar un prototipo experimental en el que se presenten los resultados obtenidos a un usuario final.

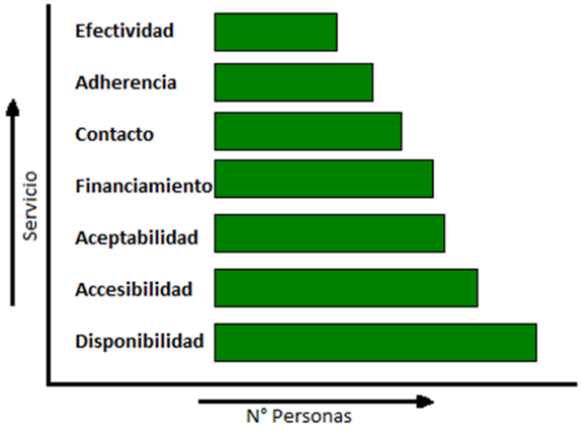


Figura 1.1: Etapas de cobertura del sistema de salud.
Fuente: Comisión Nacional de Investigación Científica y Tecnológica. [80]

Este proyecto se relaciona con la memoria en cuestión debido a que cuenta con un modulo de preguntas apreciativas relativas a la depresión. Por lo tanto, se planteo el desafió de elaborar un indicador de riesgo de depresión en base a preguntas de una encuesta con un fin investigativo.

1.2. Justificación

En el mundo existen más de 300 millones de personas que sufren depresión [57], según el informe de la OMS el 4,4 % de la población mundial sufre depresión y este número ha

aumentado un 18 % en los últimos 10 años, y es aún mayor en países con ingresos medios o bajos. El informe de la OMS también resalta que la población adulta mayor es la más propensa a sentir los efectos de esta enfermedad. En Latinoamérica, Chile es el 3ª país con más casos de depresión en América Latina con un 5 % de su población afectada con este trastorno [54].

Un ejemplo de las consecuencias según el comunicado de la OMS en conjunto con el Banco Mundial, la depresión y la ansiedad le cuestan al mundo 1 billón de dólares al año [17]. También según estudio realizados por "The Lancet Psychiatry" por cada US$ 1 invertidos en tratamientos de la depresión y la ansiedad rinde US$ 4 en mejoras de la salud y la capacidad de trabajo [77]. La importancia de los tratamientos de la depresión tiene sentido desde el punto de vista de la salud y bienestar, pero también desde una perspectiva económica.

En Chile se entiende la depresión como aquel estado de ánimo triste que persiste pese a haberse disipado la causa externa o una expresión desproporcionada de está. Incluso, Casos severos de depresión no tienen causa externa precipitante. El paciente con depresión pierde el interés incluso de vivir, sintiéndose incapaz de realizar sus actividades [89]. Junto con la tristeza aparecen una serie de otros síntomas como son las alteraciones del sueño, del apetito, concentración y síntomas corporales que interfieren gravemente con su calidad de vida. La respuesta al tratamiento de la depresión es variable según cada individuo, lo que aún no es comprendido en su totalidad.

La depresión en adultos mayores es más frecuente en mujeres, viudas, pacientes con enfermedades crónicas, portadores de algún tipo de trastornos del sueño como insomnio, en aquellos que han vivido experiencias estresantes en su vida y aquellos que han vivido aislamiento social [89]. Según la última Encuesta Nacional de Salud del 2016-2017 la depresión afecta al 6,2 % de la muestra y en su mayoría son mujeres, esta encuesta se basa en aproximadamente de 5.000 personas y en base al instrumento DSM-IV para la detección de depresión [23],los resultados detalla el siguiente gráfico 1.2:

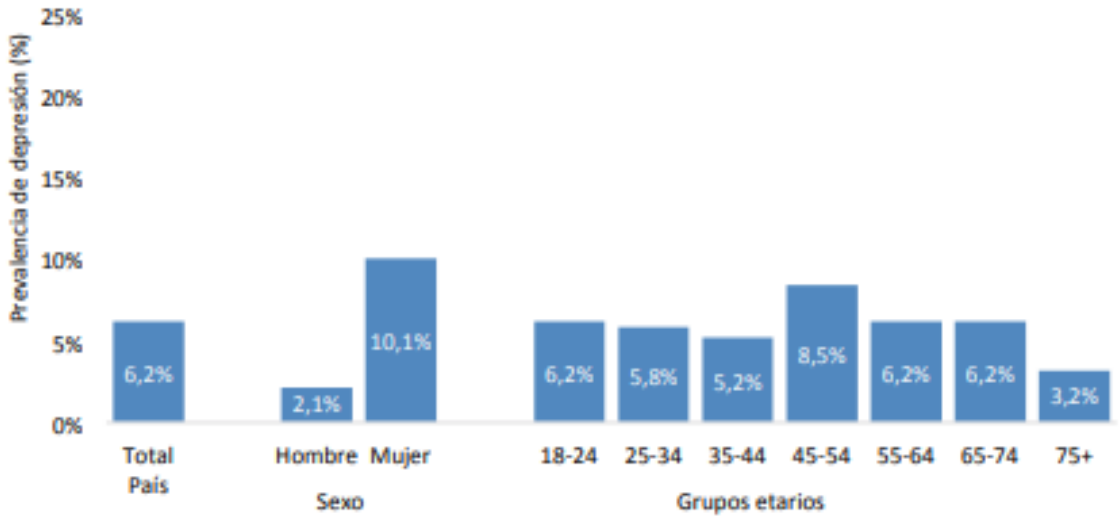


Figura 1.2: "Prevalencia depresión de los últimos 12 meses" total según sexo y edad.
Fuente: Criterio DSM-IV Encuesta Nacional de Salud 2016-17.[23]

No existen informes que calculen los costos de la depresión en Chile dado a que no solo involucra un costo económico, sino que también uno costo social, pero se sabe que los trastornos mentales y del comportamiento ocupan el primer lugar en el gasto por licencias curativas, concentrando 19,3 % de los gastos totales por subsidios de cargo de las isapres, de todas estas licencias el 53,8 % son por episodios depresivos. De las causas principales en la emisión de licencias médicas [24], donde en promedio cada licencia alcanza un costo de los $264.702.

En Chile se calcula que existen 850.000 casos de depresión, en donde los adultos mayores lideran la tasa de suicidios llegando al 14,2 por cada 100.000 habitantes en los mayores de 65 y en 17,7 por cada 100.000 habitantes en los mayores de 70 [58]. Esta tasa es aún mayor que en los jóvenes entre los 15 a 30 años que alcanza el 13,3 suicidios cada 100.000 habitante, dejando a Chile como una de las tasas más altas de Latinoamérica en lo que compete a suicidios de adultos mayores alcanzando los 935 entre 2010-2015.

Una de las principales consecuencias de los cuadros depresivos en adultos mayores es que pueden evolucionar hacia deterioro cognitivo e incluso a demencia, a pocos años de comenzar la patología depresiva, cuyos síntomas cognitivos remiten casi completamente después del tratamiento antidepresivo, un 40 % desarrolla demencia dentro de los siguientes 3 años[66]. Además, la depresión es un factor de riesgo para enfermedades como el Alzheimer y enfermedades cardiacas.

Los principales factores de riesgo para el brote de la depresión entre los adultos mayores están:

- Las enfermedades crónicas
- La pérdida de seres queridos
- El sentimiento de soledad e inutilidad
- El hospitalizarse
- La inactividad
- Antecedentes de depresión
- Bajo nivel socio económico
- Abuso y/o suspensión brusca de alcohol, tranquilizantes o drogas.
- Las enfermedades crónicas

La primera encuesta de calidad de vida del adulto mayor e impacto del pilar solidario de la subsecretaria de previsión social, ejecutada por el centro de encuestas y estudios longitudinales de la Universidad Católica, esta encuesta consideró variables tales como calidad de vida, condición de la vivienda, entorno, equilibrio en el uso del tiempo, bienestar subjetivo, redes y cohesión social, salud, seguridad, educación y competencias, además de ingresos y gastos. Esta encuesta reveló que entre las enfermedades que más sufren los adultos mayores se encuentra la depresión, ocupando el séptimo lugar [75] como muestra el gráfico 2.2.

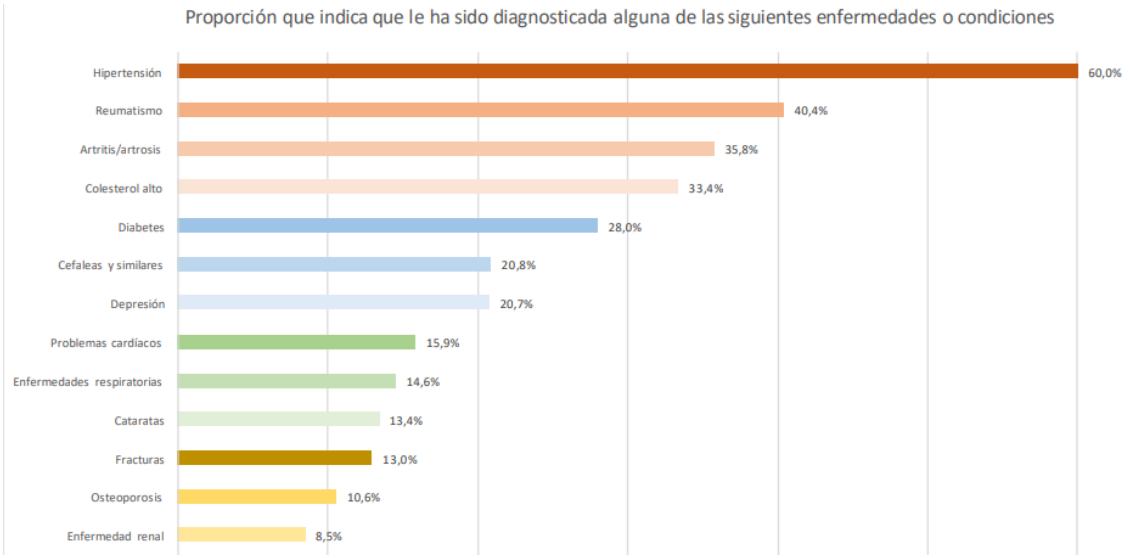


Figura 1.3: Listado de enfermedades, ordenado de mayor a menor prevalencia.
Fuente: Principales resultados de la encuesta de calidad de vida del adulto mayor e impacto del pilar solidario [75].

Además, la encuesta reveló el hecho que el 47,6 % de los adultos mayores es menos feliz que cuando era más joven, un 54,8 % se siente más inútil que cuando era más joven y un 19,3 % de estos adultos mayores se siente aislado socialmente [75] esta ultima es una de las principales condiciones de riesgo de depresión.

Otro dato relevante respecto a los adultos mayores hospitalizados, según un estudio realizado en el Hospital Militar de Santiago del General Luis Felipe Brieba Aran reveló que el 28 % [89] de sus adultos mayores que pasaban por su centro hospitalario presentaba depresión, esto es sumamente preocupante dado que la depresión es un factor que incide en la recuperación de los adultos mayores.

En Chile, la depresión unipolar es la segunda más importante condición de carga de enfermedad (índice que combina los daños letales prematuros que ocasionan los problemas de salud con sus consecuencias en términos de discapacidad [14]), la tercera para el grupo de las mujeres y la quinta en los hombres [71].Además, en Chile a pesar de las altas tasas de depresión detectadas en consultantes del nivel primario de atención, estudios han demostrado moderados niveles de detección de los médicos generales, según estudio americanos menos del 36,7 % de los médicos generales son capaces de detectar depresión en la primera consulta [11].

1.2.1. Estado del arte

En el mundo actualmente se esta manejando diferentes programas de salud mental liderados por el de la OMS llamado Plan de Acción sobre Salud Mental 2013-2020 [16], cuyos principales objetivos son reforzar un liderazgo y una gobernanza eficaces en el ámbito de salud mental, proporcionar un servicio asistencial social y de salud mental, colocar en práctica

estrategias de promoción y prevención en el campo de salud mental y por ultimo fortalecer los sistemas de información e investigaciones sobre la salud mental. En este mismo ámbito Chile por su parte ha comenzado la planificación de un nuevo Plan Nacional de Salud Mental,el propósito del plan es recoger y realizar los principales desafíos del sector y dar respuesta a las necesidades de salud mental de la población que habita el territorio nacional, este plan de salud mental y psiquiatría propone objetivos estratégicos y metas a cumplir entre el 2017 y el 2025 en 7 lineas de acción complementarias y sinergia entre si en áreas de regulación y derechos humanos, provisión de servicios de salud mental, financiación, gestión de la calidad , sistemas de información e investigación, recursos humanos y formación y participación intersectorial.

Por lo tanto, la oportunidad que se presenta en este contexto es el generar algún mecanismo que ayude a detectar si los adultos mayores dado a su contexto y entorno pudiese ser susceptible a caer en un estado depresivo que a su vez desencadene en algún trastorno mayor, la rápida detección de estas patologías podría ayudar de manera significativa a la calidad de vida de los adultos mayores.

Actualmente en el área de salud mental se ha avanzando en la utilización de técnicas de Machine Learning y Data Mining en particular en el caso de la de la depresión se han desarrollado distintos algoritmos para detectarla, como por ejemplo es el caso del desarrollo de un clasificador basado en las señales encefalogramas(EEG) para discriminar a los pacientes con depresión. Para discriminar los dos grupos se utilizo K-Nearest Neighborths, el análisis discriminante lineal y la regresión logística. La mayor precisión alcanzada de clasificación fue del 83.3 % [35].

Otro ejemplo de desarrollo de un clasificador de depresión fue el que utilizó imágenes de rostros neutros para determinar depresión , este proyecto basado en estudios que hablaban de anomalías en el procesamiento de caras emocionales en depresión mayor, desarrolló un clasificador utilizando Machine Learning, el resultado de este experimento no fue del todo concluyente aunque si se encontró patrones de activación cerebral en caras neutrales [53].

En Chile no han existido grandes avances en el desarrollo de nuevos tecnologías relativas a la depresión, uno de los pocos registros es el caso del desarrollo de un predictor de depresión mayor en Europa denominado predictD, el cual es el primer algoritmo de riesgo de depresión , este modelo en particular fue validado con casos de población de Chile. El resultado de este instrumento alcanzo 72 % de efectividad, las principales variables utilizadas por este predictor son demográficas como el sexo, edad, nivel educativo, trabajo, salud física y mental [39].

1.3. Propuesta de solución

Dado al desarrollo de las herramientas de Data Mining, Machine Learning y con los datos que se manejan de las Encuestas Nacionales de Salud que tiene por objetivo saber que enfermedades y que tratamientos están recibiendo hombre y mujeres mayores de 15 años en Chile, se tiene la oportunidad de predecir ciertos patrones que pudiesen determinar cuan propenso es un adulto mayor chileno a tener depresión, y con ello formular una alerta temprana con la

que un tomador de decisiones pueda focalizar sus esfuerzos y recursos en aquellos casos que estén más propensos a sufrir la enfermedad.

Esto significa realizar un modelo de predicción de riesgo de depresión en pacientes adultos mayores, usando algoritmos de clasificación de Machine Learning. El objetivo es dar con un sistema de alarma temprana a la falta de otros mecanismos que ayuden a detectar el riesgo de sufrir la enfermedad, además de entender y descubrir cuales son las variables que explican este trastorno y si las las nuevas técnicas de Machine Learning son capaces de determinar de forma efectiva la prevalencia de depresión en adultos mayores chilenos.

1.4. Hipótesis de investigación

La hipótesis central en el desarrollo de este proyecto es que a partir de los datos disponibles en la Encuesta Nacional de Salud es posible utilizar un algoritmos de Machine Learning para predecir el riesgo de depresión en adultos mayores chilenos.

1.5. Objetivos

1.5.1. Objetivo general

El objetivo general de esta memoria es crear un modelo de Machine Learning capaz de predecir si un adulto mayor es propenso o tiene riesgo de sufrir depresión dadas las variables de la Encuesta Nacional de Salud y encontrar patrones que pudieran servir de alerta temprana a un tomador de decisiones y determinar en que pacientes debe focalizar sus esfuerzos. Al final se pretende que el modelo arroje un indicador que entregue cuan propenso es un adulto mayor en caer en trastornos depresivos que puedan afectar su calidad de vida.

1.5.2. Objetivo específicos

En aras de cumplir el objetivo general de la memoria se detallan los hitos importantes que se deben cumplir en el desarrollo del proyecto:

1. **Estudio de las variables relevantes para detectar depresión y selección de variables para adultos mayores:** En esta etapa se espera realizar una investigación exhaustiva sobre el estado del arte de la cobertura de la depresión en Chile y detectar las principales variables que afectan en el deterioro de la salud mental en adultos mayores y por último realizar la selección y construcción de variables para el desarrollo del proyecto. Además en etapa se desarrollo la investigación de diferentes herramientas de Machine Learning.
2. **Testing de múltiples modelos de Machine Learning para predecir riesgo de depresión en adultos mayores**: En esta etapa se construirá un modelo capaz de

predecir riesgo de depresión en adultos mayores. Además, se pretende validar el modelo realizado mediante distintas métricas, calculado la efectividad del modelo.

3. **Desarrollo del indicador de riesgo de tener depresión en adultos mayores:** En base a la capacidad predictiva del modelo anterior se pretende realizar un modelo para estimar el riesgo de depresión en adultos mayores.
4. **Análisis de impacto:** En esta última etapa se espera analizar el impacto del proyecto y obtener conclusiones del trabajo realizado. Además de entregar en los aspectos que puede mejorar el modelo.

1.6. Metodología

La metodología consiste en la revisión de técnicas relacionadas con el desarrollo y validación de modelos de predicción. Para el desarrollo de la memoria se utilizará la metodología KDD. Por lo tanto, se nombran a continuación las distintas etapas de la metodología:

1. **Selección de datos:** En esta etapa se seleccionan y extrae la data disponible en registros de las respuestas de la Encuesta Nacional de Salud del año 2009-2010 y de ser necesario se complementará con la base de datos de la Encuesta Nacional de Salud del año 2016-2017.
2. **Pre-Procesamiento y transformación:** Una vez completada la etapa anterior se procede al pre-procesamiento y a la transformación de los datos, lo que incluye la transformación de variables, reducción de dimensionalidad. Esta etapa también incluye el tratamiento de valores faltantes. Posteriormente se espera generar la base de datos respectiva con la información de ya consolidada incluyendo las variables predictoras y variables objetivos.
3. **Data Mining:** En esta etapa se selecciona el algoritmo de clasificación y la estrategia para el entrenamiento con el fin de generar un modelo predictivo. Posteriormente se construye el modelo predictivo para la depresión. Finalmente se entrenará el algoritmo de clasificación que permita predecir de mejor manera depresión en adultos mayores.
4. **Análisis de resultados:** En esta etapa se quiere medir el desempeño del calificador y para ello se usará la validación cruzada, comparando cada una de las metodologías y proponer la que mejor se ajuste.

1.7. Resultados esperados

Los resultados esperados para el desarrollo de esta memoria se pueden enumerar en los siguientes:

1. **Modelo predictivo de riesgo de depresión:** En una primera instancia se quiere lograr construir un modelo capaz de predecir riesgo de depresión en adultos mayores, y con ello poder evaluar cuales son los factores de mayor riesgo que pueden hacer que

un adulto mayor caiga en depresión, para ello se utilizaran las bases de datos de la Encuesta Nacional de Salud.

2. **Elaboración de un indicador de riesgo de depresión:** Con lo anterior se pretende desarrollar un indicador capaz de ser una alerta temprana que comunique lo propenso de un adulto mayor a caer en depresión, el propósito de este indicador es hacer una medida tangible del trabajo realizado.
3. **Medición de impacto:** Se pretende evaluar el beneficio económico y social que traería consigo el poder predecir riesgo de depresión en adultos mayores.

1.8. Alcances

El objetivo de esta memoria es lograr predecir con variables extraídas de fuentes públicas, como son la Encuesta Nacional de Salud 2009-2010 y la Encuesta Nacional de Salud 2016-2017, la posibilidad de encontrar ciertos patrones que predigan si una persona puede sufrir de depresión y para ello se utilizaran datos de adultos mayores (Personas mayores de 60 años), las variables están estrictamente restringidas a las que aparecen en ambas encuestas entre las que existen variables como nivel educacional, edad, sexo, enfermedades previas, diagnósticos médicos, diagnósticos sobre la depresión, etc. Se asume que los datos de estás encuestas están previamente validados por el Ministerio de Salud y el Departamento de Epidemiología. Otro punto a considerar es que el resultado de esta memoria no pretende determinar si un adulto mayor completa su tratamiento, solo se enfocara en la detección del riesgo de episodios depresivos.

Además, se asume este proyecto de memoria dentro de los márgenes de Koreisha que está dirigido hacia adultos mayores con problemas cognitivos o de carácter mental, Por lo que no supone una implementación en algún centro asistencial. Por último aclarar, que esta implementación de Data Mining no pretende reemplazar ninguna de las herramientas existentes para detectar depresión y que este diagnóstico debe ser siempre entregado por un médico especialista.

Esta investigación empleó información de las Encuestas de Salud para vigilancia epidemiológica de la Subsecretaría de Salud Pública. El autor agradece al Ministerio de Salud de Chile, haberle permitido disponer de la base de datos. Todos los resultados obtenidos del estudio o investigación son de responsabilidad del autor y en nada comprometen a dicha institución.

Capítulo 2

Marco Teórico

2.1. Depresión

La depresión es uno de los desordenes mentales más comunes que se presenta con un estado anímico deprimido, con perdida de interés o placer, disminución de energía, sentimientos de culpa, baja autoestima, sueño perturbado, apetito y pobre concentración. Frecuentemente la depresión viene con síntomas de ansiedad. Estos problemas pueden volverse crónicos y conduce a perdidas en las capacidades de cuidar de las responsabilidades de todos los días, en el peor de los casos la depresión puede llevar al suicidio. Casi 1 millón de vidas humanas se pierden debido al suicidio todo los años, que se traduce en 3000 muertes todos los días y por cada suicidio que se llega completar 20 personas han intentado acabar con su vida [48].

En el caso de los adultos mayores las consecuencias de la depresión puede desencadenar en múltiples problemas en su calidad de vida. Los síndromes depresivos pueden asociarse en los adultos mayores con restricción de las competencias sociales debido al aumento de la morbilidad, la soledad, el aislamiento social, a consecuencias de la jubilación, los conflictos interpersonales y la pérdida de la pareja o parientes cercanos. Los síndromes depresivos endógenos en los ancianos se combinan frecuentemente con quejas somáticas difusas [51].

La depresión geriátrica se puede clasificar en 2 tipos, la primera de inicio temprano y la segunda de inicio tardío, cuyo primer episodio depresivo ocurre después de los 60 años. En la depresión de inicio tardío, existe menor frecuencia de antecedentes psiquiátricos, alteraciones de la personalidad y mayor presencia de síntomas psicóticos (alucinaciones o ideas delirantes). Además, existe mayor asociación con factores de riesgo cardiovascular y otras enfermedades crónicas como el cáncer, dolor crónico, insuficiencia renal crónica, enfermedades del sistema nervioso entre otras. Otro elemento particular de la depresión geriátrica es la presencia de síntomas cognitivos, aunque se debe destacar que no todos los pacientes adultos mayores con depresión presentan alteraciones cognitivas [66].

2.2. Variables que inciden en la depresión

Las variables clínicas y socio-demográficas asociadas con la depresión y la vejez van desde el nivel educativo, la edad, trastornos neuropsiquiátricos discapacidad física, dificultades sociales como la soledad,etc. Las principales variables de riesgo que inciden en la depresión se detallan a continuación.

2.2.1. Socio-demográficos

Dentro de la gran gama de variables socio-demográficas el sexo, edad, nivel educativo, urbanización, situación laboral, estado marital,etc. Son las que están mayormente vinculadas con la depresión, a continuación se detallan las más relevantes:

1. **Sexo:** La depresión es una de las causas principales de discapacidad tanto para hombres como para mujeres, pero la depresión es 50 % más recurrente en las mujeres que en los hombres[67]. De hecho, la depresión es la principal causa de enfermedad para las mujeres tanto en países de ingresos altos y países de ingresos bajos. La investigación en los países en desarrollo sugiere que la depresión materna puede ser un factor de riesgo para el pobre crecimiento en niños pequeños.
2. **Nivel Educativo:** Múltiples estudios revelan una correlación entre el nivel de estudios y la depresión. Según se detalla en el informe un bajo nivel de estudios puede aumentar la prevalencia de depresión hasta un 31 % [29]. También hacer hincapié en los ancianos con un nivel de estudios bajos, con alguna dependencia para las tareas diarias y con aislamiento social que presentan inclusive un riesgo mayor.
3. **Nivel Socio-económico:** Estudios revelan que dentro de adultos mayores que viven en pobreza extrema la mitad tiene algún tipo de depresión, siendo recurrentes en las mujeres.[27]
4. **Estado Marital:** Un estudio realizado sobre la correlación de variables socio-demográficas con la depresión y desordenes mentales descubrió una relación entre la depresión y las personas quienes previamente estuvieron casadas o se encuentran en estado de viudes.[4]

2.2.2. Actividad física

La actividad física según múltiples estudios revela consistentemente una alta correlación con una buena salud mental [44]. También se ha revelado que existe una correlación entre el sedentarismo con un aumento en el riesgo de tener depresión.[81]

Se han presentado estudios que demuestran que la actividad física y el ejercicio también se pueden utilizar en el tratamiento de la depresión y los trastornos de ansiedad. No se conocen todos los mecanismos responsables de las mejoras relacionadas con el ejercicio en la depresión y los trastornos de ansiedad, y es más probable que sea una interacción compleja de los mecanismos psicológicos y neurobiológicos, que median o moderan estos efectos [44]. En especial los trabajos de fuerza de alta intensidad en adultos mayores a corto, medio y

largo plazo son los que tienen mejores resultados en la superación de este trastorno, lo que se debe tener en consideración para el tratamiento de los adultos mayores con depresión [61].

2.2.3. Problemas de vista y auditivos

Las personas que tienen discapacidades típicamente pueden tener problemas prácticos o sociales en sus vidas cotidianas que también puede afectar negativamente su salud mental,en particular se a encontrado cierta correlación entre los problemas severos de visión y audición, los cuales son comunes entre los adultos mayores y que su agudeza va en incremento proporcional con el paso de los años [8]. El mismo estudio revela que los mayores de 61 años, los problemas a la vista se ven estrechamente vinculados con la depresión. Por otro lado, los problemas de audición no fueron concluyentes con respecto a la depresión pero si tuvo correlación con problemas derivados como la ansiedad.

Otro estudio afirma una relación existente entre problemas auditivos y la depresión con mayor celeridad, en especifico en la depresión de inicio tardío, llegando a afirmar incluso que se detecta depresión en los adultos mayores en un 30 % más si estos sufren algún problema auditivo [38].

2.2.4. Dolor crónico

Los síntomas físicos son comunes en la depresión mayor y puede llevar a un sufrimiento crónico y complicaciones en el tratamiento de la depresión. Los dolores asociados a la depresión incluyen dolores en las articulaciones, dolor en las extremidades, dolor de espalda, problemas gastrointestinales, fatiga, cambios en la actividad psicométríca y cambios en el apetito [83].

Un estudio realizado por la Organización Mundial de la Salud analizó la relación entre síntomas somáticos (se presenta cuando una persona siente una ansiedad extrema a causa de síntomas físicos como dolor o fatiga), utilizando 1146 pacientes en 14 países y que cumplían con el criterio de tener depresión, el 69 % reportó en su visita tener síntomas somáticos asociados con la depresión. Desafortunadamente, la depresión a menudo no se diagnostica en estos pacientes, ya que los síntomas físicos asociados con la depresión pueden interpretarse como síntomas de una enfermedad somática [74].

Aunque los criterios de diagnóstico de depresión enfatizan en lo emocional y síntomas vegetativos, la depresión mayor también se asocia con síntomas físicos dolorosos como dolor de cabeza, dolor de espalda, dolor de estómago, dolor en las articulaciones y dolor muscular.Todo esto se debe principalmente a que la depresión y el dolor comparten una vía neuroquímica común, ya que ambos están influenciados por la serotonina y la norepinefrina. La depresión y los síntomas físicos dolorosos asociados deben tratarse juntos en para lograr la remisión [83].

2.2.5. Trastornos alimenticios

La relación entre los trastornos alimenticios y la depresión según estudios, aproximadamente del 40 % al 80 % de los pacientes con trastornos alimenticios cumplen con los criterios de un historial de vida de depresión mayor, la mayoría de episodios se desarrollan en forma simultanea o paralela trastornos alimenticios [79]. La obesidad ha sido asociada con un aumento del riesgo de depresión mayor y trastorno de pánico particularmente entre las mujeres. También la asociación entre obesidad y la depresión es más fuerte entre los que tienen casos de obesidad severa.

La depresión también se puede asociar con una serie de comportamientos poco saludables como son fumar,obesidad, sedentarismo y consumo excesivo de alcohol, también existe la correlación de depresión igualmente asociado con estos comportamientos poco saludable. Otro punto es que mientras se presenta la depresión actual las personas que fueron previamente diagnosticados con depresión tienen un mayor riesgo de comportamientos adversos para la salud como la obesidad en comparación con los que nunca han sido diagnosticados con depresión. Por ultimo la depresión y la ansiedad tuvieron un efecto aditivo sobre el tabaquismo [78].

2.2.6. Trastornos del sueños

Los trastornos del sueño son muy frecuentes entre las personas que sufren depresión como muestran los estudios realizados por The Journal of Clinical Psychiatry demuestra que el 90 % de los pacientes con depresión tienen o tuvieron problemas relacionados con el sueño[84]. Aproximadamente el 40 % de los pacientes se quejan de problemas para iniciar el sueño, mantener el sueño y/o despertarse temprano por la mañana, incluso muchos pacientes reportan los tres.

La evidencia de que el sueño influye fuertemente tanto en el desarrollo de la depresión, afectando la frecuencia, severidad y duración del episodios depresivos, sugiere que los síntomas relacionados con el sueño pueden ser importantes, y factores de riesgo modificables para prevenir la depresión o lograr y mantener la remisión de la depresión. Los pacientes con trastornos del estado de ánimo que tienen trastornos del sueño deben ser evaluados cuidadosamente [28]. La evidencia reciente sugiere que las intervenciones para el insomnio, que incluyen tratamientos psicoterapéuticos y de comportamiento y farmacoterapia, pueden ser útiles para la depresión.

2.2.7. Diabetes

La principal conclusión de la revisión de estudios es que la diabetes duplica las probabilidades de depresión. Tanto los clínicos como los epidemiólogos pueden esperar que los individuos con diabetes tengan el doble de probabilidades de estar deprimidos que los individuos no diabéticos en entornos similares (es decir, los individuos seleccionados por procedimientos similares, del mismo sexo, y evaluados con métodos de evaluación de depresión comparables).

En contraste, la estimación de prevalencia debe ajustarse para moderadores como el sexo.[3]

Por último un estudio realizado con diabetes y biomarcadores vinculados a la depresión[20] descubrió que la distribución de glóbulos rojos, glucosa sérica y la bilirrubina total están directamente relacionadas con la presencia de depresión.

2.2.8. Problemas cardiovasculares

Los infartos junto con los problemas cardiovascular son la primera causa de defunción en Chile alcanzando el 27,53 % de las defunciones del 2014. Los análisis apoyan la hipótesis de que la depresión es un factor de riesgo para el desarrollo de la enfermedades coronarias. El riesgo de contraer enfermedades cardiovasculares fue 60 % mayor en los pacientes deprimidos. El análisis de diversos estudios mostró evidencia de que la depresión tiene efecto desfavorable sobre la mortalidad en pacientes con cardiopatía coronaria. La depresión,como fumar o los lípidos, parece ser un factor de riesgo altamente relevante en pacientes con enfermedades cardiovasculares. Las estrategias para los tratamientos de enfermedades cardiovasculares estandarizadas se han establecido para estos factores de riesgo como es la depresión no desencadenen en una peor condición [5].

2.2.9. Problemas respiratorios

Existe una relacion entre las distintas enfermedades respiratorias y la depresión, como demuestra un estudio realizado con pacientes con enfermedades pulmonares obstructivas crónicas que estimo una prevalencia entre un 10 % a 42 % de depresión entre estos enfermos crónicos, las diferencias se dieron en la variabilidad en los instrumentos de diagnostico y en las puntuaciones obtenidas, pero a medida que la gravedad de los problemas respiratorios aumentaba ,aumentaba considerablemente el grado de depresión, de hecho se estima que los pacientes graves tiene 2.5 veces más probabilidades de desarrollar depresión [9].Una vez que la enfermedad psiquiátrica está bajo control, es probable que la necesidad de estos pacientes de una evaluación psiquiátrica de la psicoterapia sea mayor que la de los pacientes que no han experimentado un problema similar [21].

2.2.10. Enfermedades crónicas, memoria y dependencia

La prevalencia de depresión en adultos mayores que sufren de enfermedades crónicas según se demuestra en una encuesta realizada sobre 1.672 personas con depresión en Washington, Estados Unidos [13]. Entre las enfermedades crónicas que mas se repiten en el caso de la depresión son:

1. Cáncer
2. Enfermedades Coronaria
3. Diabetes

4. Epilepsia
5. Esclerosis múltiple
6. Derrame cerebral
7. Alzheimer
8. VIH
9. Parkinson
10. lupus
11. Artritis reumatoide

Según un estudio sobre adultos mayores en España en el que participó más de 5.830 personas usando instrumentos especializados como el EURO-D (depresión) y CASP-12 (calidad de vida) observo que la depresión se encontraba en mayor frecuencia en participantes que padecían de Alzheimer (76,4 %), trastornos emocionales (73,9 %), Parkinson (57,4 %) y fracturas de cadera (55,4 %)[65], también cabe destacar que los adultos mayores que sufren enfermedades cognitivas tienen una alta probabilidad de sufrir depresión, en un estudio en la Habana con mayores de edad descubrió una asociación directa entre las enfermedades cognitivas y la depresión [86].

2.2.11. Problemas de determinante social y psicológico

La percepción del adulto mayor acerca de su estado de salud y calidad de vida son influidos severamente por su salud mental y capacidad funcional. Por tanto, la percepción de salud es un constructo asociado a otras variables psicológicas como autoestima y satisfacción con la vida y ha probado tener asociaciones significativas con otros indicadores más objetivos, como son el número de enfermedades crónicas que los adultos mayores padecen, el periodo de tiempo que han vivido con una enfermedad, la agudización de problemas crónicos, etc [88]. Son muchos los estudios epidemiológicos que se han realizado sobre el suicidio en adultos mayores y que permiten establecer la existencia de esa serie de factores que representa un mayor riesgo de suicidio como son ser varón, mayor de 60 años, con historia de un intento previo, con antecedentes familiares de suicidio o trastornos del estado del ánimo, con pérdida de pareja reciente, el transcurso de fechas señaladas con gran carga afectiva, sufrir aislamiento social, dificultades económicas, desempleo y las más preponderantes como sentimientos de incomprensión reales o imaginarios hacia el entorno y humillaciones sociales recientes [68].

2.3. Composite Internacional Diagnostic Interview-CIDI

La Composite Internacional Diagnostic Interview (CIDI) es un encuesta escrita por la Organización Mundial de la Salud (OMS) y la Administradora Americana de alcohol,abuso de drogas y salud mental (ADAMHA) en el contexto de un proyecto en conjunto en 1979 [69]. La CIDI fue diseñada para medir y diagnosticar desordenes mentales, severidad, carga de estos desordenes, evalúa el uso del servicio, medicamentos en el tratamiento y las barreras

para el tratamiento. Dentro de los trastornos que detecta se encuentra la depresión, ansiedad, ataques de pánico, etc.

Esta encuesta se utiliza como una herramienta epidemiológica que permite realizar diagnósticos de múltiples patologías, administrarlo por personal no necesariamente clínico y puntuarlo por computador.

2.3.1. Módulo depresión

DSM-IV

El DSM-IV es una clasificación categórica que divide los trastornos mentales en diversos tipos basándose en series de criterios con rasgos definitorios. Este enfoque categórico es siempre más adecuado cuando los miembros de una clase diagnostica son homogéneos, cuando existen límites claros entre las diversas clases y cuando las diferentes clases son mutuamente excluyentes.

El DSM-IV no se asume que cada categoría de trastorno mental sea una entidad separada, con limites que la diferencian de otros trastornos mentales o no mentales.Tampoco hay certeza de que los individuos que padezcan el mismo trastorno sean iguales, el clínico que maneje esta encuesta debe considerar que es muy probable que las personas con el mismo diagnostico sean heterogéneas [62].

A	Cinco (o más) de los síntomas siguientes durante el mismo período de 2 semanas y representan un cambio respecto del desempeño previo; por lo menos uno de los síntomas es (1) estado de ánimo depresivo o (2) pérdida de interés o placer. (1) Estado de ánimo depresivo la mayor parte del día, casi todos los días, indicado por el relato subjetivo o por observación de otros. (2) Marcada disminución del interés o del placer en todas, o casi todas, las actividades durante la mayor parte del día, casi todos los días. (3) Pérdida significativa de peso sin estar a dieta o aumento significativo , o disminución o aumento del apetito casi todos los días. (4) Insomnio o hipersomnia casi todos los días. (5) Agitación o retraso psicomotores casi todos los días. (6) Fatiga o pérdida de energía casi todos los días. (7) Sentimientos de desvalorización o de culpa excesiva o inapropiada (que pueden ser delirantes) casi todos los días (no simplemente autorreproches o culpa por estar enfermo). (8) Menor capacidad de pensar o concentrarse, o indecisión casi todos los días (indicada por el relato subjetivo o por observación de otros). (9) Pensamientos recurrentes de muerte (no sólo temor de morir), ideación suicida recurrente sin plan específico o un intento de suicidio o un plan de suicidio específico
B	Los síntomas no cumplen los criterios de un episodio mixto
C	Los síntomas provocan malestar clínicamente significativo o deterioro del funcionamiento social, laboral o en otras esferas importantes.
D	Los síntomas no obedecen a los efectos fisiológicos directos de una sustancia (por ejemplo, una droga de abuso,una medicación), ni a una enfermedad médica general (por ejemplo hipotiroidismo).
E	Los síntomas no son mejor explicados por duelo, es decir que tras la pérdida de un ser querido, los síntomas persisten por más de 2 meses o se caracterizan por visible deterioro funcional, preocupación mórbida con desvalorización, ideación suicida, síntomas psicóticos o retraso psicomotor.

Tabla 2.1: Criterios para detectar depresión en DSM-IV[62]

CIE-10

En la práctica, la CIE se ha convertido en una clasificación diagnóstica estándar internacional para todos los propósitos epidemiológicos generales y muchos otros de administración de salud. Esto incluye el análisis de la situación general de salud de grupos de población y el seguimiento de la incidencia y prevalencia de enfermedades y otros problemas de salud en relación con otras variables, tales como las características y circunstancias de los individuos afectados[18]

A	El episodio depresivo debe durar al menos dos semanas.
B	El episodio no es atribuible a abuso de sustancias psicoactivas o a trastorno mental orgánico.
C	Síndrome Somático: comúnmente se considera que los síntomas "somáticos tienen un significado clínico especial y en otras clasificaciones se les denomina melancólicos o endógenomorfos - Pérdida importante del interés o capacidad de disfrutar de actividades que normalmente eran placenteras - Ausencia de reacciones emocionales ante acontecimientos que habitualmente provocan una respuesta - Despertarse por la mañana 2 o más horas antes de la hora habitual - Empeoramiento matutino del humor depresivo - Presencia de enlentecimiento motor o agitación - Pérdida marcada del apetito - Pérdida de peso de al menos 5 % en el último mes - Notable disminución del interés sexual

Tabla 2.2: Criterios para detectar depresión en CIE-10[18]

2.4. Data mining

En el desarrollo de esta memoria se enfrenta a la situación de tener una base de datos con *complex data*, la complejidad se debe a la multidimensionalidad, la inclusión de datos que podrían considerarse categóricos o continuos y la ocurrencia de *missing data* [82]

El Data Mining es una nueva disciplina que nace de la confluencia de varias otras disciplinas, impulsado por el crecimiento de las bases de datos. la motivación básica del Data Mining es que detrás de las grandes bases de datos se tiene información que es valiosa para el dueño de la base de datos [31].

El termino Descubrimiento de Conocimiento en Bases de datos o KDD por sus siglas en inglés (Knowledge Discovery in Data Base), hace referencia a un proceso de búsqueda de conocimiento en los datos el que es utilizado por muchas herramientas de Data Mining, entonces se puede definir como un proceso no trivial de identificación de patrones válidos, novedosos, potenciales y comprensibles en grandes volúmenes de datos. Entender que Data Mining y KDD son procesos distintos, KDD se refiere al proceso de descubrir información, en cambio Data Mining hace referencia al reconocimiento de patrones en una base de datos a través de distintos algoritmos.

El KDD se caracteriza por tener etapas claves y que consisten en secuencias iterativas de los siguientes etapas:

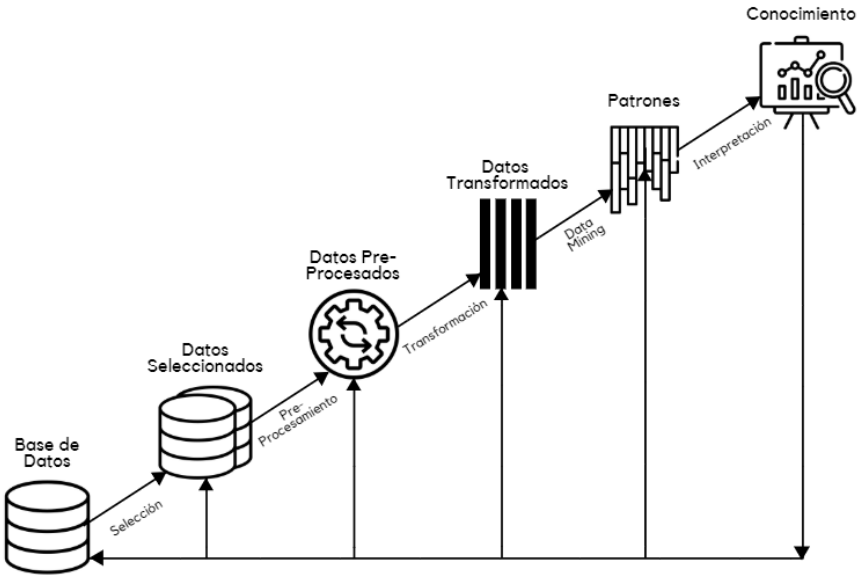


Figura 2.1: Diagrama flujo KDD
Fuente: Elaboración Propia

1. **Selección de Datos:** Como primera instancia para poder desarrollar los algoritmos de Data Mining se requiere construir un set de datos con diversas fuentes de información, seleccionando las variables con diferentes metodologías, y con esto poder revelar los patrones que ocultan posteriormente en el proceso de Data Minig, por lo tanto, estas variables se consolidan en una base de datos. Entre las fuentes más comunes de

información se encuentran los data marts y data warehouses.

2. **Pre-Procesamiento:** Para realizar una efectiva captura de los patrones ocultos en los datos se requiere garantizar la calidad de estos, para ello se pasa por una etapa de limpieza de datos, que incluye entre sus etapas la eliminación de las mediciones con valores sin información o aquellos valores que se encuentran fuera de rango.
3. **Transformación:** Es la reducción de atributos y proyección de la data, esto es, encontrar y/o crear atributos útiles que representen y caractericen la data dependiendo del objetivo del proceso, con el objetivo de transformar los datos en bruto, de manera que se facilite la extracción de información contenida en estos.
4. **Data Mining:** Este paso consiste en el análisis de datos y en la selección de algoritmos donde se realiza el aprendizaje supervisado o no supervisado con el fin de hallar patrones sobre los datos seleccionados. Este proceso incluye el desarrollo del modelo, los parámetros más apropiados según las variables y el objetivo que se quiere lograr y seleccionar un método de data mining en particular con el objetivo general del proceso de KDD
5. **Interpretación:** Esta última etapa consiste en la interpretación de los patrones encontrados, en esta etapa dependiendo de los resultados y el análisis se puede volver a cualquiera de los pasos anteriores haciendo de KDD un proceso iterativo, en esta etapa también se realizan evaluaciones sobre la calidad de las predicciones o clasificaciones que arroja el modelo, usualmente etapa incluye distintas visualizaciones de los patrones y modelos. Finalmente, después de las múltiples ideaciones se pretende concebir nuevo conocimiento [59].

2.4.1. Machine Learning, aprendizaje supervisado, no supervisado y de refuerzo

Entendiendo que aprender es obtener conocimiento de estudios, experiencias o que puede ser enseñado, también puede ser descrito como obtener conciencia por medio de la información o la observación para almacenarse en la memoria [91]. Machine Learning es un área que involucra conocimientos de la estadística y de la inteligencia artificial. En este proceso se espera que un sistema pueda aprender de la información [45] y de esta manera predecir lo que sucederá en el futuro, se requiere para esto que en los datos existan patrones que pudiesen dar luces y expliquen a que clase pertenece. Dentro de las aplicaciones de Machine Learning son el reconcomiendo de caracteres, filtros anti-spam, diagnósticos médicos, detección de fraudes, predicción del clima, segmentos de clientes. El uso de esta metodologías en el ámbito de la salud y en especial Machine Learning con depresión es muy variado donde destacan ejemplos de :

- Predictor de depresión según comportamiento de uso de aparatos móviles.
- Reconocimiento de escalas de depresión en registros de audio, visuales y análisis de texto.
- Fusionando la minería de datos, el aprendizaje automático para detectar biomarcadores asociado con la depresión.
- Clasificación de pacientes con depresión mediante técnicas de aprendizaje automático

a partir de la señal de EEG (electroencefalografía).

Dentro de Machine Learning existen múltiples formas de clasificar y predecir elementos, a continuación se dsecriben las mas relevantes:

- **Aprendizaje Supervisado:** El aprendizaje supervisado que construye un clasificador a partir de los datos de entrenamientos los que conllevan una etiqueta de la clase a la que pertenecen, el clasificador resultante es utilizado luego para asignar etiquetas de clases a los datos de prueba, estos datos de prueba están previamente etiquetados con el fin de evaluar la efectividad y precisión del clasificador.
- **Aprendizaje no Supervisado:** Son aquellos que no requieren el uso de datos de entrenamiento, es decir, el modelo aprende por su cuenta. Generalmente es usado para técnicas de clustering, que es aquel que busca asociar o agrupar datos en grupos de objetos similares.
- **Aprendizaje de Refuerzo:** Es un sistema de machine learning que se ajusta el peso según la dirección que se extiende el gradiente de refuerzo esperando un refuerzo inmediato en las tareas y mejorar su rendimiento [90].

2.5. Herramientas tecnológicas

A continuación se presentan los principales software utilizados en el desarrollo de esta memoria:

- **Lenguaje de Programación:** Para programar el software utilizado esta memoria se eligió el software Python 3.2.1 debido a que como lenguaje de programación orientada a objetos facilita de manera significativa los esfuerzos ya que permite el manejo de grandes volúmenes de datos y de un repositorio de funciones, por otro lado se utilizara la librería de desarrollo llamada Scikit-Learn[2], la que contempla diferentes herramientas de data mining y machine learning que facilitaran el desarrollo del proyecto
- **SPSS:** Para el desarrollo de esta memoria se eligio el software estadistico SPSS[50] para realizar el análisis exploratorio de variables y la selección de variables. Se utilizo este software en particular por lo intuitivo y fácil que resulta utilizarlo.

2.6. Missing Data

Dentro de los problemas que se enfrenta al desarrollo de técnicas de Data Mining, es la perdida de datos en las muestras, este tema es relevante dado que la aplicación de procedimientos inapropiados de imputación de datos puede introducir a sesgos y reducir el poder explicativo de los métodos estadísticos, incluso invalidando los hallazgos y conclusiones. Existen 3 tipos de Missing Data que se nombrar a continuación[6]:

- **Missing Completely at Random(MCAR):** Este error se basa en que cualquier ausencia de datos es independiente tanto de la misma variable como de las demás

variables del set de datos, ocurriendo de manera completamente aleatoria, por lo tanto, las observaciones con valores perdidos son una muestra aleatoria de las observaciones y síguela misma distribución de los datos [43].

- **Missing at Random(MAR):** Este caso de datos perdidos es cuando los sujetos con datos incompletos son diferentes significativamente de los que presentan datos completos de alguna variable y el patrón de ausencia de datos puede ser deducido a partir de variables de los registros observados que no muestran ausencia de datos [43].
- **Missing not at Random(MNAR)**: La perdida de datos es MNAR cuando la probabilidad de que una variable del set de datos contenga datos perdidos, depende del valor de dicha variable, es decir que el hecho de que existan observaciones perdidas contiene información acerca del valor de la variable [25].

2.7. Manejo de Missing Data

Los métodos utilizados para dar solución a este problema de falta de datos se utilizan los siguientes métodos.

- **List Wise Deletion:** este enfoque es de los más usados. El cual asumiendo un patrón de perdida MCAR, se eliminan todas las medidas que tuviesen datos faltantes, realizando el análisis con los datos restantes. Esto puede inducir a sesgos en los resultados si la perdida de los datos no son del tipo MCAR [42].
- **Imputación de Datos:** Esta técnica consiste en rellenar los datos faltantes lo que permite analizar la data por completo. Si se imputan datos con patrones MCAR puede incrementar el poder predictivo, pero no cambiar las estimaciones puntuales, en el caso de ser MNAR, para obtener resultados no sesgados se requeriría la creación de un modelo que tome en cuenta los datos faltantes. Comúnmente se utiliza para imputar la media o la mediana de las variables o una estimación vía regresión de otras variables. Existen otros métodos de imputación a través de estimadores de máxima verosimilitud, donde se calculan los estimadores usando solo los datos completos.
- **Indicadores de Ausencia:** Los datos anteriores se basan en el hecho de los datos perdidos sigan un patrón MCAR O MAR y no proveen información intrínseca sobre la perdida de la variable. Se propone una forma de modelar explícitamente la ausencia de la data mediante un indicador de ausencia construido desde una variable dicotómica que representa si la variable fue observada o no.

2.8. Trasformación de variables

Esta es la cuarta etapa del proceso KDD que se enfoca en la transformación de datos de una forma a otra para que los datos y los algoritmos de minería se puedan implementar fácilmente. Para ello se utilizan diferentes métodos de reducción y transformación de datos [73].

El principal beneficio es la estabilidad con respecto a la escala, los valores extremos para las variables continuas pueden sesgar las predicciones como lo hace en muchos modelos, especialmente los modelos lineales. Por lo tanto, esto vuelve robusto al modelo ante patrones de distribución inusuales [49]. Una de las transformaciones más comunes es la normalización que facilita la coincidencia de instancias y la integración, los valores de los atributos deben convertirse en un formato consistente y uniforme. La formula para estandarizar una variables se expone a continuación:

$$Norm = \frac{(x - x_{min})}{x_{max} - x_{min}}$$

2.9. Selección de atributos

La importancia de la selección de atributos tanto para mejorar el desempeño de los clasificadores, como para un mejor entendimiento del modelo final, dado a que si no se realizan existe un gran riesgo de sobreajuste de los datos denominado "overfitting" lo que puede desencadenar un modelo con baja capacidad predictiva, a continuación, se presentan los distintos métodos de selección:

- **Métodos de Filtro:** Estos evalúan la relevancia de los atributos observados solo con las propiedades intrínsecas de la data. Lo más común es calcular un puntaje de relevancia para cada atributo en relación a la variable dependiente, removiendo los atributos menos relevantes según un criterio de exclusión, de esto nace un sub set que funciona como conjunto de entrada de los algoritmos de clasificación. La ventaja de estos procedimientos es que son fácilmente escalables para data de alta dimensionalidad, ya que son computacionalmente simples y rápidos, además son independientes del sistema de clasificación que se utiliza (aunque en ciertos casos puede ser una desventaja). Una desventaja es que la mayoría corresponde a análisis univariados, lo cual implica que cada variable es considerada de manera separada, ignorando posibles dependencias entre variables. Dentro de los métodos de filtro se destacan:
  - **Estadístico Chi**: Este método mide que tan independiente es la distribución de cada atributo respecto de las etiquetas de las observaciones [72].
  - **Ganancia de información:** Que usa la entropía para decidir qué tan relevante es un atributo y genera más información para el clasificador.
  - **Fisher Score**: Estima la relevancia de cada atributo independiente del resto, calculando de manera explicita la diferencia absoluta entre las medias del valor del atributo en ambas clases y normaliza según la suma de la varianza intraclases para poder comparar los indicadores obtenidos para cada atributo y ver los atributos que difieren más con la clase [34].
  - **Kolmogorov-Smirnov**: Este test toma una muestra y compara la función de distribución acumulada observada de una variable con una distribución teórica determinada, que puede ser la normal, la uniforme, la de Poisson o la exponencial. La Z de Kolmogorov-Smirnov se calcula a partir de la diferencia mayor (en valor absoluto) entre las funciones de distribución acumuladas teórica y observada [41].

- **Anova**: Es un método de modelado lineal para evaluar la relación entre campos. Para los controladores clave, ANOVA prueba si el valor objetivo de media varía entre combinaciones de categorías de dos entradas. Si la variación es significativa, existe un efecto de interacción [30].

- **Correlación de Pearson**: Las correlaciones miden cómo están relacionadas las variables. Antes de calcular un coeficiente de correlación, se inspeccionan los datos para detectar valores atípicos (que pueden generar resultados equívocos) y evidencias de una relación lineal. El coeficiente de correlación de Pearson es una medida de asociación lineal. Dos variables pueden estar perfectamente relacionadas, pero si la relación no es lineal, el coeficiente de correlación de Pearson no será un estadístico adecuado para medir su asociación [7].

$$\mathbb{P}(\mathrm{i}) = \frac{\sum_{\mathrm{i}}(x_{\mathrm{i}} - \overline{x})(y_{\mathrm{i}} - \overline{y})}{\sqrt{\sum_{\mathrm{i}}(x_{\mathrm{i}} - \overline{x})^2 \sum_{\mathrm{i}}(y_{\mathrm{i}} - \overline{y})^2}}$$

  Los valores que puede tomar la correlación de Pearson fluctúan entre -1 y 1, y el nivel de correlación entre dos variables se evalúa considerando el valor absoluto de este, donde valores más altos corresponden a correlaciones más altas.

- **Métodos Wrapper:** Es un método multivariado que utiliza una función objetivo de un modelo de clasificación para jerarquizar los atributos de acuerdo a su relevancia o importancia para el modelo. Estos a su vez se dividen dos sub-categorías:

  - **Selección hacia delante (Forward Selección):** Este modelo comienza con un set de variables vacío e iterativamente se van agregando nuevas variables hasta que se encuentra un subsert óptimo de atributos.

  - **Selección hacia atrás (Backward elimination):** Este comienza con la inclusión de todos los atributos que son eliminados iterativamente hasta que se encuentra el número óptimo de atributos [10].

- **Métodos Embedded:** Para encontrar el sub set óptimo de atributos se lleva a cabo en la construcción del clasificador, al igual que los métodos Wrapper, estos métodos son específicos para cada algoritmo teniendo la ventaja que incluyen la interacciones con el modelo de clasificación. Estos métodos también pueden ser vistos como un problema de optimización, esto se puede hacer introduciendo la selección de atributos en los parámetros del modelo de manera directa [10].

- **Análisis Componentes Principales (PCA):** Es una poderosa técnica de reducción de datos, diseñada para reducir la complejidad de los datos. PCA busca combinaciones lineales de los campos de entrada que realizan el mejor trabajo a la hora de capturar la varianza en todo el conjunto de campos, en el que los componentes son ortogonales (no correlacionados) entre ellos. El objetivo es encontrar un número pequeño de campos derivados (los componentes principales) que resuman de forma eficaz la información del conjunto original de campos de entrada [36].

2.10. Balance de Base de Datos

En la fase de aprendizaje y la subsiguiente predicción de los algoritmos de Machine Learning pueden verse seriamente afectados por el problema del conjunto de datos desbalanceados. El problema de desequilibrio corresponde a la diferencia del número de muestras de diferentes clases, se considera que una base de datos se encuentra desbalanceada si el porcentaje de la clase minoritaria es un 15 % de la clase mayoritaria. para manejar este problema existen distintas metodologías que se pueden segmentar de 2 tipos.

2.10.1. Over-Sampling

Las tecnicas de Over-Sampling son aquellas que buscan sobre representar la clase minoritaria, de tal manera que la diferencia entre las etiquetas de la clase minoritaria y mayoritaria sea menor o sean equitativas.

SMOTE

Es un enfoque de muestreo en el que se sobre-representa la muestra minoritaria de la base de datos mediante la creación de ejemplos s̈intéticosën lugar de sobre representar con reemplazo. Se toma una muestra de la clase minoritaria e introduce ejemplos sintéticos a lo largo de los segmentos de línea que unen a todos los vecinos más cercanos de la clase minoritaria K. La principal desventaja de esta metodología es la introducción de cierto sesgo debido a que algunas de estas nuevas muestras sintéticas pueden estar sobre puestas sobremuestras de la clase mayoritaria.

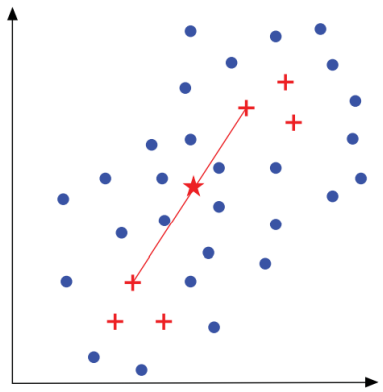


Figura 2.2: Diagrama ejemplo SMOTE.
Fuente: Local neighbourhood extension of SMOTE for mining imbalanced data[46]

La principal desventaja para este método es que se requiere que las variables fuesen continuas, pero existe una versión para variables discretas que se detalla a continuación

SMOTE-NC

Si bien el enfoque de SMOTE actualmente no maneja conjuntos de datos con todas las características discretas, se generalizó para manejar conjuntos de datos mixtos de características continuas y discretas.Esta versión asume que pueden existir variables nominales dentro de conjunto de datos y aproxima los resultados a una de las posibles opciones.[52]

2.10.2. Under-Sampling

Cluster Centroids

Metodo de Under-Sampling que remplaza el grupo mayoritario por los Cluster Centroids de un algoritmo K-Means. Este algoritmo mantiene N muestras de la clase mayoritaria ajustados al algoritmo K-Means con N Cluster Centroids de la clase mayoritaria y usando coordenadas de las n centroides como las nuevas muestras mayoritarias [55].

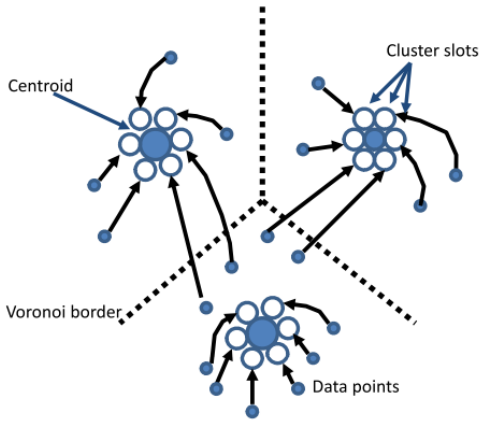


Figura 2.3: Diagrama ejemplo Cluster Centroids.
Fuente: Balanced K-Means for Clustering[47]

Nearest Neighbours

Aplica un algoritmo de vecinos más próximos para "editar" el conjunto de datos eliminando muestras que no concuerdan "lo suficiente" con su vecindario. Para cada muestra de la clase que no se muestreará, los vecinos más cercanos se computarán y, si no se cumple el criterio de selección, se eliminará de la muestra.[55]

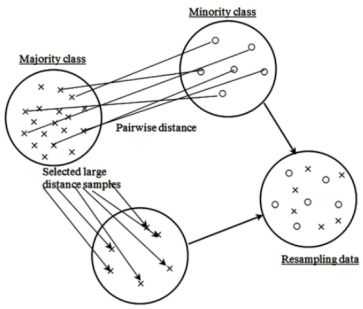


Figura 2.4: Diagrama ejemplo Nearest Neighbours.
Fuente: Balancing class for performance of classification with a clinical dataset[64]

2.11. Métodos de clasificación supervisado

Dentro del Data Mining existen múltiples métodos de clasificación basados en distintos campos de investigación, estos se pueden comparar entre sí en base a distintos parámetros[50].

- Tratamiento de distintitos tipos de atributos (categóricos, binarios, continuos, etc)
- Representación del modelo
- Cantidad y reducción de atributos
- Costos computacionales
- Interoperabilidad/Habilidad de Explicar
- Generalización, es decir, se ajustan bien a los casos menos comunes

A continuación, se enumeran los métodos más comúnmente usados en modelos de predicción.

2.11.1. Regresión Logística

La regresión logística modela la probabilidad logarítmica, de una variable binaria Y, utilizando una combinación lineal de covariables X:

$$\log(\frac{\mathbb{P}(X)}{1-\mathbb{P}(X)}) = X(\beta)$$

Esta ecuación puede ser fácilmente reordenada para obtener la probabilidad de ocurrencia de Y:

$$\mathbb{P}(X) = \frac{1}{1+\exp(-X\beta)}$$

Los coeficientes o pesos de la variable, β son generalmente ajustado utilizando máxima verosimilitud, las ventajas de utilizar este clasificador radican en que por un lado es un modelo trasparente, que es fácil de interpretar y también familiar para profesionales del aérea de la medicina, y por otro lado es un modelo que ofrece buen ajuste y desempeño, comparable con algoritmos más sofisticados [33].

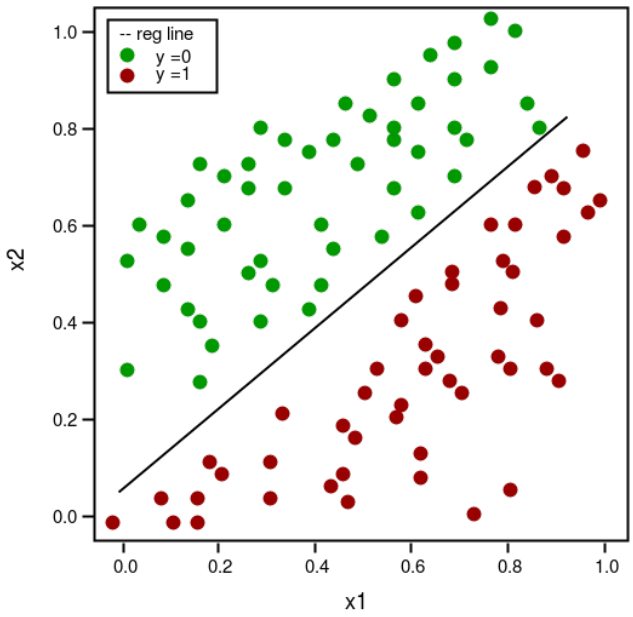


Figura 2.5: Ejemplo de Regresión Lineal
Fuente: Elaboración Propia

2.11.2. Naive Bayes

Naive Bayes es una técnica de clasificación basada en el teorema de Bayes, por lo tanto, se debe asumir dependencia entre los atributos predictores. En términos simples, Naive Bayes asume que la presencia de una variable en particular en una clase, no está relacionado con la existencia de cualquier otra variable. A pesar de que estas variables dependan unas de otras, todas estas propiedades contribuyen independientemente en la probabilidad de ocurrencia de cada clase.

La principal ventaja de que es un modelo simple de construir, de bajo consumo computacional, especialmente útil para bases de datos muy grandes. Además, es un modelo fácil de interpretar, y en general otorga un buen ajuste a los datos, incluso comparable con los métodos de Machine Learning más sofisticados. En términos simples, Teorema de Bayes propone que la probabilidad posterior de ocurrencia de un evento dado un atributo $\mathbb{P}(C|X)$, es inverso a la probabilidad previa $\mathbb{P}(C)$ por la probabilidad condicional del atributo dado $\mathbb{P}(X|C)$ [33].

$$\mathbb{P}(X) = \frac{\mathbb{P}(C|X)\mathbb{P}(C)}{\mathbb{P}(X)}$$

2.11.3. Support Vector Machine

Support Vector Machine son algoritmos de aprendizaje supervisado que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible, cunado las nuevas muestras se ponen en correspondencia con dicho modelo en función de su proximidad pueden ser clasificadas a una u otra clase.

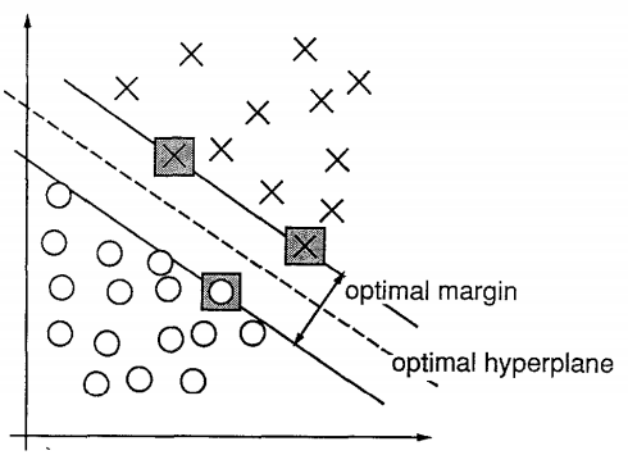


Figura 2.6: Ejemplo de Support Vector Machine
Fuente:Support-vector networks [12]

Este modelo construye un hiperplano en un espacio de dimesionalidad muy alta, que separa las muestras de las distintas clases. El propósito es encontrar el hiperplano de máxima distancia al punto más cercano de cada clase. El set de instancias más cercanos al hiperplano óptimo se le llama support vector, una de sus desventajas es que es difícil de interpretar ya que actúa como una caja negra

2.11.4. Arboles de Decisión

Un árbol de decisión es una estructura jerárquica que representa un modelo de clasificación. Con un modelo de árbol de decisión, puede desarrollar un sistema de clasificación para predecir o clasificar observaciones futuras desde un conjunto de datos de entrenamiento. La clasificación toma la forma de una estructura de árbol donde las ramas representan puntos de división en la clasificación. Los puntos de división dividen los datos en subgrupos de forma recursiva hasta que se alcanza un punto de parada. Los nodos de árbol en los puntos de parada se conocen como hojas. Cada hoja asigna una etiqueta, conocida como etiqueta clase, para los miembros de su subgrupo o clase [33].

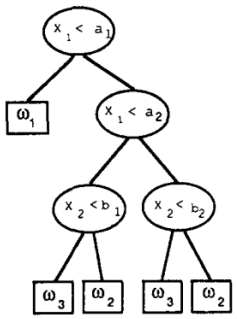


Figura 2.7: Ejemplo de Árbol de Decisión.
Fuente: Survey of decision tree classifier methodology [70]

2.11.5. Random Forest

El método de clasificación Random forest, consisten un una combinación de *tree classifiers* donde cada clasificador genera un vector aleatorio muestreado independiente de la entrada de la entrada del vector, cada *tree classifiers* emite un voto unitario para la clase mas popular para clasificar una entrada vector. al final se clasifica según la clase mas popular votada de todos los predictores [60].

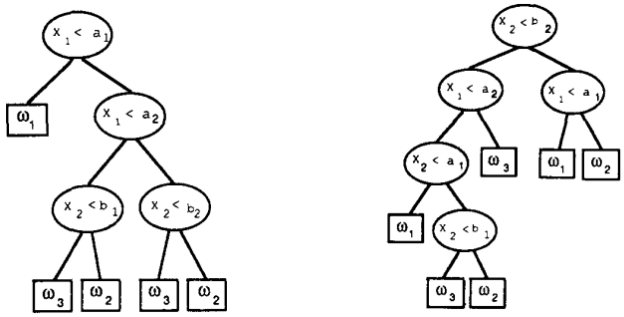


Figura 2.8: Ejemplo de Random Forest
Fuente: Survey of decision tree classifier methodology [70]

2.12. Métodos de clasificación no supervisado

Los métodos de Clasificación no Supervisados son aquellos que se usan donde no hay una variable de resultado para predecir o clasificar. Por lo tanto, no hay .ªprendizaje "de los casos en que tal resultado la variable es conocida. Reglas de asociación, reducción de dimensión. Los métodos y las técnicas de agrupación son todos sin supervisión.

En el aprendizaje no supervisado no se dispone de datos etiquetados para el entrenamiento y un algoritmo recibe un conjunto de entradas no etiquetadas, es decir, se desconocen las salidas. Su fin es buscar agrupamientos basados en características similares para encontrar alguna estructura o forma de organizarlos[40].

2.12.1. K-Nearest Neighbors

K-Means es un método de aprendizaje no supervisado su función es agrupar datos sin etiqueta y encontrar sus centros a partir de dos parámetros: el conjunto de datos inicial y el número deseado de clústeres o agrupaciones [40].

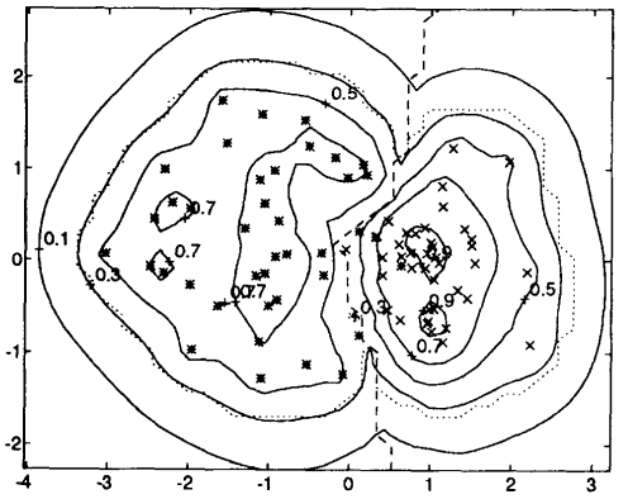


Figura 2.9: Ejemplo de K-Means
Fuente: A k-nearest neighbor classification rule based on Dempster Shafer theory.[19]

2.13. Validación de modelo

Para evaluar el desempeño de los diferentes modelos en minería de datos dada la necesidad de crear y desarrolla modelos precisos capaces de clasificar con el menor sesgo posible se deben considerar los siguientes aspectos:

- Estrategias para el entrenamiento y testeo del modelo
- Tipos de error y medidas de desempeño
- Medidas de ajuste y aplicabilidad

Para lo siguiente se dividen se sugieren distintas metodologías, en un principio se deben definir tanto datos de entrenamiento y datos de tal forma que ambos grupos sean conjuntos mutuamente excluyente, generalmente para validar se usa dos tercios de los datos para el entrenamiento y un tercio para el testeo, el problema de esto último es que ambos conjuntos deben ser representativos, además requiere un gran volumen de datos para validar el modelo, y por últimos se requiere un conjunto de prueba suficientemente robusto y representativo de la muestra.

Una solución a esto es el método de validación cruzada de n-folds (k-fold cross-validation), el cual consiste en la división de la muestra en n subconjuntos disjuntos, en este método con n-1 conjuntos y se evalúa con el conjunto que se dejó fuera del entrenamiento. Esto se realiza n veces, dejando en cada iteración un subconjunto fuera del conjunto de entrenamiento y dejándolo como conjunto de testeo, este tipo de testeo genera modelos que después de n iteraciones tengan un buen desempeño.

2.13.1. Tipo de error

Para realizar la evaluación del modelo relacionado se han desarrollado múltiples medidas de error para estimar y comparar el desempeño en los modelos de clasificación. Para la compresión de estas medidas se requiere de conocer de la matriz de confusión representada a continuación:

Matriz de Confusión		Clase Real	
		Positivo	Negativo
Clase Predicha	Positivo	Verdadero Positivo (TP)	Falso Positivo(FP)
	Negativo	Falso Negativo(FN)	Verdadero Negativo(TN)

Figura 2.10: Matriz de confusión.
Fuente: Elaboración Propia

La Matriz de confusión de un modelo información sobre la clasificación del set de testeo y la predicha por el modelo representado. La figura 2.10 representa la matriz de confusión binario. La matriz de confusión contiene la siguiente información:

- **Verdadero Positivo:** Número de casos en que el clasificador predijo la etiqueta positiva efectivamente poseen etiqueta positiva.
- **Falsos Positivos:** Número de casos en que el clasificador predijo etiqueta positiva cuando posee una etiqueta negativa.
- **Verdadero Negativo:** Número de casos en que el clasificador predijo etiqueta negativa cuando efectivamente poseen etiquetas negativas.
- **Falsos Negativos:** Número de casos en que el clasificador predijo etiqueta negativa cuando posee etiqueta positiva.

2.13.2. Indicadores de rendimiento

Los indicadores mayormente utilizados para evaluar y comparar modelos se describen a continuación:

- **Accuracy:** El Accuracy, es definido como el porcentaje de aciertos sobre la totalidad de la base, explicitando e identificando términos definidos en la matriz de confusión, la expresión que lo define es:

$$Accurancy = \frac{TP + TN}{TP + FP + TN + FN}$$

  Esta métrica es una de las más utilizadas, pero presenta el grave problema de no diferenciar los tipos de error que pueden cometerse. Puesto que un clasificador orientado a maximizar este indicador cuando la base tiene alto desbalance, tendría tendencia a clasificar todas las observaciones como la clase mayoritaria
- **Balance Accuracy:** A diferencia del anterior, este indicador pondera la tasa de acierto de cada una de las clases, esto es la cantidad de aciertos sobre la totalidad de elementos

de la clase, donde i con i = 1, 2 son los ponderadores asociados a cada clase, cumpliendo lógicamente $\alpha_1 + \alpha_2 = 1$

$$Balanced Accuracy = \alpha_1 * \frac{TP}{TP + FN} + \alpha_2 * \frac{TN}{TN + FP}$$

- **Recall (Especificidad):** Corresponde a la proporción de observaciones pertenecientes que son correctamente predichos como positivos:

$$Recall = \frac{TP}{TP + FN}$$

- **Precisión (Sensibilidad):** Ese indicador denota la proporción de las observaciones etiquetadas como positivas que corresponden a la clase positiva:

$$Recall = \frac{TP}{TP + FP}$$

- **F-Measure:** Este indicador, es una combinación de los 2 últimos indicadores descritos (Precisión y Recall), tomando la medida armónica entre ellos:

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

  Esta medida es un poco más robusta que cada una de las anteriores por sí sola, puesto que considera explícitamente los dos tipos de errores en su definición.

- **AUC:** Este procedimiento es un método útil para evaluar la realización de esquemas de clasificación en los que exista una variable con dos categorías por las que se clasifiquen los sujetos [32]. Este indicador es derivativo de la formulación de la Curva ROC y representa el área bajo la curva de dicha curva, la cual En un sentido estadístico, el AUC equivale a la probabilidad de que el clasificador evaluará una instancia positiva elegida aleatoriamente, de igual o mejor forma que una instancia negativa aleatoriamente elegida.

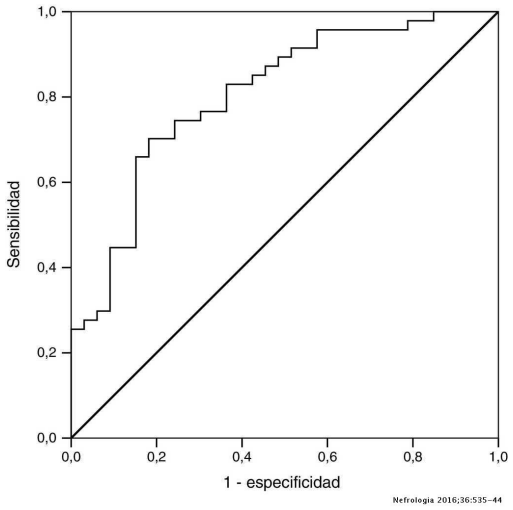


Figura 2.11: Ejemplo de Curva ROC
Fuente: Elaboración Propia

Los puntos del gráfico con un alto Recall y un bajo False Positive Rate representan buen desempeño del modelo, por lo tanto, entre más corrida esté la curva hacia la esquina superior izquierda del gráfico, mayor es el desempeño del modelo. Para cuantificar esto,se utiliza una medida de desempeño llamada AUC (Area under curve), que mide el área bajo la curva ROC en el gráfico.

- **Cumulative Accuracy Profile:** La validación del rendimiento del modelo de calificación se realiza comparando la curva del modelo de calificación, con la curva de un modelo perfecto y la curvade un modelo aleatorio (imperfecto). La curva CAP del modelo de calificación usualmente estará entre la del modelo perfecto y el modelo aleatorio, ocupando una parte del área formada por la curva CAP del modelo perfecto y la del modelo aleatorio. El área de porcentaje ocupada por la curva CAP del modelo de calificación se denomina Relación de precisión (AR). Un AR de alrededor del 40 % al 60 % se considera un modelo con un poder predictivo razonable.

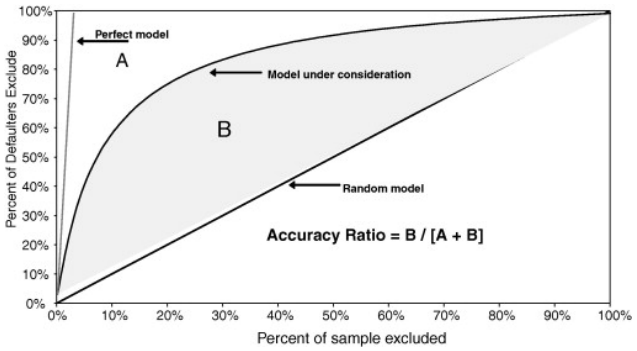


Figura 2.12: Ejemplo de Cummulative Accuracy Profiles
Fuente: Testing rating accuracy [22]

- **Log-Loss:** Mide el rendimiento de un modelo de clasificación donde la entrada de predicción es un valor de probabilidad entre 0 y 1. El objetivo de nuestros modelos de Machine Learning es minimizar este valor. Un modelo perfecto tendría una pérdida de registro de 0. La pérdida de registro aumenta a medida que la probabilidad predicha difiere de la etiqueta real.

  Por lo tanto, predecir una probabilidad de .012 cuando la probabilidad real es 1 sería malo y resultaría en una gran pérdida de registro.A medida que la probabilidad predicha se acerca a 1, la pérdida de registro disminuye lentamente. Sin embargo, a medida que disminuye la probabilidad predicha, la pérdida de registro aumenta rápidamente. La pérdida de registro penaliza ambos tipos de errores, pero especialmente aquellas predicaciones que son confiables y erróneas.

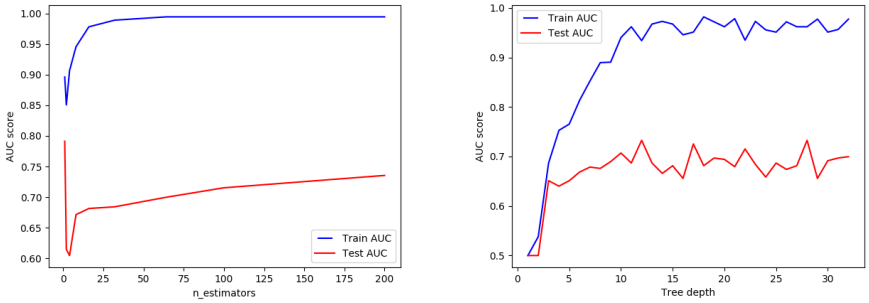
  - N - Numero de Observaciones
  - y_{i} - Indicador binario (0 o 1)
  - p_{i} - probabilidad predicha por el modelo

$$logloss = -\frac{1}{N}\sum_{\mathrm{i}}^{N}(y_{\mathrm{i}}log(p_{\mathrm{i}}) + (1-y_{\mathrm{i}})log(1-p_{\mathrm{i}}))$$

2.14. Overfitting

Overfitting o Sobre-Ajuste es un riesgo que se encuentra en los algoritmos de aprendizaje automático en que el ruido en los datos en la etapa de aprendizaje dicho algoritmo comienza a memorizar varias peculiaridades de la base de datos, por lo tanto el algoritmo aprende de estar particularidades y no encuentra una regla general predictiva, y como consecuencia al modificar pequeños parámetros a lo que pretendemos predecir reduce significativamente la precisión del algoritmo.

Para evitar problemas se debe seleccionar variables que no estén excesivamente correlacionadas entre ellas. Además en los casos de arboles de decisión y Random forest se realizan test con el fin de determinar el numero máximo de nodos sin caer en sobre-ajuste, y en Random Forest también se realizan test para determinar el numero máximo de arboles de decisión que debe formar para no caer en dicho problema. Por ultimo también a K-Nearest Neighbors se le realiza un trabajo similar limitando el numero de vecinos mínimos que requiere para formar dichos clusters,y no caer en sobre ajuste.



(a) Test N arboles de desición (b) Test profundidad de nodos

Figura 2.13: Ejemplos de los test de sobre ajuste en Random Forest.
Fuente: Elaboración Propia

2.15. Validación cruzada

El método de validación cruzada para el elección y evaluación de predicciones estadística. El concepto de tal evaluación es viejo en su forma más primitiva pero útil, consiste en división controlada o no controlada de la muestra de datos en dos submuestras, la elección de un predictor estadístico o método de clasificación, en una submuestra y entonces la evaluación de su desempeño se mide contra el la otra submuestra [76].

En el caso de esta memoria se subdivide el conjunto de datos original en k subconjuntos de igual tamaño, uno de los cuales se utiliza como conjunto de prueba mientras los demás forman el conjunto de entrenamiento. Luego la precisión general del clasificador es calculada como el promedio de las precisiones obtenidas con todos los subconjuntos de prueba[26]

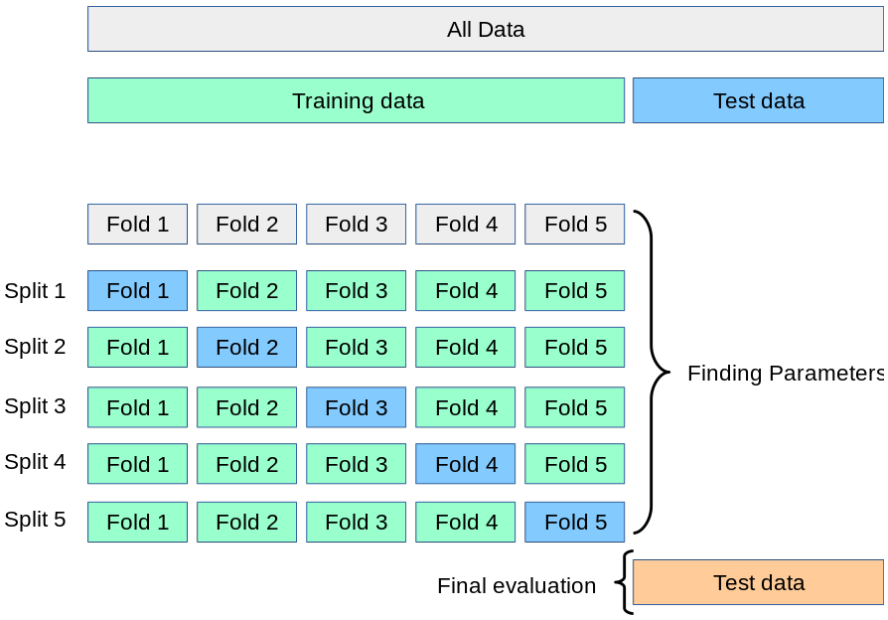


Figura 2.14: Diagrama Cross Validation

Fuente: Cross-validation: evaluating estimator performance [1]

Capítulo 3

Modelo de Predicción

En este proyecto se pretende elaborar múltiples predictores de riesgo de depresión con las variables de la Encuesta Nacional de Salud del 2009-2010 y validarlo de ser posible con los datos entregados en la Encuesta Nacional de Salud del año 2016-2017. Se realizaron distintas etapas en el proceso de formulación de un modelo predicitivo, desde la limpieza y selección de datos, pasando por los distintos metodologías de balance de datos y la implementación de lo modelos. Por último, la selección del los mejores modelos y obtención de los indicadores en cuestión.

3.1. Encuesta Nacional de Salud

La Encuesta Nacional de Salud (ENS), es una herramienta que utiliza el Ministerio de Salud para saber qué enfermedades y qué tratamiento están recibiendo hombres y mujeres mayores de 15 que viven en Chile [15]. La información que arroja esta encuesta es de vital importancia para formular los planes de prevención, atención y las políticas de salud para las personas que lo necesiten. Para ello cuenta con los siguientes objetivos:

- Estimar la prevalencia nacional de los problemas de salud y factores de riesgo definidos para la encuesta.
- Describir estas prevalencias según : sexo, edad, zona urbana/rural, regiones, entre otros.
- Describir la distribución poblacional de los parámetros antropométricos y de laboratorio.
- Calcular la prevalencia de desdentamiento y caries en adultos y describir su comportamiento a nivel nacional, según sexo, edad, zona urbana/rural, regiones y otros.
- Calcular la prevalencia nacional de rezago y retraso del desarrollo psicomotor infantil reportado por el cuidador principal de niños/as de 7 meses a 4 años 11 meses.
- Calcular la prevalencia nacional de problemas de salud mental priorizados en una submuestra de 18 años y más.
- Calcular la prevalencia del nivel de discapacidad por dificultades funcionales de la vida cotidiana y de integración asociadas a los problemas de salud estudiados a nivel nacional

y según subgrupos de enfermedades prioritarias.

- Calcular la prevalencia de cobertura nacional de atención del sistema de salud chileno para un grupo seleccionado de enfermedades prioritarias.
- Desarrollar la seroteca y orinoteca del la ENS en el ISP con las muestras de suero y orina de la encuesta.

En el desarrollo de esta memoria se utilizaron las encuestas nacionales de salud de los años 2009-2010 y 2016-2017 que tienen las siguientes diferencias.

Encuesta 2009-2010	Encuesta 2016-2017
Cuenta con etiqueta de depresión	No cuenta con etiqueta de depresión
Resultados de múltiples exámenes	No cuenta con todos los resultados de exámenes
Cuenta con 1386 variables y 5293 muestras	Cuenta con 1158 variables y 6233

Tabla 3.1: Tabla comparativa ENS 2009-2010 y la ENS 2016-2017

Debido a que entre las encuestas no existen exactamente los mismos exámenes y que estos fueron modificados por el ministerio, se utilizaran solo las preguntas e instrumentos que contiene la encuesta dejando de lado dichos exámenes.

Por último, recalcar existe una cantidad de preguntas que no se encuentran en ambas encuestas y otras que a pesar de que ser relativamente similares muchas veces la escala de posibles respuesta con la que se presentan hace imposible la validación, por lo tanto para la validación de los modelos realizados a posteriori con la base de datos de los años 2016-2017 se realizara dependiendo si las varaibles del modelo cuentan con un homologo en la base de datos del 2009-2010 y que contengan una escala comparativa de respuestas.

3.2. Análisis exploratorio

En una primera instancia se realizo un análisis exploratorio de las variables de la Encuesta Nacional de Salud del 2009-2010, para ello se utilizo el software SPSS y Python, obteniendo los siguientes resultados.

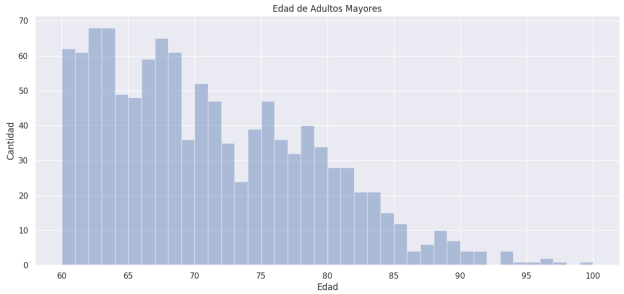
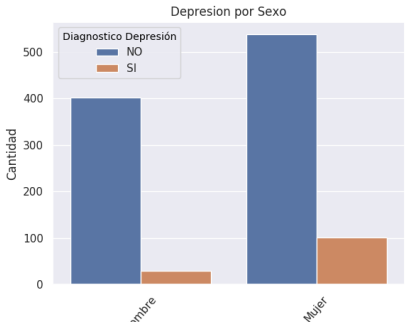
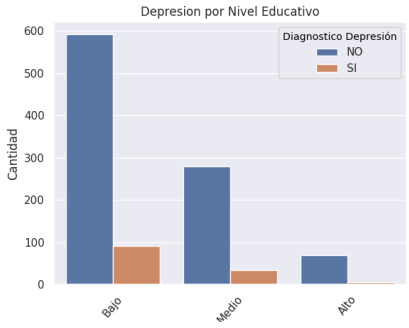


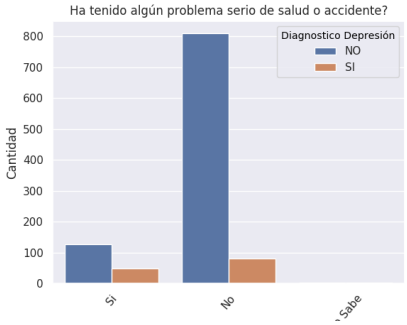
Figura 3.1: Histograma de dispersión edades adulto mayor
Fuente: Elaboración Propia



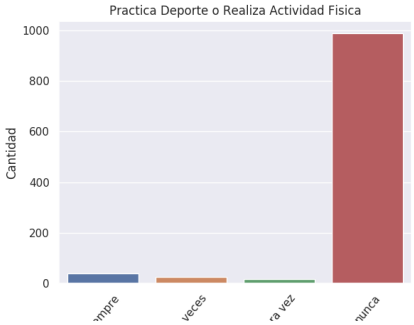
(a) Depresión por sexo



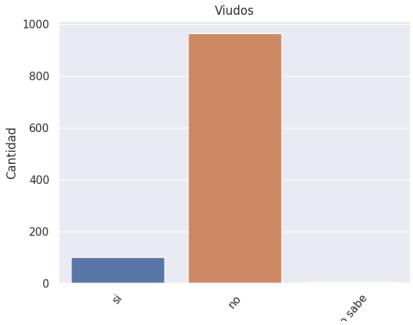
(b) Depresión por nivel educativo



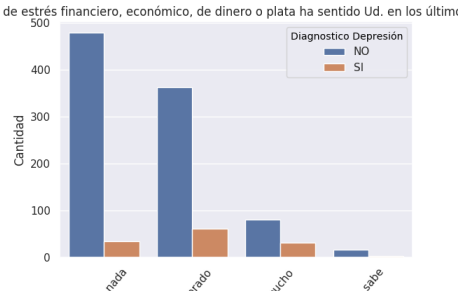
(c) Problemas de salud



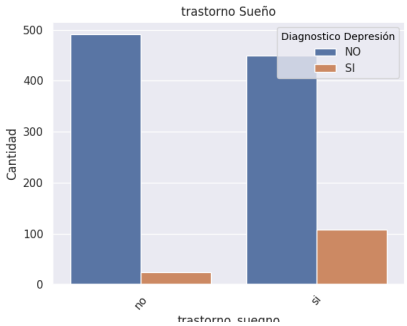
(d) Actividad deportiva



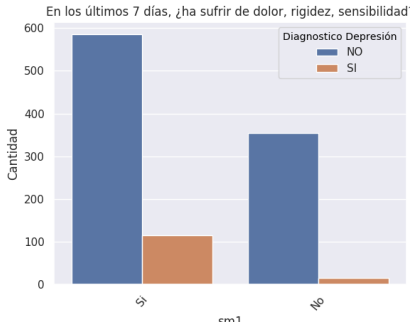
(e) Viudos



(f) Problemas económicos



(g) Problemas de sueño



(h) Sufre de dolor

Figura 3.2: Resultados Modelo Socio-Demográfico.
Fuente: Elaboración Propia

Como se puede ver en la figura (3.1) la dispersión de edades en la encuesta no es uniforme, teniendo en cuenta que solo se focalizara en adultos mayores. Por lo tanto, del total de 5293 solo se dejaron las muestras que tuvieran más a 60 años reduciendo la muestra a un total 1070. También existe el antecedente que el Ministerio de Salud utilizo para la formulación de esta base de datos factores de expansión para los adultos mayores debido a que no se cuenta con una alta representación de los más longevos. También reveló que se cuenta con 1070 muestras de adultos mayores de los cuales 131 tienen depresión, lo que significa que aproximadamente el 12,2 % de los adultos mayores de la encuesta tiene depresión.

Otro punto importante al observar la figura (3.2) es que la mayoría de los encuestados son mujeres alcanzando el 60 % de la muestra y al mismo tiempo se ve que cuentan con una tasa mayor de depresión. También se observa una tendencia que advierte que entre menor nivel educativo mayor es la probabilidad de sufrir depresión, el nivel educacional de los adultos mayores alcanza un 30 % que tiene estudios medios completos y solo el 7,15 % tienes estudios superiores completos. Esta última variable es relevante dado que se encuentra relacionado con el nivel de vida que puede alcanzar un adulto mayor.

En la figura 3.2(d) se ve la poca actividad física al mes que realizan los adultos mayores en Chile, ya sea por cultura, por la falta de motivación o dificultades físicas este habito trae consigo múltiples consecuencias en la calidad de vida de las personas, y se correlaciona en gran medida con los problemas de salud o enfermedades que puedan tener. También se puede ver en la figura 3.2(c) la relación entre los problemas de salud y la presencia de depresión en adultos mayores, cuya tasa es notoriamente superiores en aquellos que hayan tenido un accidente o alguna enfermedad el último año.

En cuanto al estado marital la mayoría se encuentran en pareja y una pequeña porción de estos vive solo, aunque la mayoría de los adultos mayores dice no tener problemas económicos, en la figura 3.2(f) se deja ver una relación clara entre los problemas económicos y el riesgo de sufrir depresión, esto se justifica principalmente por los gastos en salud y medicamentos. Se estudió también unas de las principales secuelas de la depresión como son los trastornos del sueño en la figura 3.2(g) que revelo que aproximadamente la mitad de los adultos mayores sufre de trastornos del sueño.

Por último, destacar una relación entre el dolor y la depresión, como se puede ver en la figura 3.2(h), la mayoría de los adultos mayores sufre de alguna molestia física y claramente se observa una relación entre la cantidad de casos de depresión y los adultos mayores que aseguran tener algún dolor en el último tiempo.

3.3. Construcción etiqueta

La construcción de la etiqueta de depresión en la Encuesta Nacional de Salud de 2016-2017 se debe a que base no cuenta con esta etiqueta. Por lo tanto, se elaboró una etiqueta artificial según el módulo de preguntas depresión basados en el DSM-IV de la encuesta (figura 3.3). Para el diagnostico de depresión se debe cumplir con los siguientes puntos:

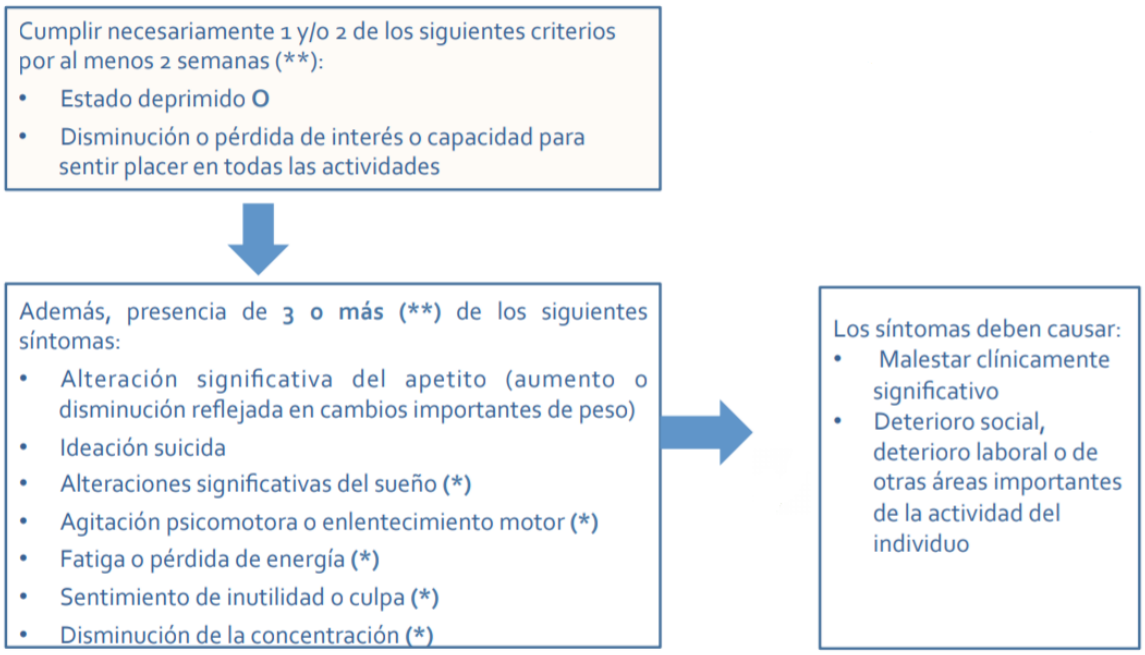


Figura 3.3: Condiciones para etiqueta de depresión según modulo DSM-IV de la encuesta nacional de salud 2016-2017

Fuente: Criterio DSM-IV Encuesta Nacional de Salud 2016-17.[23]

Para esto, solo se consideró a los sujetos que cuentan con el instrumento de diagnostico de depresión completo y que fueran adultos mayores (mayores de 60). Primero se contabilizo la cantidad de síntomas que tenía cada uno de los adultos mayores y se segmentaron según los 3 criterios, posteriormente según las reglas de DSM-IV para depresión se procedió a generar la etiqueta para cada uno de los adultos mayores.

En esta base de datos la distribución de depresión resulto en una menor cantidad de casos de casos respecto a la Encuestas Nacional de Salud 2009-2010 alcanzando aproximadamente el 7,2 % de casos de depresión en adultos mayores 3.4.

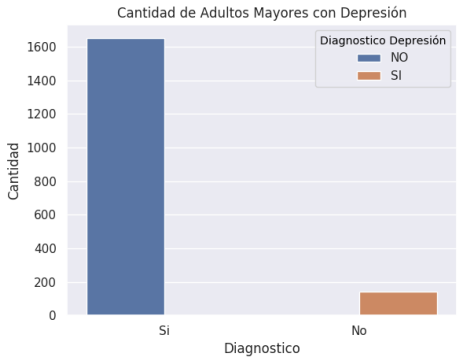


Figura 3.4: Distribución etiqueta de depresión según DSM-IV en Encuesta Nacional de Salud 2016-2017

Fuente: Elaboración Propia

3.4. Missing Data

Para el tratamiento de *Missing Data* se consideró para ambas bases de datos solo a los adultos mayores que ya contaran con la etiqueta de depresión. A continuación se contabilizaron todas las variables con valores nulos o que no tuvieran concordancia (*outliers*), como se ve en la tabla **??**. Además, para ello solo se consideraban las variables que contaran con más del 80 % de sus datos para no generar demasiado sesgo en los modelos de Machine Learning.

Nombre	Cantidad
Resultados de Trastornos del Sueño	154
Resultados de Riesgo Cardiovascular	51
Resultado de Diabetes Insulino	74
Glaucoma Ocular	20
Peso(Kg)	21
Otros	361

Tabla 3.2: Missing Data

Posteriormente a esta labor se prosiguió a impugnar datos mediante las siguientes técnicas:

1. **K-Nearest Neighbor (KNN)**: Este tipo de impugnación se utilizó en los casos en que existía alguna herramienta dentro de la encuesta con la cual relacionar la variable y poder impugnarla, como ejemplo, sea la variable trastornos del sueño, se utilizó las variables y el instrumento para determinar si un sujeto tiene trastornos del sueño para impugnar los resultados faltantes.
2. **Promedio**: En caso que fueran datos aislados respecto al resto de variables, como por ejemplo el nivel educacional, se procedió a impugnar con el promedio de redondeado al entero más cercano.
3. **Moda**: La moda se utilizó en el caso de tener columnas con variables binarias, como por ejemplo son los casos de sexo o resultados de ciertas patologías.

Como último paso se prosiguió reemplazar los valores nulos con los nuevos valores y con ello se generó una base sin missing values, además en esta etapa se quitaron outliers de dicha base de datos.

3.5. Selección variables relevantes en depresión

En la recopilación de infracción sobre depresión se da a entender que existen múltiples factores que se combinan, potencian y aumentan conjuntamente la vulnerabilidad de una persona de sufrir depresión o de incorporar un comportamiento suicida[56]. Por lo tanto, las variables que pueden ser factor de depresión no son claras y pueden variar según la realidad personal, social, de la comunidad, etc.

A cada uno de estos subconjuntos de variables se les procedió realizar 4 test estadísticos, dos de ellos para ver el comportamiento de cada una de estas variables con respecto a la distribución de la variable depresión, con ello vislumbrar las más relacionadas con dicho trastorno, los test utilizados para ellos fueron los de Chi-Cuadrado y el test de Kolmogorov-Smirnov. Se utilizaron también dos test para evaluar la relación entre variables con la finalidad de evitar problemas de sobre-ajuste en el entrenamiento del algoritmo. Los test utilizados para ello fueron los tests de Anova y de Correlación de Pearson (se descartó aquellas variables con una correlación mayor a |0,4|). A continuación, se la serie de pasos que se siguieron para la selección de variables.

Figura 3.5: Diagrama de proceso de selección
Fuente: Elaboración Propia

3.5.1. Socio-demográficos

En el caso de las variables socio-demográficas han demostrado que pueden influir en la propensión en que los adultos mayores pudiesen presentar algún síntoma depresivo como puede ser la edad, el sexo, el nivel educativo, el nivel socio económico entre muchas otras variables. Las variables socio-demográficas utilizadas de la encuesta nacional de salud se muestran en la tabla (Anexo A.2).

Primero, se tomaron todas las variables socio-demográficas de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor es posible validar la hipótesis de independencia de las frecuencias observada entre las variables socio-demográficas y la variable depresión. Este test reveló (Anexo A.3) la preponderancia de variables como el sexo(*SEXO*), la edad (*EDAD*) y nivel educativo (*NEDU, ns9c*) en el de riesgo de sufrir depresión, y aparecen variables que hacen referencia a si el entrevistado viven en un zona urbana o rural (*ZONA*),esto se puede deber al hecho que muchas de las muestras corresponden a sectores urbanos y tan solo el 15 % a rurales. Por último, se encontró una pequeña relación entre el número de personas que viven en el hogar(*nper*) y la depresión, esto hace sentido de la mirada que el aislamiento social es uno de los cuantos gatillantes de síntomas depresivos.

Posteriormente se realizo el test de **Kolmogorov-Smirnov**, se tomaron todas las variables socio-demográficas y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (anexo A.4) se vuelve a ratificar la importancia de el sexo (*SEXO*) como factor de depresión y el nivel educativo mediante la variable (*ns9c*), aparecen nuevas opciones como son la relacion que tiene el sujeto con el jefe de hogar (*ns3*) junto con alguna actividad laboral que estuviese realizando (*ns10*).

Por ultimo, se efectuó un test de **ANOVA** (anexo A.5) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.1) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|. Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables socio-demográfica arrojo similitud entre variables como (*Zona*) con las variables (*ns20*), (*ns20*) que están asociadas al tipo de vivienda.

3.5.2. Actividad física

El caso de las variables sobre la actividad física han demostrado que pueden influir en la propensión en que las personas pudiesen contener algún síntoma depresivo, en los adultos mayores la bibliografía rescata la posibilidad que el trabajo físico de fuerza ayuda en gran medida a la superación de la depresión. Las variables de actividad física que destacan en la encuesta son por ejemplo la cantidad de actividad física que tiene a la semana un adulto mayor o si en su trabajo lo obliga a ejercer una actividad física entre muchas variables(Anexo A.6).

Primero, se tomaron todas las variables de actividad física de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de actividad física y la variable depresión. Este test reveló (Anexo A.7) que contra lo formulado en la múltiple bibliografía recopilada las variables de actividad física no presentan según las pruebas relación con trastornos depresivos, esto es justificado eventualmente debido a la baja cantidad de actividad física que realizan los adultos mayores de la base de datos y por lo tanto no se puede descartar lo contrario(figura 3.2).

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de actividad física y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov en congruencia a lo obtenido mediante el test de Chi-Cuadrado, arrojó que todas de las variables de actividad física eran significativamente diferentes y no se veían relacionadas con la etiqueta de depresión, pero tampoco se puede descartar que no tengan relación con esta por la baja actividad física que realizan los adultos mayores en Chile.

Por último, se efectuó un test de **ANOVA**(anexo A.8) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.2) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables actividad física arrojo

similitud entre variables (*a2*) y (*a14*) ambas tenían relacion con la frecuencia y cantidad de días que una persona realizaba actividad física.

3.5.3. Determinantes sociales y psicológicos de salud

Las variables sobre la percepción del estado de salud y aquellas determinantes de carácter social y apoyo psicológico en los adultos mayores son sumamente relevantes dado a que definen el grado de aislamiento social y el constructo mental que pudiese tener. Estas variables albergan algunas de las principales causas por las que un adulto mayor pudiese caer en depresión. Muchas de estas variables son sumamente relevantes y se pueden ver en la tabla(anexoA.9)

Primero, se tomaron todas las variables de determinantes sociales y psicológicos de salud de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observadas entre las variables de determinantes sociales y psicológicos y la variable depresión.Este test reveló (Anexo A.10) resultados que refuerzan la bibliografía que relaciona los problemas personales ya sean laborales o familiares*p8e*, financieros (*dataF1_ 2p7*), estres, ansiedad (*dataF1_ 2p6*), problemas de salud (*p8f*) o la pérdida de un ser querido (*p8g*) con depresión. Además, se encontró una relación depresión entre los adultos mayores y el barrio donde se encuentra su hogar (*p2a*), lo que refuerza la importancia de crear entornos seguros y agradables para los adultos mayores.

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de determinantes sociales y psicológicos de salud y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.11) ratificó los resultados expuestos en la sección anterior donde las variables asociadas a la de ansiedad (*dataF1_ 2p6*), estrés,problemas económicos (*dataF1_ 2p7*), problemas de salud (*p8f*) están fuertemente vinculadas con los trastornos depresivos, lo que concuerda con la bibliografía.

Por último, se efectuó un test de **ANOVA**(anexo A.33) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Ademas, se les realizo un test de **Correlación de Pearson** (anexo A.1) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables de determinante social arrojó similitud entre variables (*dataF1_ 2p3*) y (*dataF1_ 2p4*) donde amabas variables tocan el apoyo que pueda tener una persona con su entorno familiar y amigos. Además, se aprecia una relacion entre los niveles ansiedad de una persona (*dataF1_ 2p6*) y si este cuenta con apoyo de un familiar (*dataF1_ 2p4*).

3.5.4. Perspectiva de salud y auto-reporte

El envejecimiento supone cierto número de cambios fisiológicos, anatómicos, psicológicos y sociales, es decir, una declinación en la función del organismo como un todo. El deterioro de las capacidades biopsicosociales del adulto mayor trae consigo un cambio en la posición y función que desempeña la sociedad y por lo tanto se ve fuertemente vinculado en la relación que tiene de su entorno y consigo mismo, que en el caso de ser severamente negativos están correlacionado con la probabilidad de contraer síntomas depresivos o incluso llegar a desarrollar dicho trastorno. En la tablaA.13 se muestras las principales variables asociadas a la perspectiva de salud.

Primero, se tomaron todas las variables sobre la perspectiva de salud y auto-reporte de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables sobre la perspectiva de salud y auto-reporte y la variable depresión. Este test revelo (Anexo A.14) que la mayoría de las variables sobre la perspectiva de salud que se presentan en la encuesta tienen una mayor o menor relación con la depresión, pero al mismo tiempo se encuentran profundamente correlacionadas entre ellas. Las principales variables fueron relacionadas con el estado de ánimo actual(*cd31*), seguido por los problemas ligados a la movilidad física y las dificultades con las que llevar las tareas cotidianas por ultimo las variables(*cd7* relacionadas con los conflictos en su entorno social y/o laboral(*cd27*).

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables sobre la perspectiva de salud y auto-reporte y se contrastaron con respecto a la etiqueta de depresión de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.15) arrojaron en términos generales resultados similares a los entregados por el test Chi-cuadrado, teniendo entre las principales las variables las que resumen el estado anímico y emocional (*cd31*) de los adultos mayores Pero también cabe destacar que salieron entre las más relacionadas aquellas como el dolor (*cd20*) y las molestias físicas de los adultos mayores.

Por último, se efectuó un test de **ANOVA**(anexo A.33) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.3) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación,se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables de perspectiva de salud y auto-reporte arrojó similitud entre una multitud de variables, ya que a mayoría se encontraba relacionada de una manera u otra, ejemplo son la relación entre los problemas emocionales (*cd7*) y la presencia de problemas físicos en las personas (*cd13*).

3.5.5. Problemas cardiovasculares

A pesar que la relación entre la depresión y los problemas cardiovasculares se explica estableciendo a la depresión como un factor de riesgo en el desarrollo de enfermedades coronarias o que aumenta el riesgo de sufrir otro infarto por aquellos que ya hayan sufrido uno, existe una relación importante entre estas variables que no se puede pasar por alto. Las principales variables se ven en la tabla (anexo A.17)

Primero, se tomaron todas las variables de problemas cardiovasculares de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de problemas cardiovasculares y la variable depresión. Este test revelo (Anexo A.18)cierta relación significativa entre los dolores cerca del pecho(*d1*) y aparición de trastornos depresivos, junto con el riesgo cardiovascular(RCV_MALTO) el cual está relacionado con el sedentarismo, exceso de peso, tabaquismo, hipertensión, diabetes y múltiples malos hábitos de salud[37].

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de problemas cardiovasculares y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.20)acotaron aun más las variables relacionadas a la depresión ,quedando la relacion entre el riesgo cardiovascular con la depresión (RCV), además del dolor crónico (*d1*) en el pecho como potenciales y algunos accidentes vasculares , trombosis o derrames cerebrales(*af2_2*) como variables de riesgo. Por último quedo en manifiesto los reportes de colesterol alto que al igual que el riesgo cardiovascular se relacionan con obesidad, tabaquismo y malos hábitos de salud.

Por último, se efectuó un test de **ANOVA**(anexo A.19) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.5) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables de problemas cardiovasculares arrojó similitud entre la cantidad de infartos (RCV) y los riesgos cardiovasculares (*d8_4*).

3.5.6. Problemas respiratorios crónicos

Los problemas respiratorios tienen una relación con la depresión, estudios arrojan que los pacientes con alguna enfermedad pulmonar crónica son más propensos a caer en depresión en relación a otros que no la tienen, llegando a tener inclusive 2.5 veces más opciones de caer en estos trastornos. Las variables relativas a problemas respiratorios se ven en la tabla (anexo A.21).

Primero, se tomaron todas las variables de problemas respiratorios de salud de la en-

cuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables problemas respiratorios y la variable depresión. Este test reveló (Anexo A.22) que las variables de problemas respiratorios que se vieron mayormente relacionadas con depresión son aquellas que afectan en el día a día a los adultos mayores y en sus tareas cotidianas, como puede ser el asma(*sospecha_asma*), problemas respiratorios al caminar(*respir_tos*), tos o silbidos al respirar(*respir_sibil*).

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las problemas respiratorios y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.23) reveló que la principal variable relacionada con depresión son aquellas asociadas con problemas en la capacidad de caminar (*r13*), de todos modos esto se debe principalmente a la orientación que tienen las preguntas de la encuesta nacional de salud, pero si se puede corroborar la existencia y la relación entre problemas respiratorios y depresión entre los adultos mayores.

Por último, se efectuó un test de **ANOVA**(anexo A.24) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.6) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables de problemas respiratorios arrojó similitud entre variables asociadas a la movilidad (*cd14*) y los problemas respiratorio cronicos (*cd14*).

3.5.7. Hábitos alimenticios

Los trastornos alimenticios es una de las principales causas de depresión lo que se vincula estrechamente en los test de depresión DSM-IV y CIE-10, por otro lado, el estado nutricional y la dieta que lleva a una persona pueden estar vinculados con ciertos trastornos alimenticios, basándonos en las variables que contiene la encuesta nacional de salud, emergió la probabilidad de encontrar ciertas relaciones entre la alimentación de un sujeto con depresión, las variables utilizadas para esto se muestras en la tabla (anexo A.25)

Primero, se tomaron todas las variables sobre hábitos alimenticios de salud de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables sobre hábitos alimenticios y la variable depresión. Este test reveló (Anexo A.26) una pequeña relación entre ciertos hábitos alimenticios con la depresión, como fue el consumo de frutas y verduras(*nofrutyverd7d,die4*),pero en general los hábitos alimenticios con los que contaba la encuesta nacional de salud no tuvieron gran impacto en que las personas sufrieran depresión.

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las varia-

bles sobre hábitos alimenticios y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov demostraron que las variables asociadas a hábitos alimenticios de la encuesta nacional de salud no son del todo concluyentes para determinar depresión, principalmente dado a que la depresión se ve determinada por cambios drásticos en estos hábitos y para evaluarlo de mejor manera se requeriría tener datos longitudinales o otro tipo de preguntas similares a las realizadas en los test DSM-IV o CIE-10.

Por último, se efectuó un test de **ANOVA**(anexo A.27) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.7) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables de Hábitos alimenticios arrojó similitud entre variables asociadas a la cantidad y frecuencia de consumo de ciertos alimentos como frutas, pescado y vegetales.

3.5.8. Problemas de audición y visión

Los problemas de audición y visión son unos de los problemas que más se reitera entre los adultos mayores y se agudizan al pasar de los años, según estudios realizados con estos tipos de discapacidad se dice encontrar una fuerte relacion entre los problemas de audición y visión con la aparición de depresión afirmando que las personas que sufren de este tipo de problemas tienen un 30 % más probabilidades de tener depresión, las variables se pueden ver en la figura A.28.

Primero, se tomaron todas las variables de problemas auditivos y de visión de la encuesta nacional de salud del año 2009-2010 y se realizo un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de problemas auditivos y de visión y la variable depresión. Este test revelo (Anexo A.26) que la percepción propia de visión(*aud_1*) y de audición (*v2*)de los adultos mayores es la variable que más afecta en las personas que tienen trastornos depresivos, lo que se respalda en la bibliografía que asegura que los problemas de audición y visión se correlacionan con la depresión.

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de problemas auditivos y de visión y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.23) El test de kolmogorov-Smirnov concuerdan con el test de Chi-cuadrado resaltando las variables de percepción de audición(*aud_1*) y visión(*v2*) por sobre las mismas variables de diagnostico como son la mala visión (*usa_lentes,mala_vision,cataratas,glaucoma*)o los diagnósticos de problemas auditivos(*prob_aud_1mas,prob_aud_los3,escucha_normal*).

Por ultimo, se efectuó un test de **ANOVA**(anexo A.32) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Ademas, se les realizo un test de **Correlación de Pearson** (anexo A.8) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación,se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables visuales arrojó similitud entre variables asociadas a la calidad de vista y preguntas relacionadas con las dificultades de mirar lejos, leer o si ocupaba gafas.

3.5.9. Trastornos del sueño

Los trastornos del sueño como los trastornos alimenticios son una de las principales consecuencias e indicios de que una persona esta sufriendo depresión, Aproximadamente, el 90 % de los pacientes con depresión se quejan de un deterioro tanto en la cantidad como en la calidad del sueño. La alteración del sueño asociada más frecuentemente a un episodio depresivo mayor es el insomnio. Son habituales los problemas para iniciar y mantener el sueño. Con menor frecuencia, existen sujetos depresivos que se quejan de exceso de sueño (hipersomnia) en forma de episodios de sueño nocturno prolongado o de un aumento del sueño diurno. Algunas veces, el trastorno del sueño es la razón por la que el paciente depresivo acude en busca de tratamiento. La variables asociados a estos trastornos se observan en la tabla(anexosA.34)

Primero, se tomaron todas las variables de sobre los trastornos del sueño de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de sobre los trastornos del sueño y la variable depresión.Este test revelo (Anexo A.35) resultados muy coherentes respecto a lo que se esperaba por la recopilación bibliográfica dado a la relevancia de los trastornos del sueño dentro de la depresión, donde la principal sensación es la de no poder descansar o despertar agotado y sin energías(*ts4*).

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de sobre los trastornos del sueño y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.36)volvieron a ratificar el hecho de la importancia y relacion de los trastornos del sueño en la depresión, especialmente despertar sin energías(*ts4*), pesadillas(*ts10*), parálisis del sueño(*ts9*) y dificultades para mantenerse despierto durante el dia(*ts3*).

Por último, se efectuó un test de **ANOVA**(anexo A.37) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.9) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación,se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables asociadas al

sueño arrojó similitud entre variables entre la calidad del sueño (*ts5*) y si ha tenido sueños desagrables en el ultimo tiempo (*t10*).

3.5.10. Dolor crónico y fracturas

Estudios revelan la correlación que existe entre los dolores crónicos y las fracturas con los adultos mayores que sufren depresión, inclusive la depresión es capaz de generar dolores somáticos en los adultos mayores que se dirigen a un centro asistencial, lo diagnostican con otro tipo de patología cuando el origen del dolor es debido a la depresión. Además hay que tener en consideración que el dolor y la depresión comparten una vía neuroquimica común, las variables de la encuesta nacional de salud se pueden ver en la tabla (anexo A.38)

Primero, se tomaron todas las variables de dolores crónicos y fracturas de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las las variables de dolores crónicos y fracturas y la variable depresión.Este test revelo (Anexo A.39) que las principales variables de dolores relacionadas con la depresión son los dolores vinculados a las articulación(*sm3_5,sm3_9,sm3_2*,sm3_10) y la sensaciones de dolor en general(*sm1*), por ultimo el tener un pasado con múltiples fracturas no fue vinculante con la depresión.

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las las variables de dolores crónicos y fracturas y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.40 arrojaron la preponderancia de los dolores en las articulación y en la sensación de dolor en los últimos días. También, cabe destacar la preponderancia que tuvo en este test el numero de caídas que tuvo en el ultimo año el adulto mayor(*o5*).

Por último, se efectuó un test de **ANOVA**(anexo A.41) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Ademas, se les realizo un test de **Correlación de Pearson** (anexo A.10) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación,se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables de dolor crónico arrojó similitud asociadas al dolor en las distintas articulaciones como son rodilla, muñecas, etc.

3.5.11. Alcoholismo y tabaquismo

El tabaquismo y alcoholismo esta asociado a múltiples hábitos que desencadenan problemas de salud, enfermedades crónicas, problemas familiares y sociales que se relacionan con la depresión. Las variables de la evaluación de estos en la encuesta nacional salud se ven en

la tabla (anexo A.42).

Primero, se tomaron todas las variables de alcoholismo y tabaquismo de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de alcoholismo y tabaquismo y la variable depresión.Este test revelo (Anexo A.44 y A.43) dentro de las variables relacionadas al tabaquismo 2 que fueron más relevantes (*ta11*), como son el hecho de que el adulto mayor viva en un ambiente con de humo de cigarrillo dentro de su hogar y el hecho de ser un fumador activo (*fum_ act30d*). Por otro lado, las variables relacionadas con el alcohol fueron más preponderantes que las relacionadas al tabaquismo, en especial si este consumo se desarrolla de manera periódica(*m7p9*).

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de alcoholismo y tabaquismo y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.45) arrojo ninguna relación de depresión con el tabaquismo, pero dejo en evidencia la relación entre el consumo periódico de alcohol(*m7p9*) y depresión en adultos mayores.

Por último, se efectuó un test de **ANOVA**(anexo A.46 y A.47) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Ademas, se les realizo un test de **Correlación de Pearson** (anexo A.11) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables asociadas al tabaco y al alcohol arrojó similitud entre variables asociadas con la frecuencia de consumo de cigarrillo. Además, de una relacion entre la edad a la que comenzó a fumar (*m7p9*) y si aun persiste fumando (*m7p9*).

3.5.12. Diabetes

Múltiples estudios recalcan la relación que existen entre la depresión y la diabetes, algunos de estos declaran que los sujetos con diabetes duplican la probabilidad de depresión en comparación con otro que no tenga en ambientes similares. Las variables relacionadas con la diabetes en la encuesta nacional se muestran en la tabla A.48.

Primero, se tomaron todas las variables de diabetes de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de diabetes y la variable depresión.Este test reveló (Anexo A.49) que las principales variables relacionadas con la depresión son los niveles de azúcar *di1* y el padecer diabetes, pero la relación en general de estas variables son bastante limitadas.

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las varia-

bles de diabetes y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov fueron contundentes al revelar que ninguna de las variables relacionadas a diabetes fueron realmente determinantes de que un adulto mayor sufra depresión.

Por último, se efectuó un test de **ANOVA**(anexo A.50) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Además, se les realizo un test de **Correlación de Pearson** (anexo A.12) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación,se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En el caso de las variables diabetes arrojó similitud entre si el entrevistado tiene diabetes (*dataHR2p2*) y si esta en tratamiento de la diabetes (*di6*).

3.5.13. Problemas digestivo

A pesar que los problemas digestivos no están asociados a depresión, existen ciertas patologías como colon irritable y patologías intestinales como estreñimiento, diarrea, hinchazón, dolores y molestias estomacales que se relacionan profundamente con problemas de ansiedad y esta se encuentra profundamente vinculada con la depresión, por lo tanto se decidió evaluar si existe alguna relacion, las variables se ven en la tabla (anexo A.51).

A continuación, se tomaron todas las variables de problemas digestivos de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de problemas digestivos y la variable depresión. Este test reveló (Anexo A.52) que las principales variables relacionadas con la depresión, son los problemas relacionados con dolores(*m9p1*) y problemas al defecar(*m9p11*).

Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de problemas digestivos y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.53) Los resultados del test de Kolmogrov-Smirnov son similares a los obtenidos por el test de Chi-Cuadrado donde resaltan los dolores abdominales (*m9p1*) y los problemas al momento de defecar(*m9p11*).

Por último, se efectuó un test de **ANOVA** para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Ademas, se les realizo un test de **Correlación de Pearson** (anexo A.13) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación,se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En este caso particular ninguna de las variables se correlacionaba entre ellas principalmente por las pocas variables con las que se contaba.

3.5.14. Otros

Las variables evaluadas en esta sección son aquellas que no encajaron en los grupos anteriores pero que guardan cierta relacion con la depresión, en este grupo se manejan variables relativas a la obesidad, sexo, memoria, dependencia y antecedentes de depresión, las variables se pueden observar en la tabla (anexo A.54).

Primero, se tomaron todas las variables de sobre obesidad, sexo, memoria, dependencia y antecedentes depresivos de la encuesta nacional de salud del año 2009-2010 y se realizó un test de **Chi-cuadrado** con el fin determinar si mediante el estadístico Chi-Cuadrado y el P-valor validar la hipótesis de independencia de las frecuencias observada entre las variables de sobre obesidad, sexo, memoria, dependencia y antecedentes depresivos y la variable depresión.Este test reveló (Anexo A.55) que muchas variables relacionadas a la condición física y al sexo no son relevantes en la predicción de depresión, en cambio las variables relacionadas a la dependencia (*pe11*), antecedentes de depresión(*sd27*) y de memoria(*e1*) están estrechamente relacionadas con la aparición de trastornos depresivos.

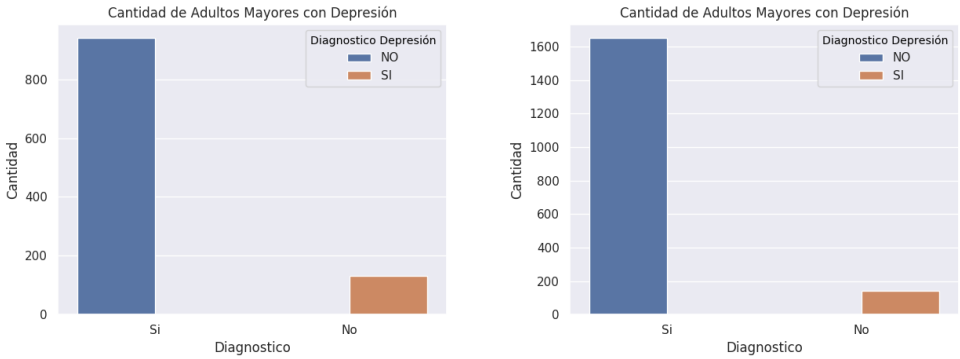
Posteriormente se realizó el test de **Kolmogorov-Smirnov**, se tomaron todas las variables de sobre obesidad, sexo, memoria, dependencia y antecedentes depresivos y se contrastaron con respecto a la etiqueta de depresión, de esta forma determinar el grado en que las dos variables son significativamente diferentes entre sí. Los resultados del test de Kolmogorov-Smirnov (Anexo A.56 corroboro las deducciones realizadas pro el test de Chi-cuadrado espacialmente la obesidad (*n3*), memoria(*e1*) y antecedentes de depresión(*sd27*). Pero en esta oportunidad los problemas de dependencia no se vieron relacionados con la depresión, y eso se debe a la cantidad de casos con dependencia en la encuesta.

Por último, se efectuó un test de **ANOVA**(anexo A.57) para descartar variables cuyas medias estuvieran demasiado relacionadas, y con ello evitar problemas en el desarrollo del algoritmo. Ademas, se les realizo un test de **Correlación de Pearson** (anexo A.14) para medir el grado de relación entre las variables y descartar aquellas que tuvieran una correlación mayor a |0,4|.Por lo tanto, si dos variables salen significativamente relacionadas entre si según el test de Anova y de correlación, se descartaba aquella que tuviera los peores resultados en los test de Chi-cuadrado y de Kolmogorov-Smirnov. En este caso particular ninguna de las variables se correlacionaba entre ellas principalmente por las pocas variables con las que se contaba y que no pertenecían al la misma área de la medicina.

3.6. Balance de base de datos

Después de la selección de variables se observó que existía una gran diferencia entre las etiquetas positivas de depresión respecto a las negativas, el desbalance dentro de la base de datos en las Encuestas Nacionales de los años 2009-2010 y 2016-2017 afecta de manera drástica la forma y la eficiencia del algoritmo de Machine Learning, ya que al algoritmo le es imposible encontrar patrones por la baja densidad de las etiquetas positivas, por lo tanto realizar técnicas de balance para la etapa de aprendizaje es fundamental para tener resultados que sean congruentes y un algoritmo eficiente. Como se puede ver a continuación en la figura

3.6 la diferencia entre etiquetas es evidente en ambas bases, donde las etiquetas positivas representan el 12,1 % de la base del 2009-2010 y el 7,2 % de la base del 2016-2017.



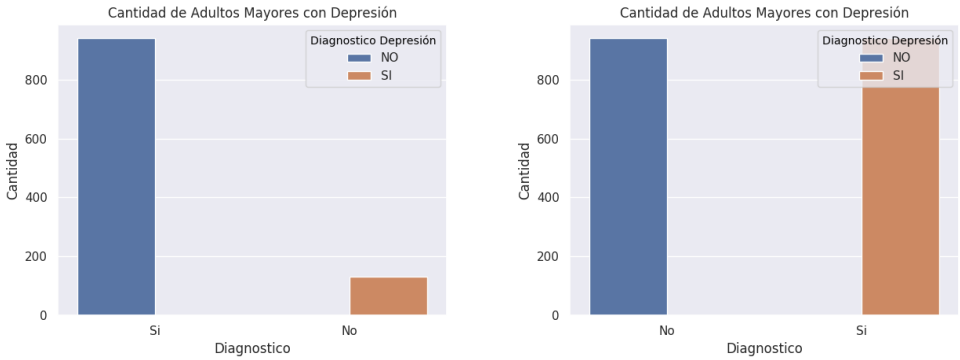
((a)) Etiquetas del ENS 2009-2010 ((b)) Etiquetas del ENS 2016-2017

Figura 3.6: Desbalance en ENS 2009-2010 Y 2016-2017
Fuente: Elaboración Propia

A continuación, se detallan las formas en que se procedió a balancear la base de datos y las distintas metodologías aplicadas.

3.6.1. Over-Sampling

Dado el caso de desbalance se tiene las técnicas de Over-Sampling, estas técnicas tienen como propósito balancear las bases de datos , incluyendo datos sintéticos que se crean a partir de los datos minoritarios en la base de datos, en este caso particular la etiqueta minoritaria es el padecer depresión. Por lo tanto para generar el balance se utilizó la técnica SMOTE-NC (tabla3.7), que es una técnica que genera datos sintéticos con variables continuas y discretas , de esta forma se logra balancear de las etiquetas positivas y negativas Como se puede ver a continuación:

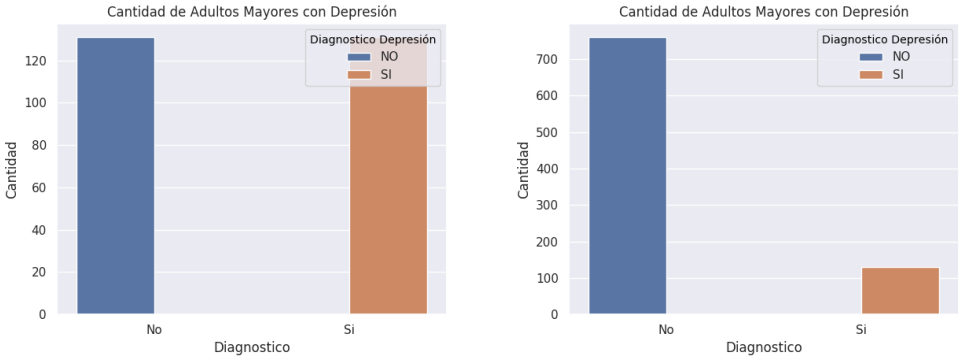


((a)) Etiquetas ENS 2010 ((b)) Etiquetas ENS 2010 SMOTE-NC

Figura 3.7: Balance etiquetas mediante Over-Sampling
Fuente: Elaboración Propia

3.6.2. Under-Sampling

En contra posición a los métodos de Over-Sampling existen técnicas que pretenden sintetizar y reducir el volumen de las etiquetas mayoritarias de tal forma generar el balance entre etiquetas positivas y negativas, a estos métodos se les denomina Under-Sampling. En este caso se utilizaron ClusterCentroids y Nearest Neighbours, la primera pretende reemplaza el grupo mayoritario por el centroide de cluster de un algoritmo K-mean, lo segundo aplica un algoritmo de vecinos más próximo(KNN) y quita los elementos del conjunto que no concuerden con su vecindario.



((a)) Etiquetas ENS 2010 ClusterCentroids ((b)) Etiquetas ENS 2010 Nearest Neighbours

Figura 3.8: Balance etiquetas mediante Over-Samplings
Fuente: Elaboración Propia

3.7. Transformación de variables

Después de seleccionar las variables relevantes de la base de datos, complementarla, quitar los *missing values* y generar el balance de datos se procede a realizar una trasformación de las variables con la finalidad que la magnitud y escala de estos no afecte el proceso de aprendizaje del algoritmo , para ello se procede a realizar una normalización de las variables según la formula:

$$Normalizada = \frac{(x - x_{min})}{x_{max} - x_{min}}$$

Posteriormente a esto la base de datos está completa y lista para la siguiente etapa de elaboración de los modelos de predicción. De todos modos, durante la etapa de elaboración de los modelos se continúo iterando la base de datos hasta encontrar la óptima que otorgaba los mejores resultados.

3.8. Modelos de predicción

Para la formulación de los modelos de predicción de riesgo de depresión se basó tanto en la bibliografía recopilada, resultado de los distintos test estadísticos realizados y la opinión de profesionales del área de la salud. En la siguiente sección se muestran los modelos más efectivos y relevantes del proceso de investigación además las variables implicadas junto con los resultados obtenidos.

Por último, dejar claro que para evaluar el rendimiento de los modelos nos basaremos principalmente en la métrica de Recall o sensibilidad, según recomendaciones del equipo médico y objetivo del proyecto que pretende superar la capacidad de detectar este tipo de patologías, dado que a opinión del experto las consecuencias de un mal diagnóstico son menores que el no detectar el trastorno.

3.8.1. Modelo basado en variables de riesgo población chilena

El modelo realizado a continuación se basó en las variables de riesgo de un modelo de predicción de riesgo de depresión desarrollado en España que se ajusto a la realidad chilena [71], en este estudio se tomaron a 2.832 pacientes de un centro de atención primaria de la ciudad de Concepción, donde se examinaron 39 variables posibles de riesgo de depresión y mediante un modelo de regresión logística se logró una eficiencia del 74,6 % según el indicador C-Index (símil de AUC en regresiones logísticas), esta base de datos contaba con 2.133 pacientes sin depresión y el modelo fue validado y comparado con el modelo español. Basado en el estudio anterior, los resultados de los análisis de variables de la sección anterior y el análisis experto, las variables para el modelo fueron las siguientes:

Variables	Etiqueta
EDAD	EDAD
NEDU	Nivel Educación
SEXO	SEXO
cd2	En general Ud. diría que su salud es:
sd27	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de depresión?
DataF1_2p3	Cuando tiene problemas, ¿tiene Ud. alguna persona en quien confiar, pedir ayuda o consejo?

Tabla 3.3: Variables de riesgo población chilena

Cada variable del modelo responde a ciertos factores de riesgo de depresión que se detallan a continuación.

- **Edad**: Esta variable es utilizada principalmente a la correlación mostrada en los test estadísticos(tablas A.3 y A.4) que demuestran la asociación con la variable dependiente. Además, en la bibliografía recopilada se utiliza debido a su fácil acceso y su poder predictivo.

- **NEDU**: El nivel educativo según la bibliografía recopilada y los test estadísticos realizados (tablas A.3 y A.4) tiene una relevancia significativa con respecto a la variable depresión, principalmente a que un a mejor nivel educativo se tiende a tener un mejor estatus de vida y más recursos cognitivos para sobreponerse a las crisis.
- **Sexo**: Dado a la diferente bibliografía recopilada que demuestra que el sexo es una de las variables socio-demográficas más vinculantes con la depresión y que existe una diferencia significativa entre hombres y mujeres en la probabilidad de tener dicho trastorno, además en el estudio de variables realizado (tablas A.3 y A.4) obtuvo valores que ratificaban su dependencia con la depresión.
- **cd2**: La percepción sobre el estado de salud, es relevante dado que la vejez está asociado con un deterioro de ciertas habilidades funcionales y cognitivas que si no se manejas de manera correcta o no se tratan pueden desencadenar en un trastorno depresivo.
- **sd27**: Los antecedentes de anteriores episodios depresivos son un buen referente de riesgo dado que existe una gran posibilidad que una persona que ya haya sufrido un episodio depresivo vuelva a caer en tales trastornos, debido a los pensamientos negativos y auto-compasivos que a generado a lo largo de los episodios anteriores.
- **DataF1_2p3**: El aislamiento social y el no contar con apoyo psicológico por parte de un ser querido o familiar cercano es un serio gatillante de pensamientos que pueden conducir a episodios depresivos.

Al conjunto de estas variables se les aplico un test de correlación para descartar variables que tuvieran una correlación superior a |0,4| y así prevenir la aparición de sobre-ajuste en los modelos de Random Forest y en Arboles de decisión, esto se muestra en la figura 3.9.

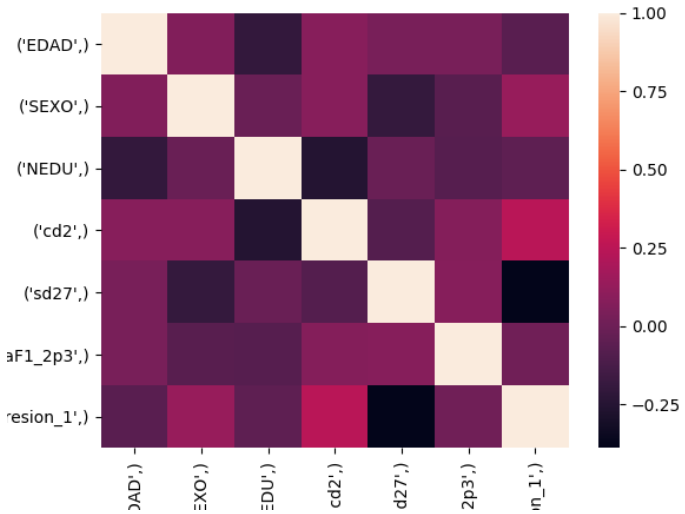
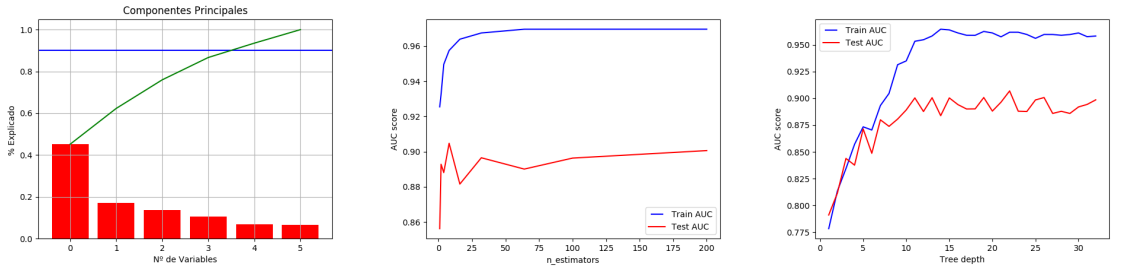


Figura 3.9: Correlación de variables del modelo.
Fuente: Elaboración Propia

Con este mismo propósito se realizó un test para determinar el número mínimo de árboles de decisión que debiesen existir en la metodología de Random Forest y así alcanzar su máximo desempeño y un test de profundidad (Max Depth) para determinar la cantidad de nodos que

debiese tener tanto Random Forest como en Arboles de Decisión para no caer en sobre-ajuste (ver figura A.19). Por último, se realizo un análisis de componentes principales para dictaminar si el número de variables que son requeridas para representar de mejor manera la muestra (ver figura A.19).



(a) Resultado PCA (b) Test N Estimadores (c) Test Máxima Profundidad

Figura 3.10: Resultados modelo con variables riesgo población.
Fuente: Elaboración Propia

Cross Validation

Los resultados expuestos corresponden a un *Cross Validation* en la que se divide la base de datos en K secciones y luego se procede a entrenar el modelo con K-1 secciones y se valida con la sección restante,esto se repite para cada una de las secciones hasta que todas hayan sido alguna vez la muestra de validación, posteriormente se obtiene el promedio de cada uno de los indicadores de rendimiento junto con su desviación estándar. En este caso particular se utilizó K=10 para la validación y lo que entrego los resultados expuestos en la tabla 3.4.

Modelo	Accuracy	Precision	Recall	AUC	F1
LOGIT	0.7878	0.8063	0.7622	0.8823	0.7822
KNN	0.8452	0.8004	0.9235	0.9192	0.8566
TREE	0.7697	0.7245	0.8778	0.8489	0.7911
NB	0.7830	0.7667	0.8158	0.8654	0.7891
SVM	0.7941	0.8294	0.7432	0.8900	0.7830
FOREST	0.8027	0.8029	0.8564	0.9040	0.8185

Tabla 3.4: Cross validation modelo variables de riesgo de depresión

Los resultados expuestos en la tabla anterior se obtuvieron mediante la técnica de balance de base de datos llamada **Nearest Neighbours**, lo que solo deja las muestras de la clase mayoritaria (las que no tienen la etiqueta de depresión) que tengan relación con su vecindario de muestras, con este balance de datos el modelo obtuvo sus mejores resultados en la etapa de validación. Teniendo en consideración que el Recall o sensibilidad es el indicador más relevante para la evaluación de los modelos, se pudo ver que, en rendimiento general, tomando en cuenta también AUC que tanto Random Forest y K-Nearest Neightbor lograron buenos niveles de predicción. Estos resultados sobresalientes se pueden entender a lo pequeña de la base de

datos y además de ser una primera etapa de entrenamiento. De todas formas, sin importar el modelo utilizado de Machine Learning el modelo mostró en una primera fase tener un gran rendimiento.

Validación ENS 2009-2010

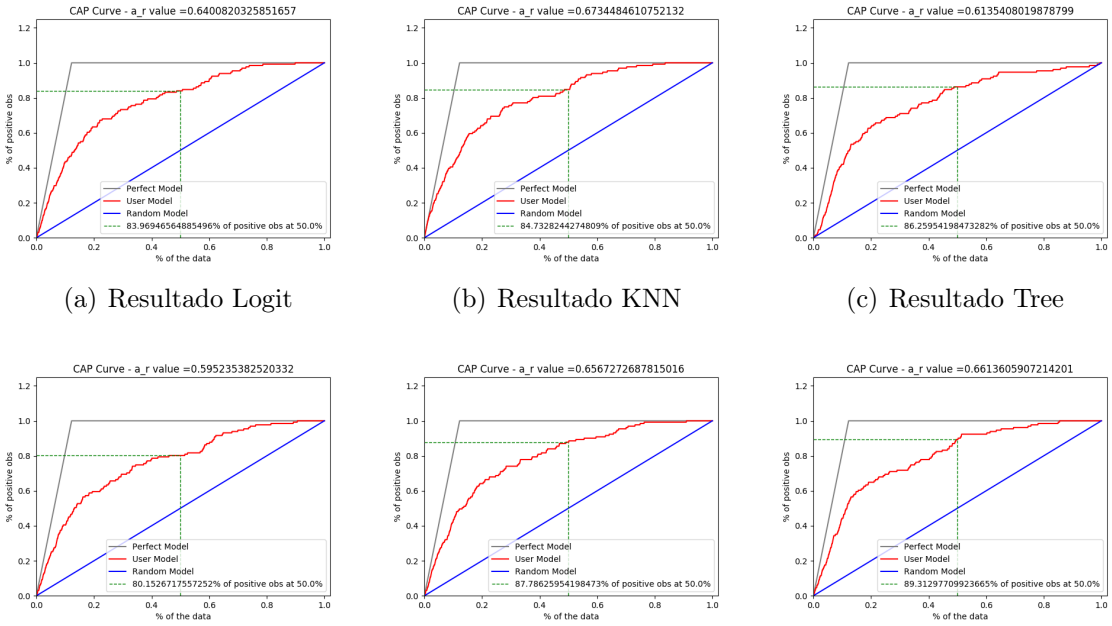
Para esta la validación con la Encuesta Nacional de Salud del 2009-2010 se consideró la base sin ningún tipo de modificación, Para evaluar el modelo previamente entrenado se consideró un nuevo indicador llamado log-loss o perdida de registro que entre menor sea su valor es mejor el poder predictivo de modelo, este indicador se utiliza especialmente en bases de datos desbalanceadas debido a que penaliza fuertemente a los falsos positivos y falsos negativos por sobre otros indicadores como Recall, Accuracy y Precision. Los resultados obtenidos de la validación se detallan en la tabla 3.6

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.865	0.527	0.454	0.488	0.719	0.308
KNN	0.867	0.412	0.454	0.432	0.672	0.323
TREE	0.864	0.626	0.458	0.529	0.761	0.310
NB	0.849	0.618	0.420	0.500	0.750	0.319
SVM	0.853	0.580	0.427	0.492	0.736	0.320
FOREST	0.872	0.634	0.483	0.548	0.769	0.305

Tabla 3.5: Validación modelo variables de riesgo ENS 2009-2010

Según los resultados expuestos se vuelve a apreciar que al igual que en la etapa de entrenamiento el modelo Random Forest como método de Machine Learning donde los resultados de AUC y perdida de registro tienen sus mejores resultados, también se puede observar una pérdida de Recall y de Precision en todos los modelos y una baja sustancial en el poder de predicción de KNN, demostrando su baja sensibilidad o la parecencia de sobre-ajuste. Logit y Arboles de Decisión por su lado a demostrado ser un modelo más flexible que en la etapa de entrenamiento y a pesar de sufrir cambios no fue tan drástico como en otros modelos.

Además, se consideró una nueva métrica para evaluar el modelo llamado Cummulative Accuracy Profiles(CAP), esta métrica estudia la capacidad de un modelo de predecir, en el eje y se puede ver el nivel predictivo del modelo y en eje x se ve la cantidad de datos con los que dispone para hacer la predicción, un modelo efectivo tiene al 50 % de datos, una capacidad de predicción mayor a 70 % y un excelente por sobre 80 %. A continuación, se presentan los resultados de los distintos modelos.



(a) Resultado Logit (b) Resultado KNN (c) Resultado Tree

(d) Resultado NB (e) Resultado SVM (f) Resultado Forest

Figura 3.11: Gráficos CAP modelo con variables riesgo población.
Fuente: Elaboración Propia

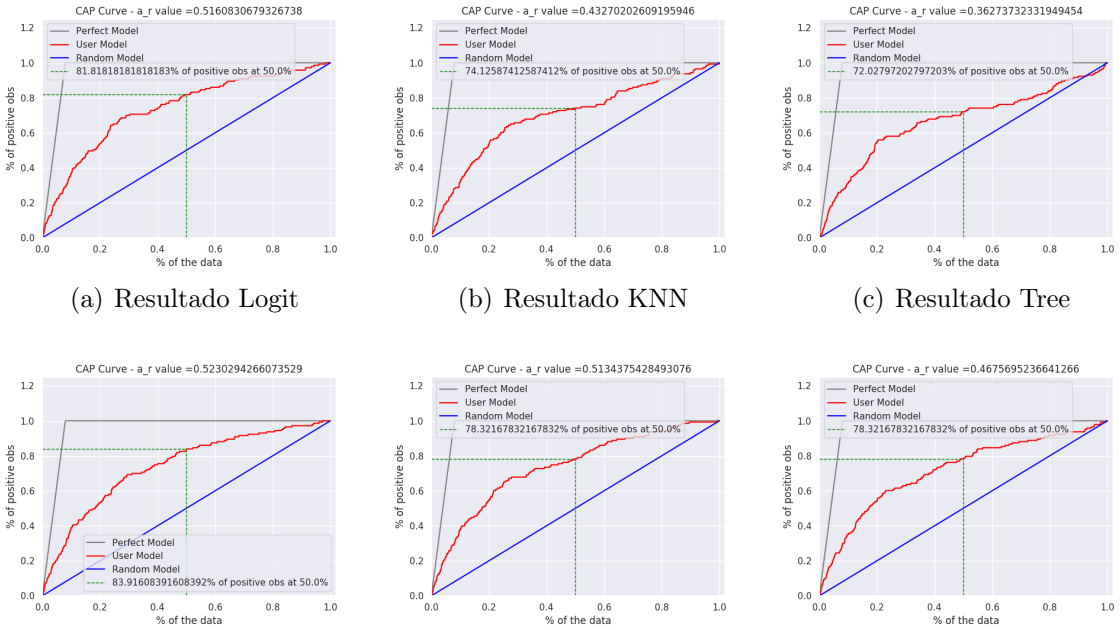
Según la figuras 3.12 se confirma lo que arrojaba la tabla de resultados anterior 3.6 donde Random Forest obtiene los buenos resultados, entregando un modelo con un alto nivel predictivo. Aunque también cabe considerar que los modelos Árbol de Decisión,KNN y Logit también tuvieron un gran rendimiento bajo las mismas métricas.

Validación ENS 2016-2017

Para la validación del modelo en base a los datos obtenidos de la Encuesta Nacional de Salud de los años 2016-2017 se realizó el cruce entre variables, y con ello se procedió a testear el modelo, estos resultados se plasman en los siguientes tablas 3.6 y gráficos 3.12.

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.888	0.294	0.296	0.295	0.617	0.276
KNN	0.867	0.350	0.256	0.296	0.631	0.304
TREE	0.910	0.175	0.362	0.236	0.574	0.268
NB	0.815	0.510	0.219	0.306	0.676	0.330
SVM	0.860	0.434	0.266	0.330	0.665	0.294
FOREST	0.837	0.483	0.240	0.320	0.675	0.302

Tabla 3.6: Validación modelo variables de riesgo ENS 2016-2017



(a) Resultado Logit (b) Resultado KNN (c) Resultado Tree

(d) Resultado NB (e) Resultado SVM (f) Resultado Forest

Figura 3.12: Gráficos CAP mdelo con variables riesgo población ENS 2016-2017.
Fuente: Elaboración Propia

Como se pudo plasmar los resultados obtenidos respecto a los obtenidos con la encuesta Nacional de Salud del año 2009-2010 fueron sustancialmente menores en las métricas generales, donde las métricas como sensibilidad se redujeron en la mayoría de los modelos incluso llegando a la mitad de lo que tenían en la primera validación. La única métrica que mantuvo su nivel predictivo fue Log-Loss, incluso mejorando en la mayoría de los casos, se debe tener en consideración que la ENS 2016-2017 tiene una menor proporción de casos de depresión respecto a la ENS 2009-2010 y debido a esto probablemente el indicador Log-Loss no se vio mayormente perjudicado.

En cuanto a poder predictivo de los modelos medido mediante Cummulative Accuracy Profiles demostró disminuir significativamente en todos los modelos en especial en Random Forest cuyo buen desempeño disminuyo tanto más de un punto en el CAP. Por otro lado los modelos KNN y Naive Bayes a pesar de disminuir su poder predictivo, su perdida fue menor al resto de los modelos.

Por ultimo cabe destacar que si consideramos los resultados obtenidos en la validación con la Encuesta Nacional de Salud del año 2009-2010 y los del estudio realizado en Chile por el centro médico de Concepción, en el cual se basó el modelo que español, los resultados son bastante similares aunque a pesar de que no se utilizaron exactamente las mismas variables y este modelo solo contaba con 6 variables respecto a las 8 que cuenta el modelo original, Random forest mostró tener un mejor rendimiento que el modelo original y deja la ventana abierta a utilizar técnicas de Machine Learning para este tipo de proyectos de salud en vez de los típicos modelos a base de regresiones logísticas.

3.8.2. Modelo basado en variables de determinante social

Este modelo emerge dado a los buenos resultados generales que tuvieron las variables sobre la determinantes sociales y psicológicas en los test realizados en la sección anterior y la bibliografía que enfatiza como estas variables son buenos indicadores de riesgo de depresión, especialmente porque entregan un buen marco de referencia del contexto en el que se puede encontrar un adulto mayor [87]. Además, estas variables recogen la percepción personal sobre el entorno que como se vio en los test anteriormente realizados tienen una gran influencia en el riesgo que una persona pudiese tener depresión, adicionalmente entregan una buena intuición sobre el apoyo social que puede estar recibiendo esa persona junto con su nivel de aislamiento.

Variables	Etiqueta
dataF1_2p1	¿Cuán de acuerdo está Ud. con la siguiente frase: "A nadie le importa mucho lo que me pasa"?.
p2a	En general, ¿cuánto confía usted en la gente de su villa, barrio o población? (Califique su percepción de confianza de 1 a 7, considerando que 1 es nada de confianza y 7 mucha confianza)
p4a	Si se le cayera su monedero o billetera en su barrio, calle, villa o población y alguien la viera, piensa ¿Ud. que él o ella se la devolvería?
dataF1_2p6	¿Con qué frecuencia se ha sentido estresado durante los últimos 2 meses, es decir, irritable, con ansiedad o sin poder dormir, debido a situaciones en la casa o en el trabajo?
dataF1_2p7	¿Qué nivel de estrés financiero, económico, de dinero o plata ha sentido Ud en los últimos 12 meses?
p8e	p8e Ha tenido algún problema grande en la familia?
p8f	p8f Ha tenido algún problema serio de salud o accidente?
p8h	p8h Se le ha enfermado o muerto alguien cercano de la familia?

Tabla 3.7: Variables modelo de riesgo adulto mayor

Cada variable del modelo responde a ciertos factores de riesgo de depresión que se detallan a continuación.

- **dataF1_2p1**: Pregunta relacionada a determinar el grado de aislamiento social que pudiese tener un individuo, esta variable fue considerada también por sus buenos resultados en el test de Kolmogorov-Smirnov(anexos A.11).
- **p2a**: El entorno social es una factor preponderante y de riesgo para los adultos mayores según el test de Chi-cuadrado(anexo A.10), esta variable se relaciona con el nivel de seguridad y cercanía con los vecinos de un villa y una mala relacion pudiese desencadenar en problemas mayores entre ellos la depresión.
- **p4a**: Pregunta relacionada a determinar el grado de aislamiento social que pudiese tener un individuo, esta variable fue considerada también por sus buenos resultados en el test de Kolmogorov-Smirnov(anexos A.11).
- **dataF1_2p6**: La ansiedad, estrés y problemas de sueño se encuentran profundamente relacionados con la depresión, y son factores de riesgo que pudiesen desencadenar si no

se controla en problemas depresivos. Esta variable tuvo una gran relacion según los test de Chi-Cuadrado y de Kolmogorov-Smirnov(Anexo A.10 y A.11)

- **dataF1_2p7**: Los problemas económicos son frecuentes dentro de la población adulto mayor en Chile. Especialmente justificado por los problemas de salud ,los costos asociados a remedios y las bajas pensiones, según el estudio de variables anteriormente realizado es una de las más relevantes de las determinantes sociales (Anexo A.10 y A.11).
- **p8e**:Los problemas en el entorno social más cercano para los adultos mayores puede desencadenar en problemas depresivos si estos se vinculan a maltrato, descalificaciones o un sentimiento de culpa.
- **p8f**: La vejez está vinculada a un deterioramiento en fisiológico que pudiese desencadenar en enfermedades, estas enfermedades si no son bien manejados pueden conllevar problemas de depresión.
- **p8h**: La muerte de una persona cercana puede ser una experiencia traumáticas en el caso de ser una persona extremadamente relevante dentro del círculo social de una persona, esto generalmente lleva a tener episodios depresivos pero si este se alimenta de otros pensamientos depresivos pudiese desencadenar en un problema mayor.

Al conjunto de estas variables se les aplico un test de correlación para descartar variables que tuvieran una correlación superior a |0,4| y así prevenir la aparición de sobre-ajuste en los modelos de Random Forest y en Arboles de decisión, esto se muestra en la figura 3.13.

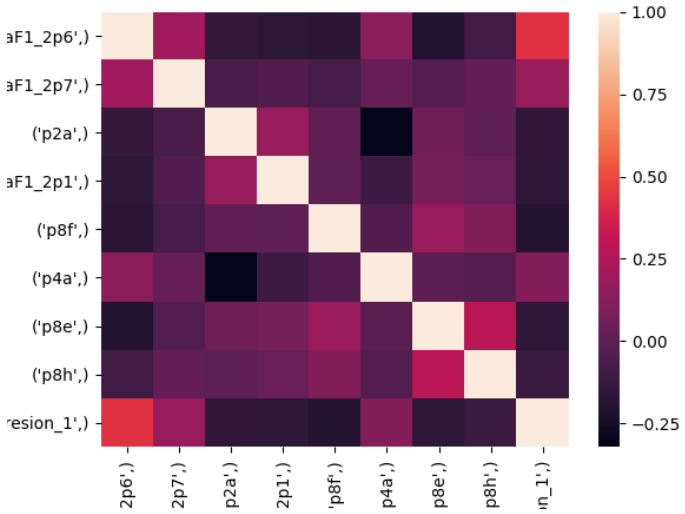
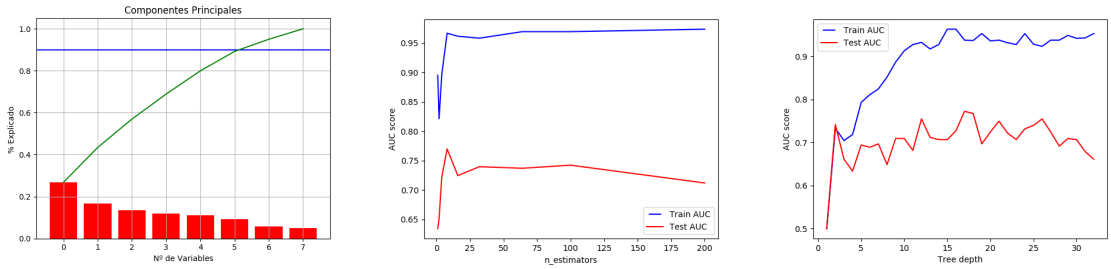


Figura 3.13: Correlación de variables de determinante social.
Fuente: Elaboración Propia

Con este mismo propósito se realizó un test para determinar el número mínimo de árboles de decisión que debiesen existir en la metodología de Random Forest y así alcanzar su máximo desempeño y un test de profundidad (Max Depth) para determinar la cantidad de nodos que debiese tener tanto Random Forest como en Arboles de Decisión para no caer en sobre-ajuste (ver figura 3.14). Por último, se realizó un análisis de componentes principales para

dictaminar si el número de variables que son requeridas para representar de mejor manera la muestra (ver figura 3.14).



(a) Resultado PCA (b) Test N Estimadores (c) Test Máxima Profundidad

Figura 3.14: Resultados modelo en variables de determinante social.
Fuente: Elaboración Propia

Cross Validation

Los resultados expuestos corresponden a un *Cross Validation* en la que se divide la base de datos en K secciones y luego se procede a entrenar el modelo con K-1 secciones y se valida con la sección restante,esto se repite para cada una de las secciones hasta que todas hayan sido alguna vez la muestra de validación, posteriormente se obtiene el promedio de cada uno de los indicadores de rendimiento junto con su desviación estándar, en este caso particular se utilizó K=10 para la validación y lo que entrego los resultados expuestos en la tabla 3.8.

Modelo	Accuracy	Precision	Recall	AUC	F1
LOGIT	0.9069	0.8626	0.4442	0.8800	0.5762
KNN	0.8946	1.0000	0.2795	0.8524	0.4283
TREE	0.8878	0.7266	0.3758	0.7973	0.4839
NB	0.8911	0.6332	0.6513	0.8691	0.6306
SVM	0.9013	0.9472	0.3448	0.8840	0.4910
FOREST	0.8968	0.8621	0.2988	0.8550	0.4236

Tabla 3.8: Cross Validation modelo variables de determinante social

Estos resultados que se detallan la tabla anterior usando la técnica de balance de base de datos llamada **Nearest Neighbours**, lo que solo deja de la muestras mayoritarias (las que no tienen la etiqueta de depresión) que tengan relación con su vecindario de muestras más cercano. Con esta metodología se obtuvieron los mejores resultados expuestos en la tabla A.68. Aun teniendo en consideración que Recall o sensibilidad es el indicador más relevante para evaluar cada uno de los modelos, en la fase de entrenamiento se destaca el modelo de Naive Bayes con un Recall del 0,65 y un AUC de 0,86 que son los mejores entre todos los modelos, aunque al mismo tiempo tiene la Precisión más baja de todas.

Validación ENS 2009-2010

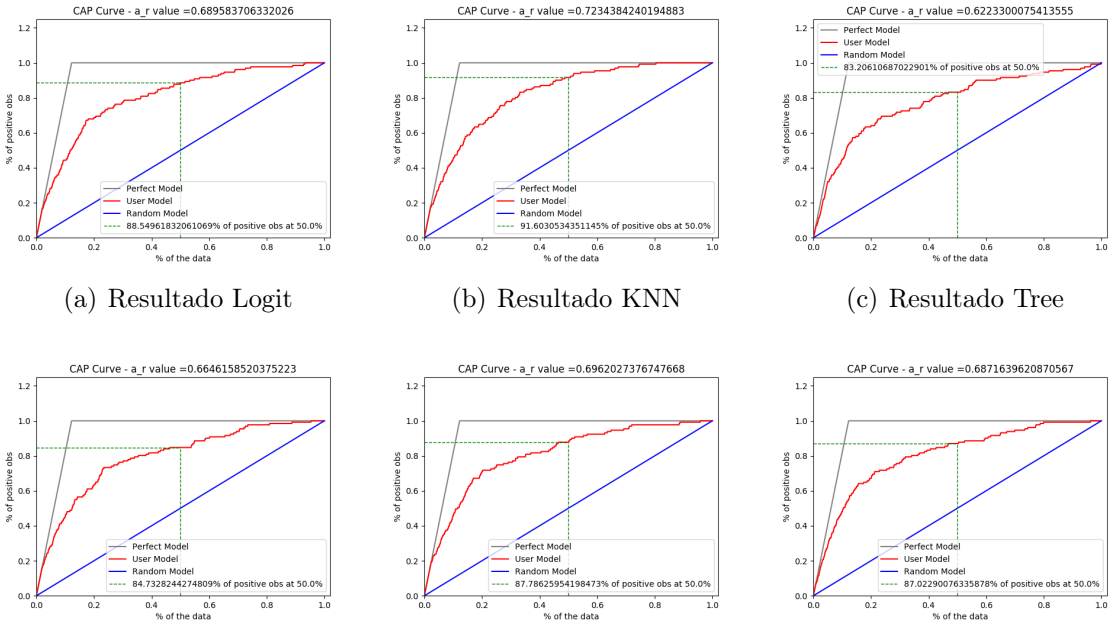
Para la validación usando Encuesta nacional de Salud del 2009-2010 sin modificar y con el modelo previamente entrenado se consideró un nuevo indicador llamado Log-Loss o perdida de registro que entre menor sea su valor es mejor el poder predictivo de modelo, este indicador se utiliza especialmente en bases de datos desbalanceadas debido a que penaliza fuertemente a los falsos positivos y falsos negativos por sobre otros indicadores como son Recall, Acurracy y Precision. Los resultados obtenidos se detallan en la tabla 3.12

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.879	0.504	0.504	0.504	0.717	0.2910
KNN	0.889	0.443	0.558	0.494	0.697	0.2848
TREE	0.891	0.443	0.569	0.498	0.698	0.2960
NB	0.853	0.565	0.425	0.485	0.729	0.3117
SVM	0.881	0.527	0.515	0.521	0.729	0.2885
FOREST	0.889	0.511	0.549	0.530	0.726	0.2876

Tabla 3.9: Validación modelo variables de determinante social ENS 2009-2010

Los resultados expuestos anteriormente demuestra que Naive Bayes sigue siendo el mejor modelo en términos de AUC pero sufre un decaimiento entre la fase de entrenamiento y de validación, al contrario el modelo de Random Forest que aumenta su nivel en términos de Recall y AUC , inclusive tiene el mejor rendimiento a nivel de Log-Loss, una mención también a Arboles de Decisión y KNN que tuvieron un gran nivel de rendimiento a nivel de Recall y Log-Loss.

Además, se consideró una nueva métrica para evaluar el modelo llamado Cummulative Accuracy Profiles(CAP), esta métrica estudia la capacidad de un modelo de predecir, en el eje y se puede ver el nivel predictivo del modelo y en eje x se ve la cantidad de datos con los que dispone para hacer la predicción, un modelo efectivo tiene al 50 % de datos, una capacidad de predicción mayor a 70 % y un excelente por sobre 80 %. A continuación, se presentan los resultados de los distintos modelos.



(a) Resultado Logit (b) Resultado KNN (c) Resultado Tree

(d) Resultado NB (e) Resultado SVM (f) Resultado Forest

Figura 3.15: Resultados modelo de determinante social ENS 2009-2010
Fuente: Elaboración Propia

Cumulative Accuracy Profiles demostró que el modelo KNN obtuvo los mejor resultado con un AR=0.72, esta métrica arrojo también un buen rendimiento por parte de Support Vector Machine con un AR=0.69 , Random Forest aparece en tercer lugar con un AR=0.68 lo que lo convierte en un modelo con alta capacidad de predicción.

Validación ENS 2016-2017

Para este caso particular no se pudo validar con la base de datos de la encuesta nacional de salud del 2016-2017 porque especialmente no todas las variables se encontraban en ambas encuestas, por lo tanto, para no perder rigurosidad metodológica se decidió no realizar la validación con esta base de datos.

Finalmente se deduce de este análisis que el modelo basado en variables de determinantes sociales y psicológicas obtuvo su mejor modelo mediante la técnica de Naive Bayes, pero dado a que nuestro objetivo es maximizar la capacidad de detección de los trastornos el mejor modelo en ese caso fue Random Forest (a pesar que Arboles de decisión tuvo mejor Recall) ya que en términos medios como son AUC y Log-Loss tuvo un gran rendimiento.

3.8.3. Modelo basado en el estudio de variables de riesgo

El modelo realizado se basa especialmente en los resultados arrojados por los distintos test realizados en las secciones anteriores, donde se escogieron las áreas con variables que demostraron mediante los test de Chi-Cuadrado y de Kolmogorov-Smirnov un gran poder predictivo y por otro lado también se eligieron aquellas que fueran según la bibliografía factores determinantes para diagnosticar depresión en los adultos mayores. Las áreas seleccionadas fueron:

- Trastornos del sueño
- Percepción de salud
- Determinantes sociales y psicologicos
- Antecedentes depresivos

Luego se itero con múltiples combinaciones de variables y se buscó diseñar de tal manera de utilizar un número mínimo de preguntas que entregar la mayor asertividad y precisión al momento de evaluar. Las variables resultantes se exponen en la tabla 3.10

Variables	Etiqueta
cd20	¿En general, durante los últimos 30 días, ¿qué grado de molestia o dolor ha tenido??
cd31	En general, durante los últimos 30 días, ¿en qué grado se ha sentido triste, decaído o deprimido?
sd27	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de depresión?
ts4	¿Despierta usted sintiéndose cansado(a) o casi tan cansado como antes de dormir, por lo menos tres días a la semana?

Tabla 3.10: Variables modelo de riesgo adulto mayor

Cada variable del modelo responde a ciertos factores de riesgo de depresión que se detallan a continuación.

- **cd20**: Esta variable fue escogida dado a su gran asociación con la variable depresión según muestran las tabla A.14 y A.15 , además según el estudio de las variables de dolor han demostrado su relación con la depresión.
- **cd31**: Un estado animo triste o decaído es uno de las principales consecuencias de la depresión y se encuentra profundamente correlacionado. Por lo tanto, una conducta por más de 30 días presenta ser una variable de riesgo.
- **sd27**: Los antecedentes de anteriores episodios depresivos son un buen referente de riesgo dado que existe una gran posibilidad que una persona que ya haya sufrido un episodio depresivo vuelva a caer en tales trastornos, debido a los pensamientos negativos y auto-compasivos que a generado a lo largo del los episodios anteriores.
- **ts4**: Los trastornos del sueño son una de las principales consecuencias e indicios de que una persona está sufriendo de depresión, y según los registros.

Al conjunto de estas variables se les aplico un test de correlación para descartar variables que tuvieran una correlación superior a |0,4| y así prevenir la aparición de sobre-ajuste en

los modelos de Random Forest y en Arboles de decisión, esto se muestra en la figura 3.9.

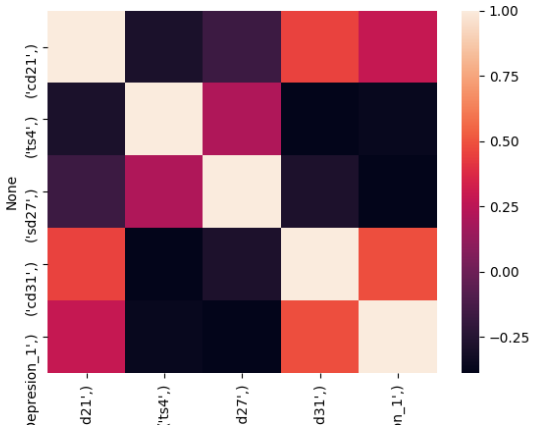
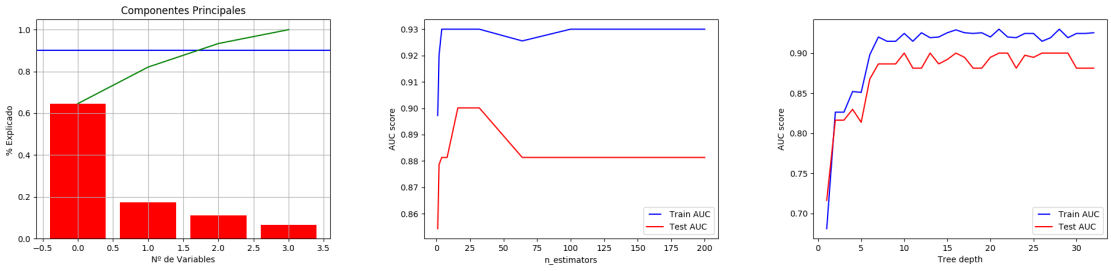


Figura 3.16: Correlación de riesgo adulto mayor
Fuente: Elaboración Propia

Con este mismo propósito anterior se realizó un test Numero Arboles de Decisión para determinar el número mínimo de árboles de decisión que debiesen existir en la metodología de Random Forest para alcanzar su máximo desempeño y un test de máxima profundidad (Max Depth) para determinar la cantidad de nodos que pudiese tener tanto Random Forest como en Arboles de Decisión (ver figura 3.17)

Por último, se realizó un análisis de componentes principales para dictaminar si el número de variables es el requerido para representar de mejor manera la muestra (ver figura 3.17).



(a) Resultado PCA (b) Test N Estimadores (c) Test Máxima Profundidad

Figura 3.17: Resultados modelo con variables riesgo población.
Fuente: Elaboración Propia

Cross Validation

Los resultados expuestos corresponden a un *Cross Valiation* en la que se divide la base de datos en K secciones y luego se procede a entrenar el modelo con K-*1* secciones y se valida con la sección restante,esto se repite para cada una de las secciones hasta que todas hayan

sido alguna vez la muestra de validación, posteriormente se obtiene el promedio de cada uno de los indicadores de rendimiento junto con su desviación estándar, en este caso particular se utilizó K=10 para la validación y lo que entrego los resultados expuestos en la tabla 3.11.

Modelo	Accuracy	Precision	Recall	AUC	F1
LOGIT	0.9314	0.8514	0.6451	0.9513	0.7273
KNN	0.9336	0.9576	0.5698	0.9495	0.7053
DTREE	0.9402	0.8502	0.7111	0.8848	0.7669
NB	0.9225	0.7019	0.8193	0.9522	0.7490
SVM	0.9347	0.8508	0.6624	0.9507	0.7398
FOREST	0.9447	0.8993	0.7040	0.9475	0.7667

Tabla 3.11: Cross Validation modelo variables de riesgo adulto mayor

Estos resultados que se detallan la tabla anterior usando la técnica de balance de base de datos llamada **Nearest Neighbours**, lo que solo deja de la muestras mayoritarias(las que no tienen la etiqueta de depresión) que tengan relación con su vecindario de muestras más cercano. Con esta metodología se obtuvieron los mejores resultados expuestos en la tabla 3.11. Aun teniendo en consideración que Recall o sensibilidad es el indicador más relevante para evaluar cada uno de los modelos,los resultados generales usando estas variables fueron muy buenos, en la fase de entrenamiento donde Naive Bayes destaco por su alto nivel en Recall y AUC. En cambio, KNN y Random forest destacaron por su alto nivel de precisión.

Validación ENS 2009-2010

Para la validación usando Encuesta nacional de Salud del 2009-2010 sin modificar y con el modelo previamente entrenado se consideró un nuevo indicador llamado log-loss o perdida de registro que entre menor sea su valor es mejor el poder predictivo de modelo, este indicador se utiliza especialmente en bases de datos desbalanceadas debido a que penaliza fuertemente a los falsos positivos y falsos negativos por sobre otros indicadores como son Recall, Acurracy y Precision. Los resultados obtenidos se detallan en la tabla 3.12.

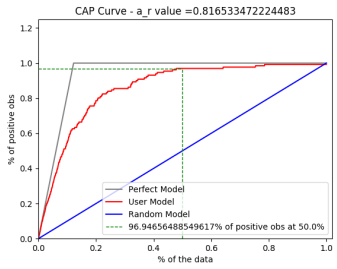
Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.890	0.695	0.538	0.607	0.806	0.267
KNN	0.888	0.664	0.534	0.592	0.792	0.288
TREE	0.895	0.740	0.554	0.634	0.829	0.283
NB	0.871	0.756	0.483	0.589	0.821	0.277
SVM	0.893	0.679	0.549	0.608	0.801	0.278
FOREST	0.899	0.695	0.572	0.628	0.811	0.304

Tabla 3.12: Validación modelo variables de riesgo adulto mayor ENS 2009-2010

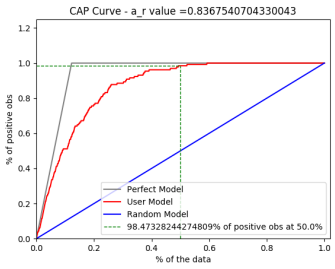
En esta etapa de validación vuelve a pasar que Random Forest es el que tuvo los mejores resultados en términos de Recall en cambio no tuvo mejor AUC o Log-Loss que un modelo

Logit o arboles de decisión, Arbol de Decisión por otro lado tuvo un gran desempeño en términos de AUC, de todos modos, a nivel general todos los modelos tuvieron rendimientos similares que con el uso de estas variables

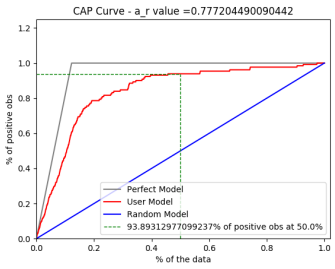
Además, se consideró una nueva métrica para evaluar el modelo llamado Cummulative Accuracy Profiles(CAP), esta métrica estudia la capacidad de un modelo de predecir, en el eje y se puede ver el nivel predictivo del modelo y en eje x se ve la cantidad de datos con los que dispone para hacer la predicción, un modelo efectivo tiene al 50 % de datos, una capacidad de predicción mayor a 70 % y un excelente por sobre 80 %. A continuación, se presentan los resultados de los distintos modelos.



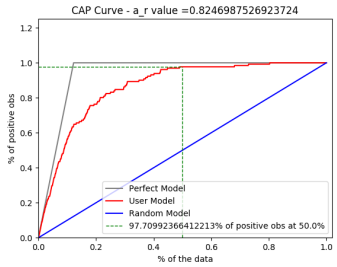
(a) Resultado Logit



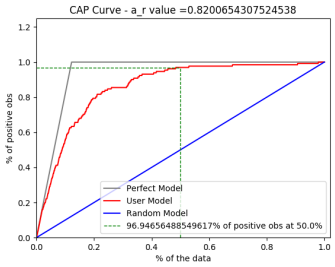
(b) Resultado KNN



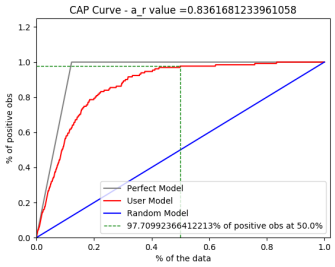
(c) Resultado Tree



(d) Resultado NB



(e) Resultado SVM



(f) Resultado Forest

Figura 3.18: Resultados Modelo de riesgo adulto mayor
Fuente: Elaboración Propia

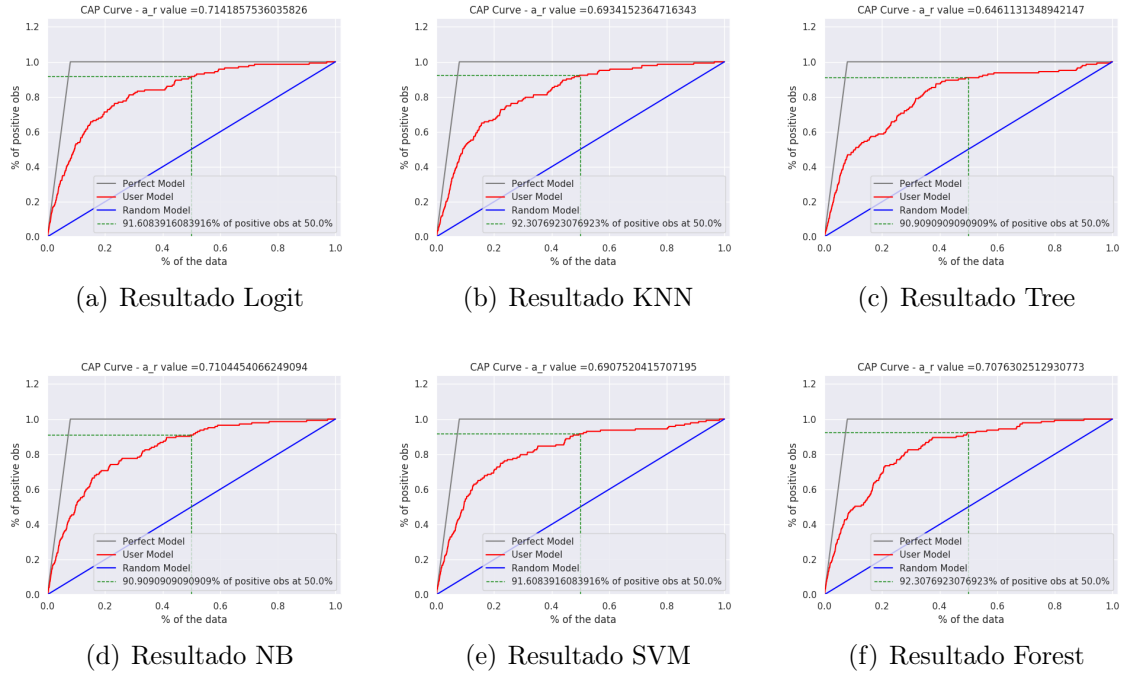
En cuanto a la evaluación mediante Cumullative Acuracy profiles determino que la mayoría de estos modelos tuvieron un rendimiento similar, y no tuvieron mayor diferencia en su poder predictivo a excepción de árboles de decisión. El mejor rendimiento general aun así se lo lleva el modelo de Random Forest.

Validación ENS 2016-2017

Para la validación del modelo en base a los datos obtenidos de la Encuesta Nacional de Salud de los años 2016-2017 se realizo el cruce entre variables, y con ello se procedió a testear el modelo, estos resultados se plasman en los siguientes tablas 3.13 y gráficos 3.19.

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.870	0.657	0.337	0.445	0.773	0.323
KNN	0.861	0.650	0.317	0.427	0.765	0.336
TREE	0.898	0.490	0.389	0.433	0.711	0.297
NB	0.832	0.706	0.280	0.401	0.774	0.326
SVM	0.876	0.650	0.350	0.455	0.773	0.330
FOREST	0.863	0.524	0.296	0.379	0.708	0.326

Tabla 3.13: Validación modelo variables de riesgo adulto mayor ENS 2016-2017



(a) Resultado Logit (b) Resultado KNN (c) Resultado Tree

(d) Resultado NB (e) Resultado SVM (f) Resultado Forest

Figura 3.19: Resultados Modelo de riesgo adulto mayor

Fuente: Elaboración Propia

Como se ha extendido en toda la evaluación de este modelo todos los resultados han sido bastante parejos, este modelo a diferencia de los anteriores tiene un mejor nivel predictivo , aunque empeora respecto a la anterior validación , existe una importante relación entre los resultados obtenidos y la forma de la base de datos.

3.9. Otros modelos interesantes

En la búsqueda de encontrar el mejor modelo que predijera la depresión en adultos mayores de una manera efectiva se tomaron distintas variables, entre estos sub-sets de variables nacieron múltiples modelos, algunos de estos modelos demostraron en particular tener un desempeño notable, aunque no mejor que los desarrollados previamente, pero si generaron

modelos que podían detectar la depresión en adultos mayores chilenos con cierta eficiencia. Otros modelos por lo contrario a pesar de contar con variables que según los test anteriores tenían cierto grado de relación lograron rendimientos mediocres o peores de lo esperado, a continuación se muestran algunos casos.

Modelos basado en variables socio-demograficas

Las variables socio-demograficas han sido testeadas multitud de veces tanto por la bibliografía como por los test de las secciones anteriores. A pesar de todo lo anterior este modelo no tuvo un desempeño comparable con los modelos de la sección anterior, ya que a pesar que contaba variables como el sexo y el nivel educativo no hacía referencia a ninguna variable que estuviera relacionado con el estado de animo de una persona lo que hacía que perdiera efectividad. su máxima capacidad en términos de AUC fue de 0,59 pero con un Recall de 0,19, por lo tanto, resulta ser un modelo bastante pobre(anexoA.66).

modelos de basados en variables asociadas a depresión en adulto mayor

Según un estudio que recalcaba los principales factores de depresión en la población adulta mayor en España, se tomaron la mayor cantidad de respuestas relacionadas a esos ámbitos y dejar aquellas que tuvieran la mayor relación con la depresión en adultos mayores [65].

Las variables utilizadas en esto fueron algunas socio-demográficas, percepción de la salud física y emocional, dolores, artritis y movilidad y por ultimo fracturas a lo largo de su vida. Este modelo tuvo un buen desempeño, pero su rendimiento no alcanzo uno tan bueno como los anteriormente expuestos, alcanzando como máximo un nivel máximo un AUC aproximado de 0.69 y un Recall de 0,47 cuando se validaba con la encuesta nacional de salud del 2009-2010(anexo A.69).

Modelo basado en sintomas de enfermedades crónicas

Dado a que las enfermedades crónicas son ampliamente disipadas por la población adulto mayor, se procedió a formalizar un modelo que contara con estas variables, aunque alcanzo un gran poder predictivo si se le vinculaba con ciertas variables emocionales como son el estado de animo de la persona y otras vinculadas a sus trastornos del sueño, cuando se aislaba solamente a variables asociadas a alguna enfermedad crónica su poder predictivo descendía drásticamente, así que no se consideró y se prefirió los expuestos anteriormente(anexoA.72).

3.10. Calculo de riesgo

Dado los modelos de Machine Learning y los diferentes sets de variables utilizadas, para cada uno de los modelos planteados en la sección anterior, se seleccionó un el algoritmo de

Machine Learning que mejor predijo depresión según las métricas. Para obtener el riesgo de depresión se utilizó comandos incluidos en los paquetes de SKlearn de Python que entrega la probabilidad de pertenecer a cada uno de los grupos dependiendo de las respuestas entregadas por los pacientes. A continuación, se muestran 2 ejemplos concretos de como predice riesgo cada uno de los modelos.

3.10.1. Modelo en base a estudios de depresión en chile

Según los resultados obtenidos en la sección anterior se postulo realizar el ejemplo con un modelo de **Random Forest**, dado a que presento el mejor Recall en la mayoría de los test.

Este modelo tiene con 6 preguntas que cuentan con las siguientes respuestas según la Encuesta Nacional de Salud 2009-2010 que se muestra a continuación en la tabla 3.14.

Etiqueta	Respuestas
EDAD	60-99
NEDU	<8 8-12 <12
SEXO	Hombre, Mujer
cd2	Excelente Muy buena Buena Regular Mala
sd27	Si, No
DataF1_2p3	Siempre Casi siempre Algunas veces Rara vez o nunca No lo necesita No sabe No responde

Tabla 3.14: Variables de riesgo población chilena

Para ello se supusieron las siguientes respuesta para cada uno de los sujetos y en la siguiente tabla se muestra la probabilidad de sufrir depresión de cada uno.

Etiqueta	Pregunta	Sujeto 1	Sujeto 2
EDAD	Edad	73	85
NEDU	Nivel Educación	<8	8-12
SEXO	Sexo	Hombre	Mujer
cd2	En general Ud. diría que su salud es:	Mala	Buena
sd27	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de depresión?	Si	No
DataF1_2p3	Cuando tiene problemas, ¿tiene Ud. alguna persona en quien confiar, pedir ayuda o consejo?	Siempre	Rara vez
%	**Riesgo de Depresión**	82,8 %	28,3 %

Tabla 3.15: Respuestas de modelo riesgo población chilena

3.10.2. Modelo en base a determinantes sociales y psicológicas

Según los resultados obtenidos en la sección anterior se postulo realizar el ejemplo con un modelo de **Naive Bayes**, dado sus buenos resultados en el test de Cross-Validation.

Este modelo tiene 9 preguntas que cuentan con las siguientes respuestas según la Encuesta Nacional de Salud 2009-2010 que se muestra a continuación en la tabla 3.16.

Etiqueta	Respuestas
dataF1_2p	Totalmente de acuerdo Más o menos de acuerdo Algo en desacuerdo Totalmente en desacuerdo No Sabe o No contesta
p2a	1-7
p4a	Muy de acuerdo De acuerdo Ni acuerdo ni en desacuerdo En desacuerdo Muy en desacuerdo
dataF1_2p6	Nunca Algunas veces Varias veces Permanentemente
dataF1_2p7	Poco o nada Moderado Alto o mucho No sabe/No responde
p8e	Si,No,No sabe
p8f	Si,No,No sabe
p8h	Si,No,No sabe

Tabla 3.16: Respuesta de modelo determinantes sociales y psicológicas

Para ello se supusieron las siguientes respuesta para cada uno de los sujetos y en la siguiente tabla se muestra la probabilidad de sufrir depresión de cada uno.

Etiqueta	Pregunta	Sujeto 1	Sujeto 2
dataF1_2p1	¿Cuán de acuerdo está Ud. con la siguiente frase: "A nadie le importa mucho lo que me pasa"?	Más o menos de acuerdo	Totalmente de acuerdo
p2a	En general, ¿cuánto confía usted en la gente de su villa, barrio o población? (Califique su percepción de confianza de 1 a 7, considerando que 1 es nada de confianza y 7 mucha confianza)	3	1
p4a	Si se le cayera su monedero o billetera en su barrio, calle, villa o población y alguien la viera, piensa ¿Ud. que él o ella se la devolvería?	De acuerdo	De acuerdo
dataF1_2p6	¿Con qué frecuencia se ha sentido estresado durante los últimos 2 meses, es decir, irritable, con ansiedad o sin poder dormir, debido a situaciones en la casa o en el trabajo?	Permanentemente	Nunca
dataF1_2p7	¿Qué nivel de estrés financiero, económico, de dinero o plata ha sentido Ud en los últimos 12 meses?	Poco o Nada	Moderado
p8e	p8e Ha tenido algún problema grande en la familia	Si	No
p8f	p8f Ha tenido algún problema serio de salud o accidente?	Si	No
p8h	p8h Se le ha enfermado o muerto alguien cercano de la familia?	Si	Si
%	**Riesgo de Depresión**	67,8 %	8,7 %

Tabla 3.17: Variables de modelo determinantes sociales y psicológicas

3.10.3. Modelo en basado en el estudio de varaibles de riesgo

Según los resultados obtenidos en la sección anterior se postulo realizar el ejemplo con un modelo de **Random Forest**. Dado a su buen funcionamiento según los test e indicadores.

Este modelo tiene 4 preguntas que cuentan con las siguientes respuestas según la Encuesta Nacional de Salud 2009-2010 que se muestra a continuación en la tabla 3.18.

Etiqueta	Respuestas
cd20	Ninguno Poco Moderado Mucho Demasiado
cd31	Ninguno Poco Moderado Mucho Demasiado
sd27	Si,No
ts4	Si,No

Tabla 3.18: Variables de modelo basado en el estudio de varaibles de riesgo

Para ello se supusieron las siguientes respuesta para cada uno de los sujetos y en la siguiente tabla se muestra la probabilidad de sufrir depresión de cada uno.

Etiqueta	Pregunta	Sujeto 1	Sujeto 2
cd20	¿En general, durante los últimos 30 días, ¿qué grado de molestia o dolor ha tenido??	Demasiado	Poco
cd31	En general, durante los últimos 30 días, ¿en qué grado se ha sentido triste, decaído o deprimido?	Demasiado	Ninguno
sd27	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de depresión?	Si	No
ts4	¿Despierta usted sintiéndose cansado(a) o casi tan cansado como antes de dormir, por lo menos tres días a la semana?	No	Si
%	**Riesgo de Depresión**	83,2 %	8,0 %

Tabla 3.19: Variables modelo basado en el estudio de varaibles de riesgo

3.11. Evaluación de impacto

Reducir los niveles de depresión es uno de los grandes desafíos que tiene Chile y el mundo para el siglo XXI, el impacto que pudiese tener un algoritmo que capaz predecir riesgo de depresión en la población no tan solo ayudaría en una primera etapa a detectarla si no que también mejorar la calidad de vida de aquellos que padecen este trastorno. Una de las razones por la cual la depresión se mantiene tan presente en la población es por la no identificación del problema y el tratamiento poco efectivo [71].Frente a este escenario es la capacidad de mejorar la eficiencia médica o de lo instrumentos existentes para detectar depresión es crucial si se quiere mejorar la salud mental de los chilenos.

En Chile dado el alto costo que implican la gran cantidad de licencias médicas [24] y los costos sociales para la población, optar por mecanismo que ayuden a los profesionales de la salud a detectar problemas de salud mental es fundamental. Para los adultos mayores el tema de la depresión y los costos asociados están muy relacionados con los costos en medicamentos y en el tratamiento de enfermedades, según el un estudio realizado en EEUU que busco determinar si los adultos mayores con síntomas depresivos incrementaban el costo general de los servicios médicos, determino después de un análisis de 5012 adultos mayores en un plazo de 4 años que los adultos mayores con síntomas depresivos gastan un 50 % más que los pacientes sin depresión[85].

Debido al amplio espectro que significa la depresión que no tan solo afecta el entorno social, si no hay estudios que revelan que afecta el tiempo de recuperación de adultos hospitalizados, acotaremos esta evaluación de impacto en el caso de los adultos mayores con depresión que acceden a un recinto asistencial por cualquier enfermedad y cuantificaremos las consecuencias de la no detección de este trastorno significa para los adultos mayores y para el recinto asistencial.

La distribución del gasto de la última hospitalización de los adultos mayores en Chile, mostró una media de $6.242.469 y una mediana de $ 3.112.627 [63], además conociendo el crecimiento anual en los costos médicos de 2,7 % [93] según la inflación anual de Chile.Además se considera los egresos hospitalarios de los adultos mayores para el año 2018 de 364.077 egresos asumiendo el crecimiento promedio anual que ha tenido los egresos hospitalarios de los adultos mayores en un 2 %. Por último, se asumió que dado un estudio que afirma que los adultos mayores que pasan por un tratamiento o se encuentran hospitalizados presentan síntomas depresivos entre un 15 % a 27 %[92].

Variable	Tasa
Crecimiento egresos hospitalarios	1,8 %
Crecimiento anual costos de hospitalización	2,7 %
Tasa de hospitalizados con depresión	21 %
Estimación de incremento de costos debido a la depresión	50 %

Tabla 3.20: Supuestos y datos relacionados para el calculo de impacto

Año	2017	2019	2030
Egresos hospitalarios	364.077	378.785	470.967
Costo	6.242.469	6.494.664	8.075.292
Adultos hospitalizados con depresión	76.456	79.544	98.903
Diferencia(Costo debido a la depresión)	$238,637,104,932	$261,863,302,464	$436,466,801,789

Tabla 3.21: Impacto de la depresión en adultos mayores

Según lo expuesto en la tabla 3.21 el impacto de la depresión en los adultos mayores es muy relevante y afecta de manera significativa en los costos que estos pudiesen tener el sistema de salud chileno, estos costos se elevan a medida que aumentan los costos de hospitalización en Chile, por lo tanto el impacto de los trastornos se vuelve más relevante a medida que pasa el tiempo en la salud publica y privada no tan solo por el impacto en los pacientes sino que también a nivel económico. y considerando los costos que pudiera ahorrar una familia estos costos podrían alcanzar los 3 millones en promedio.

Capítulo 4

Conclusiones

4.1. Conclusiones generales

Las principales conclusiones del trabajo realizado se subdividen en 4 secciones globales. Las primeras conclusiones sobre la calidad de los datos y como estos afectan al momento de elaborar los modelos y en los resultados finales, la segunda sección donde se discute sobre los descubrimientos y el poder predictivo de ciertas variables junto con su respectiva justificación, la tercera sección se analiza y compara el poder predictivo de las técnicas de Machine Learning junto con los problemas y soluciones. Por último, se responde las preguntas de investigación y se discute la hipótesis planteada.

4.1.1. Sobre calidad de los datos

Dentro del marco en el que se desarrolló esta memoria la Encuesta Nacional de Salud es la fuente de variables y preguntas de las que se dispone, pero esta base de datos cuenta con múltiples desventajas al momento de diseñar un modelo de riesgo de depresión.

La primera de estas es que las variables utilizadas son parte de instrumentos médicos con propósitos distintos a determinar si una persona tiene depresión, por lo tanto no se ajustan a una metodología y pudieran incidir en ciertos desajustes al momento de desarrollar un predictor práctico, por lo tanto se requeriría la formulación de preguntas elaboradas por especialistas en las áreas exploradas con el fin de hacerlas más propicias para el uso en personas que sufran de este trastorno y no inducir respuestas en los pacientes o sesgo en los cálculos.

Otro de los puntos a tener en consideración es sobre el desbalance entre las etiquetas de los que padecen trastornos de depresión en la encuesta nacional de salud que alcanza aproximadamente a ser el 12 % del total, lo que a posterior dificulta y entorpece la capacidad de aprendizaje del algoritmo. Además, la encuesta no cuenta con datos transversales que no permite vislumbrar la evolución del último año del afectado, se recomienda en este sentido

contar datos longitudinales con el fin de que el algoritmo pudiese discernir entre episodios depresivos y casos de depresión mayor.

También cabe destacar que la encuesta nacional de salud se desarrolló en base a factores de expansión lo que se puede ver en la distribución de edades de las muestras 3.1 donde la cantidad de adultos mayores por edad no es equivalente, por lo tanto al momento de entrenar el algoritmo y pudiese haber un sesgo de edad al no contar con los suficientes datos. Por último se requeriría tener un mayor volumen de datos y muestras con lo que entrenar y validar el modelo, y poder investigar otro tipos de procedimientos de Machine Learning como son las redes neuronales.

4.1.2. Variables relevantes

En el desarrollo del trabajo, se pudo encontrar distintas relaciones entre variables que son significativas y buenos productores de depresión, y otras que a pesar de la bibliográfica recopilada no fueron efectivas al momento de predecir depresión en adultos mayores chilenos, además de algunas conclusiones sobre los hábitos de los adultos mayores chilenos. A continuación, se repasa cada una de estas variables.

Socio-Demográficas: El resultado de las variables socio-demográfica demostró muchas preposiciones de la bibliografía, como fue el peso de las variables edad, sexo y nivel educativo. Además se encontraron nuevas correlaciones como el relativo a zonas urbanas o rurales.,se encontró una correlación entre el número de personas que viven en el hogar y la depresión en adultos mayores debido principalmente a su relación con el aislamiento social.

Actividad Física: En el caso de las variables asociadas a la actividad física y ejercicio no fueron del todo concluyentes con la aparición de depresión, aunque no se puede ratificar debido a que menos del 2 % de los adultos de la encuesta realizaban actividad física. Pero de todos modos dado los datos de la encuesta nacional de salud se puede concluir que la actividad física no es buen predictor para este caso.

Determinantes Sociales Y Psicológicos de Salud: La principal conclusión extraída de estas variables son la relevancia que las determinantes del entorno ya sean sociales o psicológicas como factor de riesgo de depresión, en la que cabe resaltar la importancia del barrio y la confianza que pueda tener en sus vecinos un adulto mayor, así mismo la importancia de los problemas del entorno familiar, de salud o el fallecimiento de un ser querido.

Perspectivas de Salud y Auto-reporte: Las conclusiones principales en el área de perceptiva de salud como mayores factores de riesgo se encuentran los problemas en la movilidad física, las dificultades de realizar tareas de la vida cotidiana y por ultimo las asociadas al estado de ánimo y el dolor que siente un adulto mayor.

Problemas Cardiovasculares: Los test realizados sobre las variables de riesgo cardiovascular reflejaron la relación que existe con la depresión especialmente aquellas variables asociadas a dolores en la zona toraxica y el riesgo cardiovascular.

Problemas Respiratorios: Los problemas respiratorios a pesar de no contar con mucha bibliógrafa que los respalden, en el caso de los adultos mayores, las dificultades o síntomas que conllevan una de estas enfermedades si que tiene relación con depresión, como son los problemas respiratorios al agitarse o caminar.

Hábitos Alimenticios: La dieta y los hábitos alimenticios demostraron no ser relevantes en la predicción de depresión, con la excepción de las porciones de frutas y verduras, de todos modos fue una relación débil, en comparación a las variables de entorno social o de perspectiva de salud.

Problemas de Audición y Visión: Los problemas de audición y visión son bastante comunes entre los adultos mayores chilenos, y por lo común también están relacionados con depresión, pero más que la patología o el diagnostico de una mala visión o audición, la percepción de la calidad de su propia audición y visión afecta en gran medida su propensión a depresión.

Trastornos del Sueño: Los trastornos del sueño están sumamente relacionados con trastornos depresivos, ansiedad y ciertas fobias, por lo tanto el estudio de estas variables en su mayoría respalda la bibliografía en cuanto a su importancia en la depresión, en particular la sensación de un mal descanso y el sentirse decaído y taciturno durante el día fueron las más relacionadas con la depresión.

Dolor Crónico y Fracturas: Una de las mayores sorpresas en el desarrollo de la memoria fue el encontrar una relación fuerte entre los dolores crónico con la depresión en adultos mayores, en especial el dolor en articulaciones como rodillas, codos , muñecas y tobillos. También se encontró relación con los dolores en el pecho y los dolores en la espalda baja.

Alcoholismo y tabaquismo: En cuanto a el alcoholismo y el tabaquismo, el alcoholismo estuvo mayormente relacionado con depresión en especial dependiendo la cantidad de alcohol que se consumía en una semana, en el caso del tabaco las principal razón de depresión fue que dicho adulto mayor viviera en un entorno con humo de tabaco.

Diabetes: Según la bibliografía que existe una gran relación entre los adultos mayores y diabetes , en especial que la diabetes tipo 2, los resultados de esta enfermedad no fueron del todo concluyentes y su intromisión en la posibilidad de tener depresión es bastante pequeño.

Digestión: Los problemas digestivo a diferente de otros problemas como los cardiovasculares y respiratorios, tuvieron menos efecto en la posibilidad de tener depresión, cabe destacar que las molestias y dolores de la zona abdominal fueron las únicas relacionadas con la depresión.

Otras: Las variables relevantes que no pertenecen a ninguno de los subgrupos anteriores, cabe destacar las variables asociadas a la movilidad, problemas cognitivos y de dependencia y por último aquellos relacionados con la memoria, en esta área destaco principalmente los antecedentes de depresión lo que está muy corroborado por la bibliografía.

4.1.3. Tecnicas de Machine Learning

En términos de técnicas de Machine Learning cada una de estas se comportó de manera distinta dependiendo la forma en que fue aplicada y bajo que variables, las principales conclusiones de cada modelo se detallan a continuación:

Logit: Los Resultados obtenidos mediante esta técnica se destacaron principalmente por el grado de flexibilidad y su estabilidad aunque no tenían los mejores resultados.

K-Nearest Neighbor: Un método bastante efectivo en las etapas de moldeamiento pero en la validación ya sea con la Encuesta Nacional de Salud del 2009-2010 o 2016-2017 tenía una radical baja en su rendimiento y lo que lo convierte un modelo poco flexible.

Arboles de Decisión: Arboles de decisión mostró un gran desempeño en especial en las etapa de validación que a pesar de no mostrar los mejores resultados a nivel general, mostró buenos niveles de predicción según indicadores como Recall o AUC.

Naive Bayes: Naive bayes mostró buenos niveles de desempeño tanto en la etapa de entrenamiento como con la etapa de validación en términos de AUC en especial en el modelo de variables de determinante social y psicológico.

Support Vector Machine: Los resultados obtenidos mediante esta técnicas fueron mediocres en la gran mayoría de los casos, en especial en términos de Log-Loss y CAP, esto se debe a lo complejo que son cada uno de los casos de depresión y cortar el con un hiper-plano la muestra del encuesta es muy complejo.

Random Forest: El método de Random Forest obtuvo muy buenos rendimientos en promedio tanto en el entrenamiento como en la validación según indicadores como Log-Loss y AUC, y bueno niveles en términos de Recall la cual es el indicador más importante de cara a los objetivos propuestos por el equipo médico.

Otra conclusión es en relación a las metodologías de balance de datos que afectan de manera sustancial el proceso de aprendizaje de los algoritmos, cada uno mostró tener ciertas ventajas y desventajas que se detallan a continuación.

Las técnicas de over-sampling demostraron ser una manera eficiente al realizar el balance en la base de datos , en el caso de la técnica SMOTE, es una técnica bastante eficiente para generar nuevos datos sintéticos distintos a los originales, lo que le da una mayor sensibilidad al modelo al cambiar la base de test, pero su principal consecuencia es generar datos ficticios sobrepuestos con datos de la muestra mayoritaria lo que genera un sesgo que entorpece el aprendizaje de el algoritmo (Anexo tabla A.58, tabla A.60, tabla A.62).

La otra técnica de over-sampling implementada fue el intentar multiplicar la cantidad de casos minoritarios repitiendo los datos de las etiquetas minoritarias, este tipo de técnica tiene la ventaja de ser bastante sencilla de implementar, pero genera dependencia de ciertas tendencias de los datos minoritarios. Por lo tanto, pierde flexibilidad al momento de validar la información.

Las técnicas de under-sampling mostraron ser una técnica eficiencia para obtener el balance en las bases de datos. El caso de la metodología de Clustering demostró ser inteligente para generar bases balanceadas con igual cantidad de casos positivos como negativos, su principal consecuencia es la gran pérdida de información debido a la formación de estos clusters que resumen la información de un conjunto de la muestra mayoritaria afectaba de manera significativa el rendimiento de los algoritmos(Anexo tabla A.59, tabla A.61, tabla A.63).

La técnica de k-nearest neighbor demostró tener una muy buena efectividad al momento de seleccionar datos de la clase mayoritario con el fin de eliminar solo aquellas que no entregaran mayor información, en el caso puntual de este experimento demostró ser muy efectivo entregando los mejores resultados tanto en la etapa de entrenamiento como de validación. Por último la técnica de under-sampling de selección aleatoria de la muestra mayoritaria que toma de manera aleatoria datos de la clase mayoritaria y la iguala con la clase minoritaria, este método demostró tener mucha perdida de información de la clase mayoritaria lo que afectaba el rendimiento y le entregaba muy poca flexibilidad en al momento de validar el modelo.

En cuanto a los indicadores utilizados, cabe destacar las ventajas y desventajas que demostraron en el experimento anterior y se detallan a continuación:

Accuracy: Este indicador demostró su poca eficacia al memento de trabajar con bases de datos desbalanceadas ya que, a pesar de los malos resultados de ciertos modelos, el Accuracy jamás bajo del 70 %, lo que lo hace poco fidedigno.

Precision: La precision o Especifidad demostró tener sus mejores resultados utilizando la metodología de balance de base de datos denominada SMOTE, principalmente ya que al elevar el número de etiquetas con depresión el modelo fue más flexible al cambio haciendo que su tasa se elevara.

Recall: Recall o Sensibilidad demostró ser un indicador relevante ya que refleja la tasa de efectividad de detección, cuyo mejor rendimiento se obtuvo mediante la técnica de balance denominada K-nearest neighbors ya que al reducir la incertidumbre y el sesgo que producían ciertas etiquetas de la clase mayoritaria el puntaje de Recall aumentaba.

AUC: Es uno de los indicadores más fidedignos de los utilizados y que entrega mayor información ya que nace de la especificad y la sensibilidad del modelo, en este caso particular fue un buen indicador para interpretar los resultados.

Log-Loss: La pérdida de información fue uno de los más eficientes al momento de evaluar una base de datos que estuviera desbalanceada ya que penalizo de manera efectiva los falsos positivos y falsos negativo, debido al desbalance la probabilidad de encontrar con estos casos era menor, por lo tanto, al sobre dimensionarlo entregaba mayor exactitud en sus estimaciones.

Cummulative Accuracy Profiles: Esta técnica utilizada en el mundo financiero para la evaluar los algoritmos de Machine Learning que detectan fraudes financieros en donde las bases de datos se encuentran sumamente desbalanceadas, en este caso fue un algoritmo muy

efectivo para entregar el poder predictivo ya sea de detectar como de rechazar depresión en los adultos mayores, además es muy intuitiva de evaluar.

4.1.4. Hipótesis

En cuanto a la hipótesis central de este proyecto que dice de la posibilidad de elaborar un predictor de riesgo de depresión utilizando la Encuesta Nacional de Salud y técnicas de Machine Learning , dado a los resultados de los modelos y el buen rendimiento de indicadores como AUC , CAP o Log-Loss lo hace comparable con otros predictores, se puede afirmar que se puede realizar dicho indicador con las variables presentadas en la encuesta, de todos modos está abierto a cambios de formulación en ciertas preguntas para que se adapte mejor a la condición que sufren personas con depresión.

Por último, la tesis de que se requiere un modelo solo para adultos mayores queda refutada principalmente porque los modelos no consideran variables únicamente relacionadas con sus patologías, principalmente porque no fueron relevantes en los test ni cuando se formulaban en modelos, por lo tanto el desarrollo focalizado no dio ninguna ventaja por sobre uno general como se puede ver al testear el modelo con todos los datos contabilizando a los menores de 60 de la encuesta los resultados obtenidos no difieren demasiado sea cual sea el modelo aplicado, a exceptuar por aquel que guarda mayor relación con patologías y enfermedades crónicas.

4.2. Trabajos futuros

Los trabajos a futuros en el diseño de un indicador de depresión principalmente van en el desarrollo de un modelo general para la depresión en consultas médicas a nivel primario o secundario. Con ello deberá ir el diseño de las preguntas por parte de un profesional junto con el desarrollo de la plataforma para el uso práctico del modelo.

Otro matiz interesante para el desarrollo de trabajos futuros seria focalizar el estudio en un aspecto o factor en particular que incida en la preponderancia a tener depresión, un ejemplo de esto sería utilizar variables relacionadas al dolor y evaluar cuan bien puede predecir depresión.

Otro gran trabajo sería graduar la sensibilidad del modelo y que sea capaz de detectar entre tener episodios depresivos leves, depresión en niveles moderados y por último depresión grave y con ello formular un modelo más preciso y que pudiera discernir y dar luces del grado de gravedad del trastorno depresivo.

Por último, desarrollar un mecanismo de recopilación de información con el cual nutrir de información al modelo, ya sean datos concretos sobre enfermedades, exámenes o imágenes y con ellos desarrollar metodologías más complejas que requieren mayor cantidad de información como son las redes neuronales o algoritmos genéticos.

Bibliografía

[1] 3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.21.3 documentation.

[2] scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation.

[3] Ryan J. Anderson, Kenneth E. Freedland, Ray E. Clouse, and Patrick J. Lustman. The Prevalence of Comorbid Depression in Adults With Diabetes: A meta-analysis. *Diabetes Care*, 24(6):1069–1078, June 2001.

[4] Ricardo Araya, Graciela Rojas, Rosemarie Fritsch, Julia Acuña, and Glyn Lewis. Common mental disorders in Santiago, Chile: Prevalence and socio-demographic correlates. *The British Journal of Psychiatry*, 178(3):228–233, March 2001.

[5] Jürgen Barth, Martina Schumacher, and Christoph Herrmann-Lingen. Depression as a Risk Factor for Mortality in Patients With Coronary Heart Disease: A Meta-analysis. *Psychosomatic Medicine*, 66(6):802, December 2004.

[6] Gustavo E. A. P. A. Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, May 2003.

[7] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[8] Virginia Bernabei, Valentina Morini, Francesca Moretti, Antonella Marchiori, Barbara Ferrari, Edoardo Dalmonte, Diana De Ronchi, and Anna Rita Atti. Vision and hearing impairments are associated with depressive–anxiety syndrome in Italian elderly. *Aging & Mental Health*, 15(4):467–474, May 2011.

[9] Paul A. Cafarella, Tanja W. Effing, Zafar-Ahmad Usmani, and Peter A. Frith. Treatments for anxiety and depression in patients with chronic obstructive pulmonary disease: A literature review. *Respirology*, 17(4):627–638, 2012.

[10] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, January 2014.

[11] L. Bentata Chocrón, J. Franch Vilalta, I. Rodríguez Legazpi, K. Auquer, L. Franch, and R. Arrizabalaga Ramírez. [The diagnosis of mental disorders by the primary care

physician]. *Atencion primaria*, 18(1):22–26, June 1996.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[13] Thomas J. Craig and Pearl A. Van Natta. Disability and depressive symptoms in two communities. *The American Journal of Psychiatry*, 140(5):598–601, 1983.

[14] Héctor Gómez Dantés, Victoria Castro, Francisco Franco-Marina, Paula Bedregal, Jesús Rodríguez García, Azalea Espinoza, William Valdez Huarcaya, Rafael Lozano, Joyce Mendes Schramm Andrade, Joaquim Goncalves Valente, et al. La carga de la enfermedad en países de américa latina. *salud pública de méxico*, 53:s72–s77, 2011.

[15] Departamento de Epidemiología. ENS - Encuesta nacional de salud, January 2016.

[16] Asamblea Mundial de la Salud. Plan de acción integral sobre salud mental 2013-2020. Technical report, 2013.

[17] Organización Mundial de la Salud. La inversión en el tratamiento de la depresión y la ansiedad tiene un rendimiento del 400 %, 2016.

[18] Organización Mundial de la Salud. *Trastornos mentales y del comportamiento,Décima revisión de la Clasificación Internacional de Enfermedades (CIE 10)*, volume 1. Meditor, Madrid, 1992.

[19] T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, May 1995.

[20] Joanna F Dipnall, Julie A Pasco, Michael Berk, Lana J Williams, Seetal Dodd, Felice N Jacka, and Denny Meyer. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PloS one*, 11(2):e0148195, 2016.

[21] Donald L. Dudley, Edward M. Glaser, Betty N. Jorgenson, and Daniel L. Logan. Psychosocial Concomitants to Rehabilitation in Chronic Obstructive Pulmonary Disease: 3. Dealing with Psychiatric Disease (as Distinguished from Psychosocial or Psychophysiologic Problems). *CHEST*, 77(5):677–684, May 1980.

[22] Bernd Engelmann, Evelyn Hayden, and Dirk Tasche. Testing rating accuracy. *Risk*, 16(1):82–86, 2003.

[23] DEPARTAMENTO DE EPIDEMIOLOGÍA. Encuesta nacional de salud 2016 – 2017, segunda entrega de resultados. *División de Planificación Sanitaria Subsecretaría de Salud Pública*, page 50, 2018.

[24] Paula Errázuriz, Camila Valdés, Paul A. Vöhringer, and Esteban Calvo. Financiamiento de la salud mental en Chile: una deuda pendiente. *Revista médica de Chile*, 143(9):1179–1186, September 2015.

[25] Shona Fielding, Peter M Fayers, Alison McDonald, Gladys McPherson, and Marion K

Campbell. Simple imputation methods were inadequate for missing not at random (mnar) quality of life data. *Health and Quality of Life Outcomes*, 6(1):57, 2008.

[26] Yaima Filiberto, Rafael Bello, Yailé Caballero, and Mabel Frías. ALGORITMO PARA EL APRENDIZAJE DE REGLAS DE CLASIFICACION BASADO EN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS EXTENDIDA. *Dyna*, 78(169), 2011.

[27] Luis Flores-Padilla, Flor Rocío Ramírez-Martínez, and Juana Trejo-Franco. Depresión en adultos mayores (AM) con pobreza extrema pertenecientes a un Programa Social en Ciudad Juárez, Chihuahua, México. *Gaceta Médica de México*, 152(4):439–443, September 2016.

[28] Peter L. Franzen and Daniel J. Buysse. Sleep disturbances and depression: risk relationships for subsequent depression and therapeutic implications. *Dialogues in Clinical Neuroscience*, 10(4):473–481, December 2008.

[29] M. J. García Serrano and J. Tobías Ferrer. Prevalencia de depresión en mayores de 65 años. Perfil del anciano de riesgo. *Atención Primaria*, 27(7):484–488, January 2001.

[30] Ellen R Girden. *ANOVA: Repeated measures*. Number 84. Sage, 1992.

[31] David J. Hand. Data Mining Based in part on the article "Data mining" by David Hand, which appeared in the Encyclopedia of Environmetrics. In *Encyclopedia of Environmetrics*. American Cancer Society, 2013.

[32] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[33] Peter Harrington. *Machine Learning in Action*. Manning Publications Co., Greenwich, CT, USA, 2012.

[34] Lajos Horváth and Emanuel Parzen. Limit Theorems for Fisher-Score Change Processes. *Lecture Notes-Monograph Series*, 23:157–169, 1994.

[35] Behshad Hosseinifard, Mohammad Hassan Moradi, and Reza Rostami. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer Methods and Programs in Biomedicine*, 109(3):339–345, March 2013.

[36] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[37] M. José Medrano, Elena Cerrato, Raquel Boix, and Miguel Delgado-Rodríguez. Factores de riesgo cardiovascular en la población española: metaanálisis de estudios transversales. *Medicina Clínica*, 124(16):606–612, April 2005.

[38] Balu Kalayam, Barnett S. Meyers, Tatsuyuki Kakuma, George S. Alexopoulos, Robert C. Young, Sandra Solomon, Rochelle Shotland, Draupathi Nambudiri, and Daniel Goldsmith. Age at onset of geriatric depression and sensorineural hearing deficits. *Biological Psychiatry*, 38(10):649–658, November 1995.

[39] M. King, C. Bottomley, J. Bellón-Saameño, F. Torres-Gonzalez, I. Svab, D. Rotar, M. Xavier, and I. Nazareth. Predicting onset of major depression in general practice attendees in Europe: extending the application of the predictD risk algorithm from 12 to 24 months. *Psychological Medicine*, 43(9):1929–1939, September 2013.

[40] Will Koehrsen. Random Forest Simple Explanation, December 2017.

[41] Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.

[42] Ting Hsiang Lin. A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality & Quantity*, 44(2):277–287, February 2010.

[43] Roderick J. A. Little. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83(404):1198–1202, December 1988.

[44] Bonita C. Long and Rosemary van Stavel. Effects of exercise training on anxiety: A meta-analysis. *Journal of Applied Sport Psychology*, 7(2):167–189, September 1995.

[45] Yen-Fu Luo and Anna Rumshisky. Interpretable Topic Features for Post-ICU Mortality Prediction. *AMIA Annual Symposium Proceedings*, 2016:827–836, February 2017.

[46] T. Maciejewski and J. Stefanowski. Local neighbourhood extension of SMOTE for mining imbalanced data. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 104–111, April 2011.

[47] Mikko I. Malinen and Pasi Fränti. Balanced K-Means for Clustering. In Pasi Fränti, Gavin Brown, Marco Loog, Francisco Escolano, and Marcello Pelillo, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science, pages 32–41. Springer Berlin Heidelberg, 2014.

[48] M Marcus, Mohammad Taghi Yasamy, M Van Ommeren, D Chisholm, and S Saxena. Depression: A global public health concern. *World Health Organization Paper on Depression*, pages 6–8, January 2012.

[49] Hugh Miller, Sandy Clarke, Stephen Lane, Andrew Lonie, David Lazaridis, Slave Petrovski, and Owen Jones. Predicting customer behaviour: The university of melbourne's kdd cup report. In *KDD-Cup 2009 Competition*, pages 45–55, 2009.

[50] Benson Mwangi, Tian Siva Tian, and Jair C Soares. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244, 2014.

[51] F. Müller-Spahn and C. Hock. Clinical Presentation of Depression in the Elderly. *Gerontology*, 40(Suppl. 1):10–14, 1994.

[52] Kevin w.Bowyer Nitesh V. Chawla. SMOTE: Synthetic Minority Over-sampling Tech-

nique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[53] Leticia Oliveira, Cecile D. Ladouceur, Mary L. Phillips, Michael Brammer, and Janaina Mourao-Miranda. What Does Brain Response to Neutral Faces Tell Us about Major Depression? Evidence from Machine Learning and fMRI. *PLOS ONE*, 8(4):e60121, April 2013.

[54] Pan American Health Organization. The burden of mental disorders in the region of the americas. *PAHO*, 2018.

[55] World Health Organization, editor. *Clasificación estadística internacional de enfermedades y problemas relacionados con la salud.* Number 554 in Publicación cientifica. OPS, Oficina Sanitaria Panamericana, Oficina Regional de la Organización Mundial de la Salud, Washington, D.C, décima revisión. [10a rev.] edition, 1995.

[56] World Health Organization et al. *Preventing suicide: A global imperative*. World Health Organization, 2014.

[57] World Health Organization et al. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017.

[58] Ronit Fischman Susana Morales Jorge Barros Orietta Echávarri, María de la Paz Maino. Aumento sostenido del suicidio en chile: un tema pendiente. *Centro de Politicas Públicas UC*, 79:2–13, June 2015.

[59] Victoria Pachón Álvarez, Jacinto Mata Vázquez, Francisco Roche Beltrán, José Cristóbal Riquelme Santos, and José María Tejera. Practical application of kdd techniques to an industrial process. 2004.

[60] M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, January 2005.

[61] Fredy Alonso Patiño Villada, Elkin Fernando Arango Vélez, and Lucidia Zuleta Baena. Ejercicio físico y depresión en adultos mayores: una revisión sistemática. *Revista Colombiana de Psiquiatría*, 42(2):198–211, April 2013.

[62] Pierre Pichot, J. J López-Ibor Aliño, and Manuel Valdés Miyar. *Manual diagnóstico y estadístico de los trastornos mentales: DSM-IV*. Masson, Barcelona, 2001. OCLC: 906688025.

[63] E.Salazar P.Olivares. Gasto en Salud y Proximidad a la muerte del Adulto Mayor en el Sistema Isapre: Análisis Econométrico (Parte II), 2016.

[64] N Poolsawad, C Kambhampati, and JGF Cleland. Balancing class for performance of classification with a clinical dataset. In *Proceedings of the World Congress on engineering*, volume 1, pages 1–6, 2014.

[65] Cristina Portellano-Ortiz, Josep Garre-Olmo, Laia Calvó-Perxas, and Josep Lluís Conde-Sala. Depresión y variables asociadas en personas mayores de 50 años en españa. *Revista*

de Psiquiatría y Salud Mental, 11(4):216–226, 2018.

[66] Claudia Dechent R. Depresión geriátrica y trastornos cognitivos. *Revista Hospital Clínico Universidad de Chile*, 19:339, 2008.

[67] Atif Rahman, Vikram Patel, Joanna Maselko, and Betty Kirkwood. The neglected 'm' in MCH programmes–why mental health of mothers is important for child nutrition. *Tropical medicine & international health: TM & IH*, 13(4):579–583, April 2008.

[68] Victoria de la Caridad Ribot Reyes, Maritza Alfonso Romero, Martha Elena Ramos Arteaga, and Antonio González Castillo. Suicidio en el adulto mayor. *Revista Habanera de Ciencias Médicas*, 11(S5):699–708, 2012.

[69] L.N. Robins, Wing JK, H.U. Wittchen, J.E. Helzer, T.F. Babor, Jack Burke, A Farmer, A Jablenski, R Pickens, and Darrel Regier. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of general psychiatry*, 45:1069–77, January 1989.

[70] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, May 1991.

[71] Sandra Saldivia, Benjamin Vicente, Louise Marston, Roberto Melipillán, Irwin Nazareth, Juan Bellón-Saameño, Miguel Xavier, Heidi Ingrid Maaroos, Igor Svab, M.-I. Geerlings, and Michael King. [Development of an algorithm to predict the incidence of major depression among primary care consultants]. *Revista Medica De Chile*, 142(3):323–329, March 2014.

[72] Albert Satorra and Peter M. Bentler. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4):507–514, December 2001.

[73] Umair Shafique and Haseeb Qaiser. A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1):217–222, 2014.

[74] G. E. Simon, M. VonKorff, M. Piccinelli, C. Fullerton, and J. Ormel. An international study of the relation between somatic symptoms and depression. *The New England Journal of Medicine*, 341(18):1329–1335, October 1999.

[75] SUBSECRETARIA DE PREVISIÓN SOCIAL. Principales resultados de la encuesta de calidad de vida del adulto mayor e impacto del pilar solidario. *Dirección de Estudios Previsionales*, 19:26, 2018.

[76] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[77] William J. Strawbridge, Stéphane Deleger, Robert E. Roberts, and George A. Kaplan. Physical Activity Reduces the Risk of Subsequent Depression for Older Adults. *American Journal of Epidemiology*, 156(4):328–334, August 2002.

[78] Tara W. Strine, Ali H. Mokdad, Shanta R. Dube, Lina S. Balluz, Olinda Gonzalez, Joyce T. Berry, Ron Manderscheid, and Kurt Kroenke. The association of depression and anxiety with obesity and unhealthy behaviors among community-dwelling US adults. *General Hospital Psychiatry*, 30(2):127–137, April 2008.

[79] Michael Strober and Jack L. Katz. Do eating disorders and affective disorders share a common etiology? A dissenting opinion. *International Journal of Eating Disorders*, 6(2):171–180, 1987.

[80] COMISIÓN NACIONAL DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA. Programa idea-primera etapa de ciencia aplicada, informe de avance anual. *Web Intelligence Center*, 3(5), 2018.

[81] Megan Teychenne, Kylie Ball, and Jo Salmon. Sedentary Behavior and Depression Among Adults: A Review. *International Journal of Behavioral Medicine*, 17(4):246–254, December 2010.

[82] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, and G. J. Gelpke. Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients. *Journal of the Royal Statistical Society. Series A (General)*, 144(2):145, 1981.

[83] Madhukar H. Trivedi. The Link Between Depression and Physical Symptoms. *Primary Care Companion to The Journal of Clinical Psychiatry*, 6(suppl 1):12–16, 2004.

[84] Norifumi Tsuno, Alain Besset, and Karen Ritchie. Sleep and Depression. *The Journal of Clinical Psychiatry*, 66(10):1254–1269, 2005.

[85] Jürgen Unützer, Donald L. Patrick, Greg Simon, David Grembowski, Edward Walker, Carolyn Rutter, and Wayne Katon. Depressive Symptoms and the Cost of Health Services in HMO Patients Aged 65 Years and Older: A 4-Year Prospective Study. *JAMA*, 277(20):1618–1623, May 1997.

[86] Mónica Valdés King, José Alberto González Cáceres, and Mansur Salisu Abdulkadir. Prevalencia de depresión y factores de riesgo asociados a deterioro cognitivo en adultos mayores. *Revista Cubana de Medicina General Integral*, 33(4):0–0, December 2017.

[87] Benjamín Vicente P, Romina Rojas P, Sandra Saldivia B, Cristhian Pérez V, Roberto Melipillán A, Nain Hormazábal P, and Rolando Pihan V. Determinantes biopsicosociales de depresión en pacientes atendidos en Centros de Atención Primaria de Concepción, Chile. *Revista chilena de neuro-psiquiatría*, 54(2):102–112, June 2016.

[88] Flavia Vivaldi and Enrique Barra. Bienestar Psicológico, Apoyo Social Percibido y Percepción de Salud en Adultos Mayores. *Terapia psicológica*, 30(2):23–29, July 2012.

[89] GÓMEZ R. y GONZÁLEZ M VON MUHLENVROCK F. Prevalencia de depresión en pacientes mayores de 60 años hospitalizados en el servicio de medicina interna del hospital militar de santiago. *Revista neuro-psiquiatria*, 4:332, 2011.

[90] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[91] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[92] Kara Zivin, Tracy Wharton, and Ola Rostant. The economic, public health, and caregiver burden of late-life depression. *The Psychiatric clinics of North America*, 36(4):631–649, December 2013.

[93] Sergio Zuniga-Jara, Oscar Sjoberg-Tapia, and Diego Opazo-Gallardo. Análisis de las proyecciones de crecimiento económico del banco central de chile: 1991-2017. *Información tecnológica*, 30(1):133–142, 2019.

Anexo y Apendices

A. Preguntas DSM-IV

Códigos	Etiqueta
sd1	Durante los últimos 12 meses ¿ha tenido Ud. dos semanas seguidas o más en que se sintió triste, desanimado(a) o deprimido(a)?
sd2	Por favor piense en ese período de dos semanas o más, durante los últimos 12 meses, cuando estos sentimientos fueron peores. Durante ese período de dos semanas, el sentirse triste, desanimado(a) o deprimido(a) ¿le duró usualmente.
sd3	Durante esas dos semanas, ¿se sintió de esa manera ...
sd4	Durante esas dos semanas o más, ¿perdió Ud. el interés en la mayoría de las cosas (como pasatiempos, trabajo o actividades que usualmente son placenteras para Ud.?
sd5	Durante esas mismas dos semanas, ¿se sintió Ud. más cansado(a) o con menos energía de lo habitual para usted?
sd6	¿Perdió o ganó peso sin desearlo, o se quedó más o menos en el mismo peso?
sd7	¿Alrededor de cuántos kilos de peso ganó y/o perdió?
sd8	¿El cambio fue mayor o igual a 5 kilos?
sd9	Durante esas dos semanas, ¿tenía Ud. más problemas de lo habitual para quedarse dormido?
sd10	Durante esas dos semanas, ¿le pasó esto ...
sd11	Durante esas dos semanas, ¿tenía Ud. más dificultad de lo habitual para concentrarse? (Si el(la) entrevistado(a) pregunta si aún estamos hablando de se período de dos semanas o más, contestele que sí.)
sd12	Las personas a veces se sienten mal consigo mismas, que valen poco, que no son suficientemente buenas. Durante esas dos semanas ¿se sintió Ud. de esa forma?
sd13	Durante esas dos semanas, ¿pensó mucho en la muerte, ya fuese en la suya, en la de alguien más o en la muerte en general?
sd13x	Vamos a repasar, Ud. mencionó que tenía dos semanas seguidas, durante los últimos 12 meses, en que se sintió triste, desanimado(a) o deprimido(a) y también tenía otros problemas como...
sd14	Durante los últimos 12 meses ¿alguna vez hubo un período que duró dos semanas o más en que Ud. Perdió el interés en la mayoría de las cosas como pasatiempos, trabajo o actividades que usualmente hace para divertirse?
sd15	Durante ese período de dos semanas, la pérdida de interés en las cosas ¿le duró usualmente ...
sd16	Durante esas dos semanas, ¿se sintió de esa manera ...
sd17	Durante esas mismas dos semanas, ¿se sintió Ud. agotado(a), o con menos energía de lo habitual para Ud.?
sd18	¿Perdió o ganó peso sin desearlo, o se quedó más o menos en el mismo peso? (Si el entrevistado pregunta si aún estamos hablando de ese período de dos semanas o más, conteste sí.)
sd19	¿Alrededor de cuántos kilos de peso ganó y/o perdió?
sd20	¿El cambio fue mayor o igual a 5 kilos?
sd21	Durante esas dos semanas, ¿tenía Ud. más problemas de lo habitual para quedarse dormido?
sd22	Durante esas dos semanas, ¿le pasó esto ...
sd23	Durante esas dos semanas, ¿tenía Ud. más dificultad de lo habitual para concentrarse?
sd24	Las personas a veces se sienten mal consigo mismas, que no son suficientemente buenas o valen poco. Durante esas dos semanas ¿se sintió Ud. de esa forma?
sd25	Durante esas dos semanas, ¿pensó mucho en la muerte, ya fuese en la suya, en la de alguien más o en la muerte en general?
sd25x	Vamos a repasar, Ud. mencionó que tenía dos semanas seguidas, durante los últimos 12 meses, en que se sintió triste, desanimado(a) o deprimido(a) y también tenía otros problemas como...
sd26	¿Cuánto han interferido estos problemas con su vida o actividades ...
sd27	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de depresión?
sd28	¿A qué edad se lo diagnosticaron por primera vez?
sd29	¿Alguna vez ha recibido tratamiento por depresión?
sd30	¿Ha estado tomando algún medicamento durante las últimas 2 semanas para la depresión?

Tabla A.1: Variables y preguntas de DSM-IV

B. Test Variables

B.1. Variables Socio-Demográficas

Códigos	Etiqueta
EDAD	EDAD (Al 31 de enero 2010)
NEDU	NEDU
ZONA	ZONA
REGION	REGION (Considera 15 regiones)
SEXO	SEXO
nper	Número de persona en el hogar incluidas en la Tabla Kish
EDAD_CODIFICADA	GRUPOS EDAD
EDAD_REPORTE	EDAD
numhogar	numhogar Número de integrantes del hogar
ns3	¿Cuál es la relación de parentesco con el jefe de hogar?
ns4	¿Cuál es su estado civil actual?
ns6	¿Asiste actualmente a algún establecimiento educacional?
ns7c	¿Cuál es el último curso aprobado?
ns8	¿A qué nivel educacional corresponde?
ns9c	(corregido) Número de años cursado completos
ns10	¿En cuál de las siguientes situaciones se encontraba?
ns19_1	Calefont o sistema de calentamiento de aguas
ns19_2	Horno microondas
ns19_3	Computador o notebook
ns19_4	Videograbador, DVD o pasapelículas
ns19_5	Refrigerador
ns19_6	Automóvil de uso particular
ns19_7	¿Tiene alguno de los siguientes bienes?: No responde
ns20	¿De dónde proviene el agua que usa esta vivienda?
ns21	¿Por dónde llega agua a esta vivienda?:
ns22	¿Cómo funciona el servicio higiénico (W.C.) de esta vivienda?:
ns23	¿De dónde llega la electricidad a esta vivienda?
ns24	¿Cuál es el principal combustible usado para cocinar?
ns25	¿Cuántas duchas tiene esta vivienda?
ns26	Tipo de vivienda:
ns27	En la cubierta del techo predomina:
ns28	El material predominante en las paredes exteriores es:
ns29	El material predominante en el piso es:

Tabla A.2: Variables Demográficas

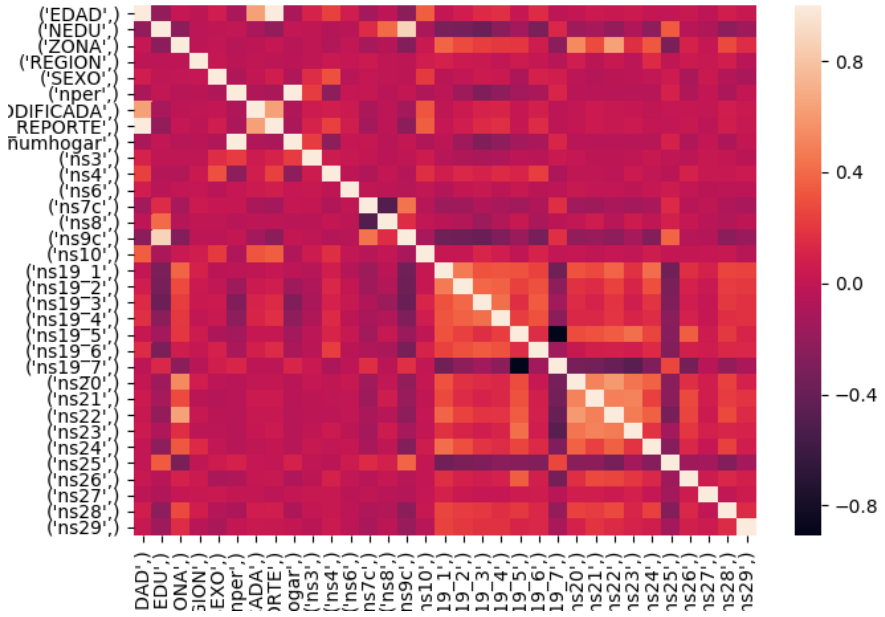


Figura A.1: Correlación de variables socio-demográficas

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
EDAD	7.972	ns13cod	0.279
NEDU	7.605	ns19_1	0.252
ZONA	2.451	ns19_2	0.248
REGION	2.109	ns19_3	0.227
SEXO	1.137	ns19_4	0.153
nper	1.109	ns19_5	0.103
c1	1.081	ns19_6	0.074
EDAD_COD	1.005	ns19_7	0.061
EDAD_REP	0.800	ns20	0.050
numhogar	0.721	ns21	0.041
ns3	0.713	ns22	0.027
ns4	0.713	ns23	0.024
ns6	0.681	ns24	0.012
ns7c	0.661	ns25	0.005
ns8	0.467	ns26	0.003
ns9c	0.467	ns27	0.003
ns10	0.422	ns28	0.002
ns11	0.390	ns29	0.001
ns12cod	0.307		

Tabla A.3: Resultados test Chi-Cuadrado variables socio-demográficas

Variable	Resultado	Variable	Resultado
ns29	0.431998104	REGION	0.141967158
ns12cod	0.418480084	ns11	0.078581759
EDAD_COD	0.345656604	ns3	0.047021721
ns9c	0.24030859	EDAD	0.025483111
ns10	0.213902976	EDAD_REP	0.020618481
SEXO	0.000183514		

Tabla A.4: Resultados test Kolorov-Smirnov variables socio-demograficas

Variable	Variable	Resultado
ns22	ns20	0.972
ns9c	ns19_1	0.909
EDAD	ns8	0.724
ns13cod	ZONA	0.568
ns19_2	ns10	0.497
ns19_4	ns10	0.324
ns19_2	ns19_4	0.180
nper	ns24	0.171
ns24	numhogar	0.171
ZONA	ns24	0.162
EDAD_COD	ns19_3	0.051

Tabla A.5: Resultados test Anova variables socio-demograficas

B.2. Variables de Actividad Física

Códigos	Etiqueta
a1	¿Exige su trabajo una actividad física intensa que implica un gran aumento de la respiración o de la frecuencia cardíaca, como transportar o levantar cargas pesadas, trabajos de construcción, etc durante al menos 10 minutos seguidos?
a2	En una semana normal, ¿cuántos días realiza usted estas actividades físicas intensas en su trabajo?
a4	¿Exige su trabajo una actividad de intensidad moderada que implica un aumento moderado de la respiración o de la frecuencia cardíaca?,
a5	En una semana normal, ¿cuántos días realiza usted estas actividades de intensidad moderada en su trabajo?
a7	...cómo se traslada o va de un sitio a otro, por ejemplo, ir al trabajo, a comprar, a la feria, a la iglesia, a dejar o buscar niños a la escuela, etc. ¿Camina o usa usted una bicicleta al menos 10 minutos seguidos en sus traslados?
a8	En una semana normal, ¿cuántos días camina o va en bicicleta al menos 10 minutos seguidos en sus traslados?
a10	¿En su tiempo libre, practica usted deportes/entrenamientos intensos que implican un gran aumento de la respiración o de la frecuencia cardíaca?
a11	En una semana normal, ¿cuántos días practica usted estos deportes o ejercicios intensos en su tiempo libre?
a13	¿En su tiempo libre practica usted alguna actividad de intensidad moderada que implique un aumento moderado (mediano) de la respiración o de de la frecuencia cardíaca (latidos del corazón)?
a14	En una semana normal ¿cuántos días practica usted estas actividades físicas de intensidad moderada en su tiempo libre?
a16_2	¿Cuánto tiempo suele pasar sentado o acostado (tendido, reclinado) en un día normal? (Mostrar TARJETA 15) (Minutos)
a17	¿En el último mes practicó deporte o realizó actividad física fuera de su horario de trabajo, durante 30 minutos o más cada vez?

Tabla A.6: Variables Actividad Física

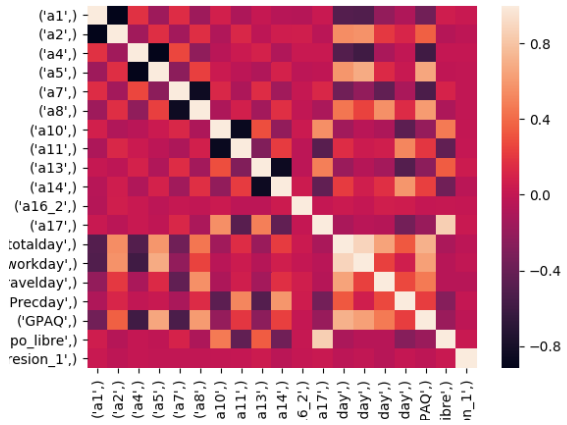


Figura A.2: Correlación de Actividad Física

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
a1	0.956	a16_2	0.046
Ptravelday	0.713	a5	0.041
a2	0.305	a4	0.039
GPAQ	0.276	sedent_t_libre	0.032
a13	0.250	a8	0.030
Ptotalday	0.148	Pworkday	0.028
a1	0.078	a11	0.021
a7	0.056	Precday	0.018
a17	0.056	a10	0.001

Tabla A.7: Resultados test Chi-Cuadrado variables actividad física

Variable	Variable	Resultado
GPAQ	a7	0.701
GPAQ	a8	0.671
a2	a14	0.631

Tabla A.8: Resultados test Anova variables actividad física

B.3. Variables Determinantes Sociales y Psicológicos de Salud

Códigos	Etiqueta
dataF1_2p1	¿Cuán de acuerdo está Ud. con la siguiente frase: "A nadie le importa mucho lo que me pasa"? (Mostrar TARJETA 20).
dataF1_2p2	¿Cuán de acuerdo está Ud. con la siguiente frase: "Es más seguro no confiar en nadie"? (Mostrar TARJETA 20).
p2a	En general, ¿cuánto confía usted en la gente de su villa, barrio o población? (Califique su percepción de confianza de 1 a 7, considerando que 1 es nada de confianza y 7 mucha confianza)
dataF1_2p3	Cuando tiene problemas, ¿tiene Ud. alguna persona en quien confiar, pedir ayuda o consejo?
dataF1_2p4	Puede recurrir confiadamente a alguien cuando tiene un gasto imprevisto, emergencia económica u otra situación grave o catastrófica?
p4a	Si se le cayera su monedero o billetera en su barrio, calle, villa o población y alguien la viera, piensa ¿Ud. que él o ella se la devolvería?
p4b	¿Diría usted que esta villa, barrio o población es un lugar donde los vecinos se preocupan unos de otros?
dataF1_2p6	¿Con qué frecuencia se ha sentido estresado durante los últimos 2 meses, es decir, irritable, con ansiedad o sin poder dormir, debido a situaciones en la casa o en el trabajo?
dataF1_2p7	¿Qué nivel de estrés financiero, económico, de dinero o plata ha sentido Ud. en los últimos 12 meses?
p8a	¿Ha experimentado alguno de los siguientes eventos en los últimos 12 meses? Se ha divorciado o separado?
p8b	Ha perdido el trabajo o jubilado?
p8c	Le han andado mal los negocios?
p8d	Le ha sucedido un hecho de violencia?
p8e	Ha tenido algún problema grande en la familia?
p8f	Ha tenido algún problema serio de salud o accidente?
p8g	Se le ha muerto el esposo(a) o pareja?
p8h	Se le ha enfermado o muerto alguien cercano de la familia?
p8i	Ha tenido algún otro estrés importante?

Tabla A.9: Variables Determinantes Sociales y Psicológicos de Salud

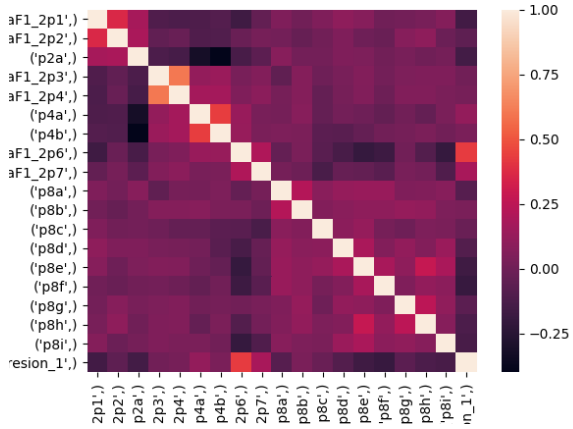


Figura A.3: Correlación de Variables Sociales y Psicológicas

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
dataF1_2p6	59.529	p8i	0.626
dataF1_2p7	9.725	p4b	0.308
p2a	4.027	p8g	0.243
dataF1_2p1	3.850	p8d	0.239
p8f	3.492	dataF1_2p4	0.181
p4a	2.826	p8a	0.109
p8e	2.524	p8b	0.076
p8h	2.433	dataF1_2p3	0.042
dataF1_2p2	0.934	p8c	0.019

Tabla A.10: Resultados test Chi-Cuadrado de variables sociales y psicológicas

Variable	Resultado	Variable	Resultado
p4a	4.2.E-03	dataF1_2p1	7.3E-06
p8h	1.5.E-03	p8f	3.3E-06
p8e	2.5.E-04	dataF1_2p7	2.0E-06
p2a	1.9.E-05	dataF1_2p6	7.0E-20

Tabla A.11: Resultados test Kolmogorov-Smirnov de variables sociales y psicológicas

Variable	Variable	Resultado
dataF1_2p4	dataF1_2p6	0.33061893

Tabla A.12: Resultados test Anova de variables sociales y psicológicas

B.4. Perspectiva de Salud y Auto-reporte

Códigos	Etiquetas
cd1	¿Cómo se siente con su vida en general (con su trabajo, familia, bienestar, salud, amor)?
cd2	En general Ud. diría que su salud es:
cd3	Su salud actual, ¿lo(a) limita para realizar esfuerzos moderados como mover una mesa, barrer, pasar la aspiradora o caminar más de 1 hora?
cd4	Su salud actual ¿lo(a) limita para realizar esfuerzos moderados como subir varios pisos por la escalera?
cd5	Durante las últimas 4 semanas, ¿con qué frecuencia ha tenido alguno de los siguientes problemas en su trabajo o en sus actividades cotidianas a causa de su SALUD FISICA ... ¿Hizo menos de lo que hubiera querido hacer?
cd6	Durante las últimas 4 semanas, ¿con qué frecuencia ha tenido alguno de los siguientes problemas en su... a causa de su SALUD FISICA... ¿Tuvo que dejar de hacer algunas tarea s en su trabajo o en sus actividades cotidianas?
cd7	¿Con qué frecuencia ha tenido alguno de los siguientes problemas en su trabajo o en sus actividades cotidianas a causa de algún problema EMOCIONAL?
cd8	¿Hizo su trabajo u otra actividad con menos cuidado que el de costumbre por algún problema emocional?
cd9	Durante las últimas 4 semanas, si ha tenido algún dolor, ¿hasta que punto éste ha interferido con sus tareas normales (incluido el trabajo dentro y fuera de la casa)?
cd10	En las últimas 4 semanas. En cada pregunta responda a lo que más se parezca a como se ha sentido Ud. durante últimas 4 semanas... ¿Con qué frecuencia se sintió tranquilo(a) y calmado(a)?
cd11	¿Con qué frecuencia se sintió con mucha energía?
cd12	¿Con qué frecuencia se sintió desanimado(a) o deprimido(a)?
cd13	Durante las últimas 4 semanas, ¿con qué frecuencia su salud física o los problemas emocionales han dificultado sus actividades sociales (como por ejemplo visitar amigos o familiares)?
cd14	La siguientes preguntas se refieren a su estado de salud general, incluida la salud física y la mental. En general, ¿cómo calificaría hoy su estado de salud?
cd15	En general, durante los últimos 30 días, ¿qué grado de dificultad ha tenido para realizar las tareas del trabajo y del hogar?
cd16	En general, durante los últimos 30 días, ¿qué grado de dificultad ha tenido para desplazarse de un lugar a otro?
cd17	¿Y qué grado de dificultad ha tenido para realizar actividades intensas, como correr 3 km o andar en bicicleta 12?
cd18	En general, durante los últimos 30 días, ¿qué grado de dificultad ha tenido para asearse, bañarse, lavarse las manos, vestirse,etc.?
cd19	¿Y qué grado de dificultad ha tenido para cuidar y mantener su aspecto general (maquillarse, peinarse, rasurarse, etc.)?
cd20	En general, durante los últimos 30 días, ¿qué grado de molestia o dolor ha tenido?
cd21	¿Y cuánto malestar en el cuerpo ha sufrido?
cd22	En general, durante los últimos 30 días, ¿qué grado dificultad ha tenido para concentrarse o recordar cosas?
cd23	¿Y qué grado de dificultad ha tenido para aprender una nueva tarea (por ejemplo un juego nuevo o una nueva receta,etc)?
cd24	En general, durante los últimos 30 días, ¿que grado de dificultad ha tenido para relacionarse con otras personas o para participar en actividades comunitarias?
cd25	¿Y que grado de dificultad ha tenido para enfrentarse a conflictos y tensiones con otras personas?
cd26	¿Utiliza usted lentes? (si la respuesta es Sí aclarar que las preguntas que siguen son tomando en cuenta sus lentes)
cd27	Durante los últimos 30 días ¿que grado dificultad tuvo para ver y reconocer de lejos su micro o colectivo ?
cd28	¿Y qué grado de dificultad tuvo para ver y reconocer un objeto que estuviera a la distancia de su mano o al leer el diario?
cd29	Durante los últimos 30 días, ¿en que medida tuvo problemas como quedarse dormido durante el dia, despertarse frecuentemente durante la noche o despertarse demasiado temprano en la mañana?
cd30	¿Y que tanta dificultad tuvo para sentirse descansado o repuesto durante el día?)
cd31	En general, durante los últimos 30 días, ¿en qué grado se ha sentido triste, decaído o deprimido?
cd32	¿Y en qué grado ha tenido preocupación o ansiedad?
WHS_SaludGeneral	PROMEDIO Preg14, Preg15 (Modulo III. F1)
WHS_Movilidad	PROMEDIO Preg16, Preg17 (Modulo III. F1)
WHS_CuidadoPersonal	PROMEDIO Preg18, Preg19 (Modulo III. F1)
WHS_DolorMalestar	PROMEDIO Preg20, Preg21 (Modulo III. F1)
WHS_Cognici_n	PROMEDIO Preg22, Preg23 (Modulo III. F1)
WHS_ActividadesSociales	PROMEDIO Preg24, Preg25 (Modulo III. F1)
WHS_Vista	PROMEDIO Preg27, Preg28 (Modulo III. F1)
WHS_Sue_oyEnergiaVital	PROMEDIO Preg29, Preg30 (Modulo III. F1)
WHS_EstadodeAnimo	PROMEDIO Preg31, Preg32 (Modulo III. F1)
WHS_Global	PROMEDIO DE TODAS LAS PREGUNTAS 14-32 MENOS LA 26 (Modulo III. F1)

Tabla A.13: Variables de Perspectiva de Salud y Auto-reporte

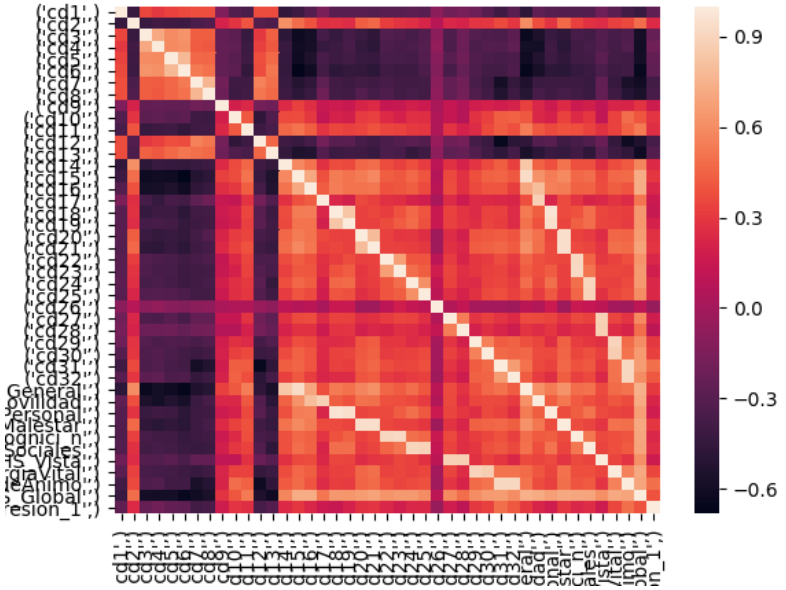


Figura A.4: Correlación de variables sobre la perspectiva de salud
Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
cd31	65.997	cd8	15.534
WHS_Animo	55.473	cd11	15.120
cd32	46.139	WHS_Salud	14.209
cd30	34.270	cd3	13.778
WHS_Energia	30.732	cd6	12.867
cd29	28.195	cd22	12.849
cd15	27.937	cd18	12.170
cd24	25.204	cd4	12.037
WHS_ActSocia	24.274	cd5	12.016
cd25	23.380	cd27	11.995
cd19	23.148	WHS_Cogn	11.687
cd10	21.070	cd23	10.529
cd7	20.635	WHS_Movi	9.811
cd16	19.907	cd9	7.800
cd21	18.648	WHS_Vista	7.342
WHS_Global	18.561	cd14	6.319
WHS_Dolor	18.121	cd17	5.970
cd20	17.598	cd1	4.645
cd12	17.458	cd2	4.206
WHS_Cuidado	16.901	cd28	3.964
cd13	15.608	cd26	1.819

Tabla A.14: Resultados test Chi-Cuadrado variables sobre la perspectiva de salud

Variable	Resultado	Variable	Resultado
cd28	1.2.E-02	cd5	2.6.E-12
cd18	1.0.E-02	cd11	7.6.E-13
cd19	5.4.E-04	cd20	3.4.E-13
WHS_Cuidado	4.0.E-04	cd21	2.3.E-13
WHS_Vista	3.1.E-05	cd6	9.4.E-14
cd17	1.9.E-05	WHS_Dolor	8.4.E-15
cd27	2.4.E-06	cd15	1.3.E-15
cd23	2.5.E-07	cd30	1.1.E-16
cd4	6.1.E-08	WHS_Salud	6.4.E-17
cd25	3.9.E-08	cd13	3.5.E-18
WHS_Movi	1.6.E-08	cd29	4.4.E-19
cd2	1.2.E-08	cd10	1.2.E-19
cd9	7.0.E-09	cd8	6.4.E-22
cd22	6.1.E-09	cd12	3.9.E-22
cd3	4.1.E-09	WHS_Energia	1.4.E-23
WHS_Cogn	4.1.E-09	cd7	3.8.E-26
cd24	4.5.E-10	WHS_Global	1.6.E-26
WHS_ActSocia	3.8.E-10	cd32	1.8.E-27
cd14	3.6.E-10	cd31	7.8.E-35
cd16	2.9.E-10	WHS_Animo	3.7.E-36
cd1	4.1.E-11		

Tabla A.15: Resultados test Kolgomorov-Smirnov sobre perspectiva de salud

Variable	Variable	Resultado	Variable	Variable	Resultado
cd13	cd7	0.945	cd10	WHS_Animo	0.37
cd10	WHS_Global	0.934	cd31	WHS_Cogn	0.341
cd28	WHS_Animo	0.89	WHS_Cuidado	cd18	0.336
cd10	cd31	0.883	WHS_Vista	cd27	0.329
cd12	cd3	0.859	cd28	WHS_Vista	0.322
WHS_Cogn	cd15	0.857	cd19	WHS_Cuidado	0.313
cd3	cd6	0.844	cd29	cd15	0.31
WHS_Dolor	cd20	0.815	cd31	cd15	0.294
cd21	WHS_Dolor	0.814	cd2	cd6	0.291
WHS_ActSocia	cd24	0.809	cd16	WHS_Energia	0.284
WHS_Animo	WHS_Cogn	0.808	cd12	cd17	0.261
cd11	cd20	0.803	cd10	WHS_Cogn	0.258
WHS_Global	cd31	0.803	WHS_Global	WHS_Animo	0.257
cd25	WHS_ActSocia	0.802	cd1	cd13	0.233
cd32	cd15	0.787	cd10	WHS_Energia	0.229
cd27	WHS_Energia	0.77	cd10	cd15	0.224
WHS_Vista	cd23	0.765	WHS_Vista	WHS_Animo	0.212
WHS_Global	WHS_Vista	0.724	cd28	cd23	0.21
cd28	WHS_Cogn	0.723	cd22	cd32	0.207
cd10	WHS_Vista	0.707	cd29	WHS_Cogn	0.203
WHS_Animo	cd15	0.687	cd1	cd7	0.188
cd29	cd22	0.684	WHS_Energia	cd31	0.187
cd12	cd6	0.671	WHS_Global	WHS_Energia	0.185
cd21	cd20	0.648	cd10	cd27	0.178
WHS_Energia	cd23	0.648	cd31	cd32	0.169
cd25	cd24	0.64	WHS_Global	WHS_Cogn	0.16
cd32	WHS_Cogn	0.63	cd16	cd23	0.154
cd11	WHS_Dolor	0.621	WHS_Global	cd27	0.148
cd28	cd15	0.621	cd27	cd31	0.147
WHS_Vista	cd31	0.611	WHS_Global	cd15	0.146
cd28	cd31	0.606	WHS_Vista	WHS_Cogn	0.14
cd12	cd2	0.51	cd22	cd15	0.137
cd10	cd28	0.506	cd28	cd29	0.136
cd10	cd23	0.5	cd29	WHS_Animo	0.134
cd27	cd23	0.494	WHS_Vista	cd15	0.126
cd6	cd17	0.493	WHS_Animo	cd23	0.125
WHS_Global	cd23	0.487	cd10	cd32	0.12
cd16	cd27	0.487	cd16	WHS_Vista	0.083
cd11	cd21	0.476	cd2	cd17	0.08
WHS_Animo	cd32	0.475	WHS_Cogn	cd23	0.079
cd2	cd3	0.472	cd28	WHS_Energia	0.078
cd31	WHS_Animo	0.47	Depresion_1	cd18	0.074
cd28	cd32	0.437	cd23	cd15	0.073
WHS_Vista	WHS_Energia	0.433	cd22	WHS_Cogn	0.07
cd29	cd32	0.433	cd28	cd27	0.064
cd31	cd23	0.425	WHS_Global	cd32	0.062
WHS_Global	cd28	0.41	WHS_Vista	cd32	0.061
cd3	cd17	0.409	cd19	cd18	0.054

Tabla A.16: Resultados test Anova variables sobre perspectiva de salud

B.5. Variables Corazón

Códigos	Etiqueta
d1	¿Ha sentido alguna vez dolor o molestias en el pecho?
d6	¿Alguna vez un médico o doctor le ha dicho que tuvo o que sufrió un infarto al corazón?
d8_1	¿Ha recibido tratamiento por un infarto al corazón? Enunciado: Sí, me operaron
d8_2	¿Ha recibido tratamiento por un infarto al corazón? Enunciado: Sí, me pusieron una malla o stent por la pierna o brazo (angioplastía)
d8_3	¿Ha recibido tratamiento por un infarto al corazón? Enunciado: Sí, me dieron medicamentos
d8_4	¿Ha recibido tratamiento por un infarto al corazón? Enunciado: No
d8_5	¿Ha recibido tratamiento por un infarto al corazón? Enunciado: No sabe/No responde
d9	¿Alguna vez un médico o doctor le ha dicho que tuvo o que sufrió un accidente vascular o trombosis cerebral?
d11	¿Recibió tratamiento por un accidente vascular o trombosis cerebral? (Si ha tenido más de uno especificar que nos referimos al ULTIMO)
d12	¿Ha estado tomando algún medicamento durante las últimas 2 semanas por un accidente vascular o trombosis cerebral?
d13	¿Alguna vez un médico o doctor le ha dicho que tuvo o que sufrió una enfermedad vascular periférica o a las arterias de sus piernas?
af2_1	Infarto o ataque al corazón
af2_2	Accidente vascular, trombosis, o derrame cerebral
af2_3	Arritmia maligna o muerte súbita
RCV	RIESGO CARDIOVASCULAR

Tabla A.17: Variables Corazón

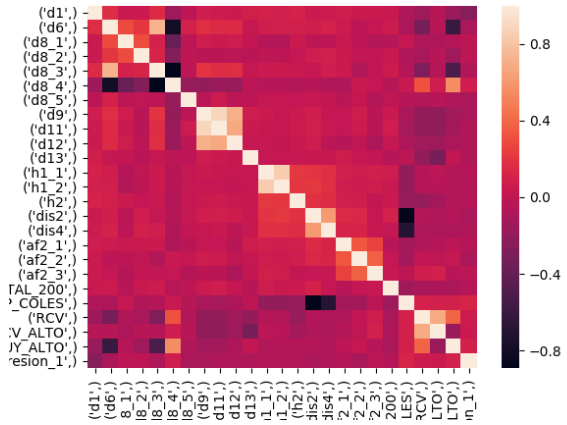


Figura A.5: Correlación de Problemas Cardiovasculares

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
d1	27.600	af2_1	0.509
AUREP_COL	10.267	d6	0.508
RCV_MALTO	9.862	d8_3	0.358
d8_4	5.297	d9	0.309
h1_2	3.031	h2	0.302
RCV_ALTO	1.518	d13	0.276
dis2	1.437	d11	0.134
h1_1	1.404	d12	0.100
af2_2	1.158	d8_5	0.023
COLESTEROL	0.868	af2_3	0.015
RCV	0.544	d8_1	0.001
dis4	0.539	d8_2	0.000

Tabla A.18: Resultados test Chi-Cuadrado de problemas cardiovasculares

Variable	Variable	Resultado
RCV_ALTO	dis4	0.959
d12	RCV	0.475
d9	d6	0.279

Tabla A.19: Resultados test Anova de problemas cardiovasculares

Variable	Resultado	Variable	Resultado
RCV	0.032535878	AUREP_COL	0.002193553
af2_2	0.003081579	d1	3.60E-15
dis2	0.002193553		

Tabla A.20: Resultados test Kolmogórov-Smirnov de problemas cardiovasculares

B.6. Síntomas Respiratorios Crónicos

Variable	Variable
r1	hora vamos a preguntarle por algunos síntomas del pulmón y respiración. ¿Tiene Ud. habitualmente tos sin que esté resfriado(a)?
r5	Tiene Ud. habitualmente flemas (expectoración, desgarro, pollos) que vengan de su pulmón o de los bronquios, o flemas difíciles de sacar sin que esté resfriado(a)?
r9	Ha tenido ud. alguna vez silbidos, pitos o sibilancias en el pecho en los últimos 12 meses?
r12	¿Tiene Ud. alguna incapacidad para caminar, que no sea por una causa de enfermedad del pulmón o del corazón?
r13	¿Ha sentido ud. ahogo o falta de aire cuando camina apurado o en una pequeña subida?
r14	¿Tiene ud. que caminar más lento que personas de su edad en un camino plano debido a falta de aire o ahogo?
r15	¿Tiene ud. que detenerse a tomar aire cuando camina por un camino plano a su paso normal?
r16	¿Tiene ud. que detenerse a tomar aire cuando camina por un camino plano después de andar unos 100 metros, una cuadra o algunos minutos?
r17	¿Su falta de aire es tan fuerte que no lo deja salir de su casa o no lo deja cambiarse de ropa?
r18	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de bronquitis crónica, enfisema pulmonar, enfermedad pulmonar obstructiva crónica o epoc?
sospecha$_a$*sma*	SOSPECHA ASMA
respir_tos	TOSEDOR CRONICO
respir_sibil	SIBILANCIAS
respir_incap_camin	INCAPACIDAD PARA CAMINAR

Tabla A.21: Variables enfermedades respiratorias

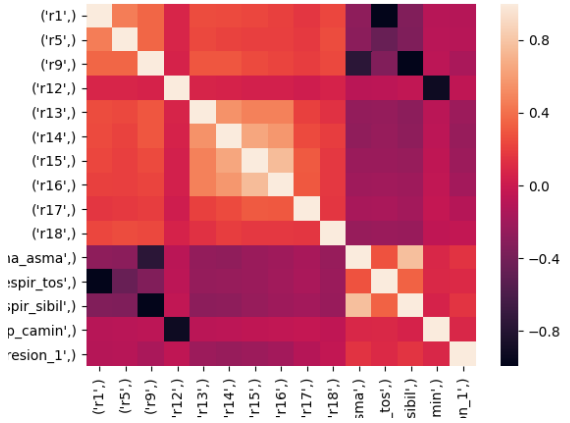


Figura A.6: Correlación de Problemas Respiratorios
Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
respi_sibil	21.495	r16	4.252
sospecha_asma	19.753	r9	2.690
r13	8.857	r1	1.427
respir_tos	8.610	r5	0.894
respir_camin	7.216	r17	0.203
r14	7.175	r12	0.156
r15	5.235	r18	0.082

Tabla A.22: Resultados test Chi-Cuadrado de síntomas respiratorios crónicos

Variable	Resultado	Variable	Resultado
r1	2.9.E-02	r16	2.4E-07
respir_tos	2.9.E-02	r15	4.2E-10
sospecha_asma	1.3.E-02	r13	4.4E-12
r9	4.2.E-04	r14	1.1E-12
respi_sibil	4.2.E-04		

Tabla A.23: Resultados test Kolorov-Smirnov de síntomas respiratorios crónicos

Variable	Variable	Resultado
r16	r14	0.984
r15	r9	0.953
r15	r5	0.524
r5	r9	0.485
r16	r5	0.457
r14	r5	0.444
r14	r1	0.188
r16	r1	0.181
r15	r16	0.174
r15	r14	0.168
r16	r9	0.156
r14	r9	0.150

Tabla A.24: Resultados test Anova de síntomas respiratorios crónicos

B.7. Variables Dieta

Códigos	Etiqueta
die1	¿Con qué frecuencia come pescado o mariscos (cualquier tipo de preparación o presentación?
die2	¿Con qué frecuencia consume leche, queso, quesillo, yogurt, postres de leche o mantequilla o margarina con leche?
die3	En una semana típica, ¿cuántos días come Ud. frutas?
die4	¿Cuántas porciones de frutas come en uno de esos días? (Mostrar TARJETA 19a)
die5	En una semana típica, ¿cuántos días come Ud. verduras, hortalizas o ensaladas de verduras? (No considerar papas ni legumbres)
die6	¿Cuántas porciones de verduras u hortalizas o ensaladas de verduras come en uno de esos días? (No considerar papas ni legumbres) (Mostrar TARJETA 19b)
die7	¿Con qué frecuencia consume usted algún tipo de cereal integral como pan integral, cereal integral o alimentos que contengan harinas integrales?
VitaminaB12d	VITAMINAB12 (DEFICIT)
VitaminaB12e	VITAMINAB12 (EXCESO)
porcdiafrutas	PORCION DIARIA FRUTAS
porcdiaverdu	PORCION DIARIA VERDURAS
porcdiafrutasyverd	PORCION DIARIA FRUTAS y VERDURAS
grsdiafrutas	GRAMOS DIARIOS (FRUTAS) [Porciones de 80gr]
grsdiaverduras	GRAMOS DIARIOS (VERDURAS) [Porciones de 80gr]
grsdiafrutaverdu	GRAMOS DIARIOS (FRUTAS y VERDURAS) [Porciones de 80gr]
frutas7d	frutas7d (Consumo de frutas todos los días)
verdu7d	verdu7d (Consumo de verduras todos los días)
nofrutyverd7d	nofrutyver7d (Consumo de frutas y verduras todos los días)

Tabla A.25: Variables Dietas

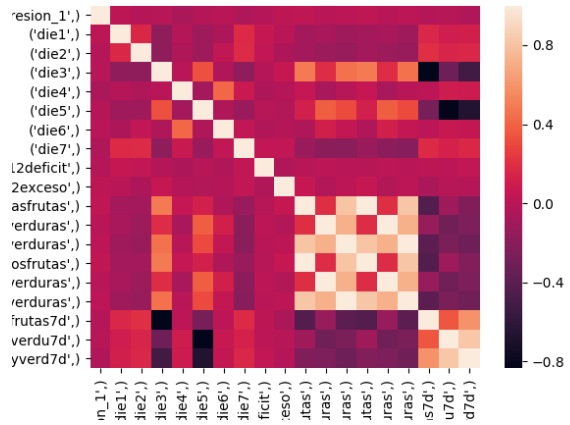


Figura A.7: Correlación de Hábitos Alimenticios

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
nofrutyverd7d	1.353	verdu7d	0.064
die4	1.330	porcdiafrutaverdu	0.027
VitaminaB12d	0.574	grsdiafrutaverdu	0.026
VitaminaB12e	0.570	GR_SAL24	0.017
frutas7d	0.261	die5	0.016
die1	0.136	die3	0.016
die7	0.112	die6	0.001
die2	0.087	porcdiariasverdu	0.001
porcdiafrutas	0.076	grsdiaverdu	0.001
grsdiafrutas	0.075		

Tabla A.26: Resultados test Chi-Cuadrado sobre hábitos alimenticios

Variable	Variable	Resultado
porcdiaverdu	grsdiaverdu	1.000
porcdiafrutas	grsdiafrutas	0.993
porcfrutasverd	grsfrutaverdu	0.986
porcdiaverdu	grsdiafrutas	0.658
grsdiafrutas	grsdiaverdu	0.658
porcdiaverd	porcdiafrutas	0.650
porcdiafruta	grsdiaverdu	0.650

Tabla A.27: Resultados test Anova sobre hábitos alimenticios

B.8. Variables de Problemas Audición y Visión

Códigos	Etiqueta
au1	¿Considera que escucha en forma normal por los dos oídos?
au2	¿Es capaz de seguir un programa de TV a un volumen aceptable para los demás?
au3	¿Es capaz de seguir una conversación de tres o más personas?
prob_aud_1mas	¿TIENE UNO O MÁS PROBLEMAS DE AUDICION? (ENTRE VARIABLES AU1, AU2 Y AU3)
prob_aud_los3	¿TIENE LOS TRES PROBLEMAS DE AUDICION? (ENTRE VARIABLES AU1, AU2 Y AU3)
escucha_normal	ESCUCHA NORMAL (RECODIFICACION AU1)
sigue_prog_TV	SUGUE PROGRAMA TV (RECODIFICACION AU2)
conversa_3pers	SIGUE CONVERSACION DE TRES O MAS PERSONAS (RECODIFICACION AU3)
v1	Ahora le voy a hablar de su vista. ¿Usted tiene lentes?
v2	¿Usted piensa que su vista es? (Si tiene lentes, aclarar que la pregunta se refiere a la visión con esos lentes)
v3	¿Cuándo fue la última vez que cambió sus lentes? Anote año de cambio. (Si nunca ha tenido anote 8888, si no sabe anote 9999)
v5	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de cataratas?
v7	¿Alguna vez se ha operado de cataratas?
v9	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de glaucoma?
v11	¿Alguna vez ha recibido tratamiento por glaucoma?
v12	¿Ha estado usando algún medicamento durante las últimas 2 semanas para el glaucoma?
usa_lentes	USA LENTES (RECODIFICACION V1)
mala_vision	MALA VISION (RECODIFICACION V2)
cataratas	CATARATAS (RECODIFICACION V5)
glaucoma	GLAUCOMA (RECODIFICACION V9)

Tabla A.28: Variables audición

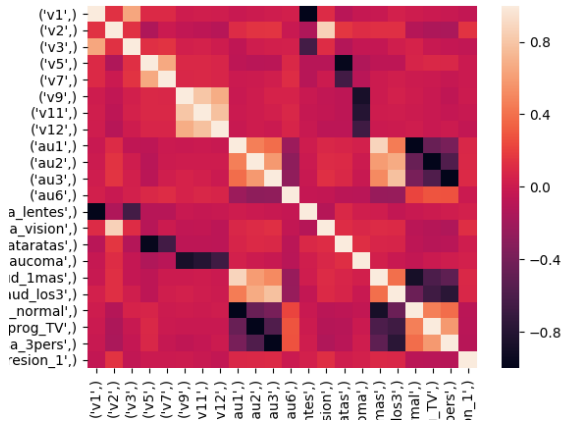


Figura A.8: Correlación de problemas audición y visión

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
au3	8.422	prob_aud_1	4.223
au2	6.775	escucha_normal	3.110
prob_aud_3	6.007	sigue_TV	2.205
au1	4.359	conversa_3	2.017

Tabla A.29: Resultados test Chi-Cuadrado problemas audición

Variable	Resultado	Variable	Resultado
mala_vision	6.247	cataratas	0.433
v2	5.451	v12	0.216
glaucoma	1.754	v11	0.036
v3	1.648	v5	0.014
v1	1.645	v7	0.000
usa_lentes	0.433	v9	0.000

Tabla A.30: Resultados test Chi-Cuadrado de problemas visión

Variable	Resultado	Variable	Resultado
au1	4.8.E-02	prob_aud_1	3.2.E-02
escucha_normal	4.8.E-02		

Tabla A.31: Resultados test Kolmogorov-Smirnov de problemas audición

Variable	Resultado	Variable	Resultado
v2	9.5.E-04	mala_vision	9.5.E-04

Tabla A.32: Resultados test Kolmogorov-Smirnov de problemas visión

Variable	Variable	Resultado
v2	v3	0.959
v9	v11	0.934
v9	v12	0.475
v11	v12	0.387
v1	cataratas	0.279

Tabla A.33: Resultados test Anova de problemas visión

B.9. Variables de Trastornos del Sueño

Códigos	Etiqueta
ts1	¿Le han dicho que ronca todas o casi todas las noches?
ts2	¿Le han dicho que cuando duerme deja de respirar por momentos?
ts3	¿Le cuesta trabajo mantenerse despierto durante el día, por lo menos tres días a la semana?
ts4	¿Despierta usted sintiéndose cansado(a) o casi tan cansado como antes de dormir, por lo menos tres días a la semana?
ts5	En promedio, ¿cuántas horas duerme los días de semana?
ts6	En promedio, ¿cuántas horas duerme los fines de semana?
ts7	Antes de acostarse ¿tiene una sensación irresistible de mover las piernas?
ts8	¿Ha tenido la sensación de pérdida de fuerza en las piernas de forma brusca que se desencadena por situaciones emocionales (risa, alegría, disgusto)?
ts9	¿Ha despertado con la sensación de no poder moverse por algunos segundos?
ts10	¿Ha tenidos sueños desagradables que se viven como muy reales?
trastorno_suegno	TRASTORNO SUEÑO

Tabla A.34: Variables Trastornos del Sueño

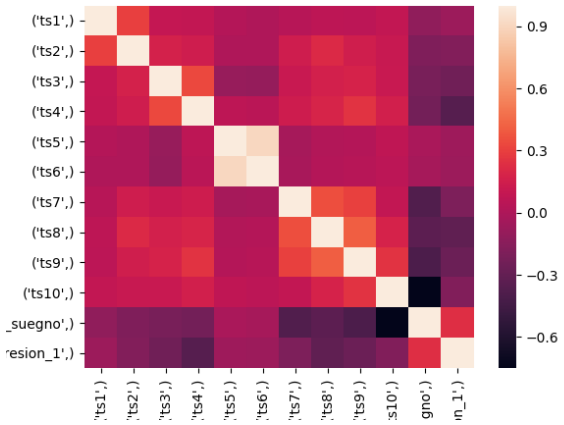


Figura A.9: Correlación de trastornos del sueño

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
ts4	21.495	ts7	4.252
trastorno_sueno	19.753	ts1	2.690
ts3	8.857	ts6	1.427
ts8	8.610	ts5	0.894
ts10	7.216	r17	0.203
ts9	7.175	r12	0.156
ts2	5.235	r18	0.082

Tabla A.35: Resultados test Chi-Cuadrado

Variable	Resultado	Variable	Resultado
ts6	2.4.E-03	ts10	2.16E-09
ts5	2.1.E-03	ts9	3.13E-10
ts7	1.7.E-05	ts8	4.91E-13
ts2	1.2.E-07	trastorno_sueno	7.31E-14
ts3	3.0.E-09	ts4	1.47E-21

Tabla A.36: Resultados test Kolorov-Smirnov de trastornos del sueño

Variable	Variable	Resultado
ts5	ts10	0.740

Tabla A.37: Resultados test Anova de trastornos del sueño

B.10. Dolor Crónico y Fracturas

Códigos	Etiqueta
sm1	En los últimos 7 días, ¿ha tenido algún problema, es decir, dolor, rigidez, sensibilidad (dolor a la presión), hinchazón en sus huesos, músculos, articulaciones o coyunturas?
sm3_1	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Cuello
sm3_2	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Hombro
sm3_3	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Espalda arriba
sm3_4	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Espalda abajo
sm3_5	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Codo
sm3_6	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Muñeca
sm3_7	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Dedos mano
sm3_8	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Cadera
sm3_9	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Rodilla
sm3_10	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Tobillo
sm3_11	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Dedos pie
sm3_12	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: Otra articulación
sm3_13	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: No contesta
sm3_14	¿En qué lugar tuvo el problema? Por favor muestre en este diagrama el o los lugares donde Ud. ha tenido la molestia en los últimos 7 días: No sabe
sm7_1	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? Artritis reumatoidea
sm7_2	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? Artrosis de cadera
sm7_3	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? Artrosis de rodilla
sm7_4	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? Gota
sm7_5	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? Ninguna de las anteriores
sm7_6	¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? No sabe/No responde
o1	Ahora le voy a hablar sobre sus huesos. Alguna vez en su vida, ¿ha tenido alguna fractura o se le ha roto algún hueso?
o2	¿Cuántas veces? (si son más de 10 veces, aclarele al entrevistado que las preguntas que vienen se refieren a las últimas 10 veces)
o5	¿Cuántas veces se ha caído el último año? (si no sabe o no recuerda anote 99)

Tabla A.38: Etiquetas Variables dolor crónico y fracturas

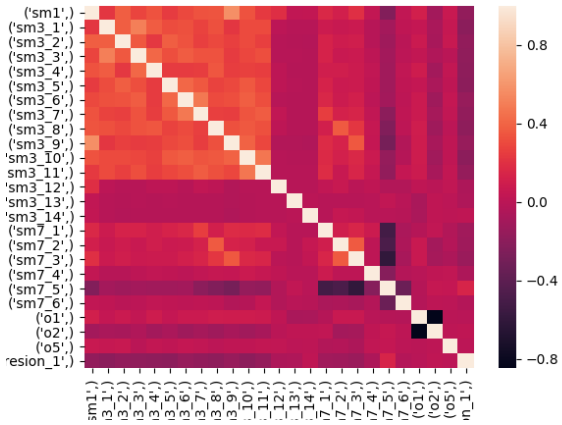


Figura A.10: Correlación de dolor crónico y fracturas

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
sm1	21.495	sm3_11	2.553
sm7s_5	19.753	sm7_3	2.326
sm3_9	8.857	sm7_1	1.186
sm3_2	8.610	sm7_2	1.062
sm3_10	7.216	sm3_12	0.236
sm3_4	7.175	o2	0.163
sm3_8	5.235	sm7_4	0.063
sm3_7	1.404	o5	0.062
sm3_3	1.158	sm7_6	0.027
sm3_5	0.868	o1	0.006
sm3_1	0.544	sm3_13	0.005
sm3_6	0.539	sm3_14	0.001

Tabla A.39: Resultados test Chi-Cuadrado dolor crónico y fracturas

Variable	Resultado	Variable	Resultado
sm7_3	3.0.E-02	sm3_8	3.60E-05
sm3_11	1.8.E-02	sm3_2	3.34E-05
sm3_6	4.8.E-03	sm3_4	2.90E-05
sm3_1	1.8.E-03	sm3_10	2.09E-05
sm3_5	1.2.E-03	o5	1.69E-05
sm3_3	7.1.E-04	sm3_9	1.07E-05
sm7_5	5.2.E-04	sm1	4.18E-07
sm3_7	1.3.E-04		

Tabla A.40: Resultados test Kolorov-Smirnov variables dolor crónico y fracturas

Variable	Variable	Resultado
sm3_7	sm3_10	0.954
sm3_7	sm3_4	0.909
sm3_10	sm3_4	0.864
sm3_1	sm7_1	0.647
sm3_1	sm7_2	0.637
sm3_11	sm3_5	0.615
sm3_11	sm7_1	0.604
sm3_3	sm7_3	0.513
sm3_3	sm3_6	0.377
sm7_2	sm7_1	0.353
sm3_1	sm3_11	0.329
sm3_5	sm7_1	0.307
sm3_5	sm7_3	0.244
sm3_11	sm7_2	0.148
sm3_1	sm3_5	0.139
sm3_6	sm7_3	0.124
sm3_11	sm7_3	0.095
sm3_3	sm3_5	0.069
sm3_5	sm7_2	0.051
sm3_6	sm3_10	0.050

Tabla A.41: Resultados test Anova variables dolor crónico y fracturas

B.11. Alcoholismo y Tabaquismo

Códigos	Etiqueta
ta1	¿Ha fumado por lo menos 100 cigarrillos en toda su vida?
ta2	ctualmente fuma Ud. cigarrillos?
ta3	mo promedio, ¿cuántos cigarrillos fumó al día durante los últimos 30 días? (Si no sabe o no esta seguro anote 99)
ta4	uánto tiempo pasa desde que se despierta hasta que se fuma su primer cigarrillo?
ta5	qué edad comenzó a fumar?
ta6	esea Ud. dejar de fumar?
ta7	uándo fue la última vez que fumó, es decir desde cuándo no se fuma por lo menos un cigarrillo diariamente?
ta9	formularemos algunas preguntas relacionadas con el hábito de fumar en su hogar. De lunes a jueves, ¿cuánto tiempo como promedio diario acostumbra estar Ud. en ambientes cargados de humo de tabaco fuera de su casa o su trabajo?
ta10	e viernes a domingo, ¿cuánto tiempo como promedio diario acostumbra estar Ud. en ambientes cargados de humo de tabaco fuera de su casa o su trabajo?
ta11	n este hogar, ¿se permite fumar dentro de la casa?
ta12	Alguna persona fuma habitualmente en la vivienda?
ta13	n su lugar de estudio o de trabajo, ¿está Ud. expuesto al humo de cigarrillo?
m7p1	¿Ha consumido alguna bebida que contenga alcohol, en los últimos 12 meses?
m7p2	¿En los últimos 12 meses ¿con qué frecuencia ha tomado al menos una bebida alcohólica?
m7p3	Cuando bebe alcohol, ¿Cuántos vasos (tragos) suele tomar en promedio al día?
m7p9	¿Qué tan seguido toma usted alguna bebida alcohólica?
m7p10a	. ¿Cuántos tragos suele tomar usted en un día típico de consumo de alcohol?
POH_1_SEM	CONSUMO ALCOHOL ÚLTIMA SEMANA
POH_CONS_DIA	CONSUMO DIARIO DE ALCOHOL
ML_VA_SEM	ML VASO SEMANA (ALCOHOL)
POH_EX_U_M	PREVALENCIA EXCESO OH ULTIMO MES
NUM_VA_DIA	NUMERO DE VASOS DIA (ALCOHOL)
ta1_ad	¿HA FUMADO MÁS DE 100 CIGARRILLOS EN SU VIDA?
cat_tabaq3	CATEGORÍA DE TABAQUISMO
fum_act30d	fum_act30d (Fumador del último mes)

Tabla A.42: Variables Tabaco y Alcoholismo

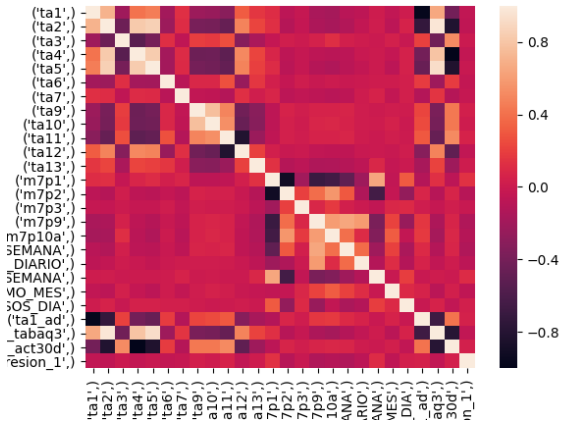


Figura A.11: Correlación de alcoholismo y tabaquismo

Fuente: Elaboración Propia

Test Chi-Cuadrado

Variable	Resultado	Variable	Resultado
ta11	2.000	cat_tabaq3	0.347
fum_act30d	1.896	ta4	0.211
ta1_ad	1.889	ta10	0.072
ta9	0.756	ta5	0.051
ta6	0.664	ta7	0.016
ta1	0.548	ta3	0.012
ta2	0.476	ta13	0.001
ta12	0.461		

Tabla A.43: Resultados test Chi-Cuadrado de tabaquismo

Variable	Resultado	Variable	Resultado
POH_1_SEM	12.551	m7p10a	2.636
m7p1	11.482	$POH_EX\_U\_M$	2.020
m7p9	6.048	ML_VASO_SEM	0.659
POH_CONS_DIA	4.944	NUM_VA_DIA	0.146
m7p2	4.032	m7p3	0.108

Tabla A.44: Resultados test Chi-Cuadrado de alcoholismo

Variable	Resultado	Variable	Resultado
POH_1_SEM	0.030	m7p1	0.001
m7p9	0.018	m7p2	0.001
m7p10a	0.005	m7p3	0.001
ML_VA_SEM	0.002		

Tabla A.45: Resultados test Kolorov-Smirnov de alcoholismo

Variable	Variable	Resultado
ta9	ta10	0.969
ta5	ta12	0.318
ta2	cat_tabaq3	0.389

Tabla A.46: Resultados test Anova de tabaquismo

Variable	Variable	Resultado
m7p9	POH_1_SEM	0.105

Tabla A.47: Resultados test Anova de alcoholismo

B.12. Diabetes

Códigos	Etiqueta
dataHR2p2	El entrevistado es diabético?
di1	tes de esta entrevista ¿alguna vez un profesional de la salud le ha medido (tomado, chequeado) el azúcar en la sangre?
di3	lguna vez un doctor, una enfermera u otro profesional de la salud le ha dicho a Ud. que ha tenido o que tiene o que padece de Diabetes (azúcar alta en la sangre)?
di5	lguna vez ha hecho algún programa, tratamiento o cambio en el estilo de vida para la diabetes o azúcar alta en la sangre?
di6	n estos momentos está llevando o haciendo algún programa, tratamiento o cambio en el estilo de vida (dieta, ejercicios, bajar de peso) para mantener controlada su diabetes/glicemia/azúcar?
AR_DIABE_C	AUTOREPORTE DE DIABETES CORREGIDO
diabetesinsulino	diabetesinsulino (Diabético insulino dependiente)
glubasalayun88	glubasalayun88 (Glucosa basal en ayuno con punto de corte 88)

Tabla A.48: Variables Tabaco y Alcoholismo

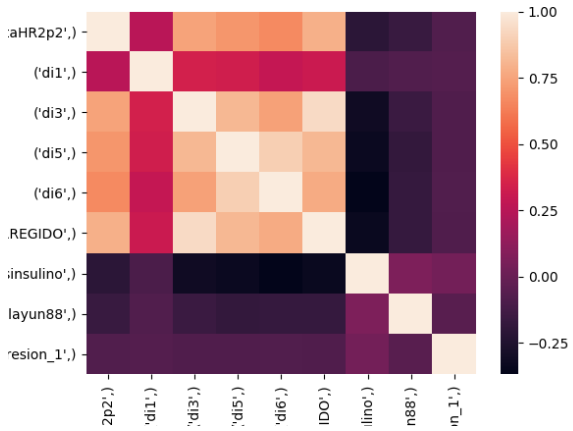


Figura A.12: Correlación de sobre diabetes

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
di1	1.940	glubasalayun88	0.950
diabetesinsulino	1.405	di6	0.909
AR_DIABE_C	1.224	di3	0.795
dataHR2p2	1.007	di5	0.671

Tabla A.49: Resultados test Chi-Cuadrado

Variable	Variable	Resultado
di6	dataHR2p2	0.794
AR_DIABE_C	dataHR2p2	0.136
AR_DIABE_CO	di6	0.080

Tabla A.50: Resultados test Anova

B.13. Problemas Digestivo

Códigos	Etiqueta
m9p1	. En los últimos 5 años ¿Ha tenido un dolor abdominal que se inicie después de las comidas?
m9p10	En los últimos 3 meses ¿Ha notado sangre roja, fresca, al obrar o defecar?
m9p11	En los últimos 3 meses ¿Ha notado cambios permanentes en la frecuencia o consistencia de sus deposiciones (fecas, heces, caca, al obrar o dar del cuerpo)?
m9p12	¿Le han hecho alguna vez en su vida una endoscopía digestiva alta?
hda	HEMORRAGIA DIGESTIVA ALTA (EN ÚLTIMOS 3 MESES)

Tabla A.51: Variables Digestivo

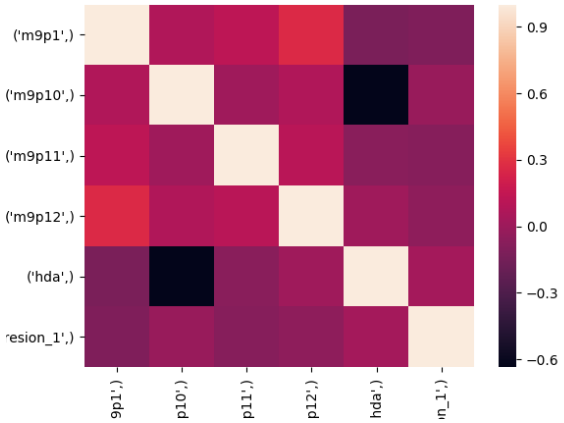


Figura A.13: Correlación de problemas digestivos

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
m9p1	2.807	m9p12	0.521
m9p11	1.428	m9p10	0.006
hda	1.216		

Tabla A.52: Resultados test Chi-Cuadrado de problemas digestivos

Variable	Resultado	Variable	Resultado
m9p11	2.2E-02	m9p1	1.6E-03

Tabla A.53: Resultados test Kolmogorov-Smirnov problemas digestivos

B.14. Otros

Códigos	Etiqueta
e1	eer: En esta encuesta le vamos a preguntar sobre algunos problemas de salud. Nos gustaría empezar con algunas preguntas sobre su memoria. ¿Ud. diría que su memoria es excelente, muy buena, buena, regular o mala?
m5p1	Peso (Kilos) [Min 30 - Max 150]
m5p2	Talla (Centímetros) [Min 80 - Max 210]
m5p3	Circunferencia de Cintura (Centímetros) [Min 30 - Max 200]
m5p4	m5p4. Circunferencia de Cuello (Centímetros) [Min 20 - Max 100]
m8p1	m8p1. ¿Alguna vez en la vida, ha tenido relaciones sexuales?
m8p6	m8p6. ¿Con cuántas personas ha tenido relaciones sexuales en el transcurso de su vida?
m8p7	m8p7. ¿Ha tenido relaciones sexuales durante los últimos 12 meses?
sd27	Alguna vez un médico o doctor le ha dicho que tiene o que padece de depresión?
pe11	e estimado por encuestador a partir de t10a, t10b,..., t10k

Tabla A.54: Variables Varias

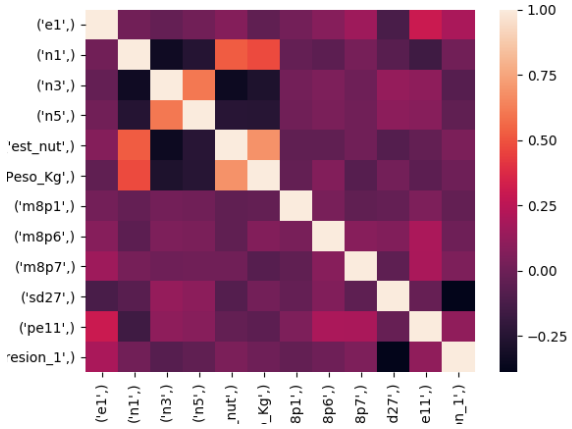


Figura A.14: Correlación de sobre diabetes

Fuente: Elaboración Propia

Variable	Resultado	Variable	Resultado
sd27	15.383	n5	0.196
pe11	8.001	n1	0.025
e1	2.529	m8p6	0.015
n3	1.179	Peso_Kg	0.000
m8p1	0.781	ta7	0.016
est_nut	0.480	ta3	0.012
m8p7	0.423	ta13	0.006

Tabla A.55: Resultados test Chi-Cuadrado otras

Variable	Resultado	Variable	Resultado
n3	4.8.E-02	sd27	4.6E-24
e1	1.7.E-07		

Tabla A.56: Resultados Kolmogorov Smirnov otras

Variable	Variable	Resultado
Peso_Kg	n3	0.137

Tabla A.57: Resultados test Anova otras

C. Resultados validación con ENS 2009-2010 con otros métodos de balance

C.1. Modelo basado en variables de riesgo población chilena

SMOTE

Modelo	Accuracy	Precision	Recall	AUC	F1	Log-loss
LOGIT	0.835	0.855	0.415	0.559	0.843	0.369
KNN	0.858	0.840	0.456	0.591	0.850	0.338
TREE	0.838	0.817	0.416	0.552	0.829	0.396
NB	0.838	0.832	0.419	0.558	0.836	0.403
SVM	0.827	0.870	0.404	0.552	0.846	0.366
FOREST	0.844	0.847	0.430	0.571	0.845	0.350

Tabla A.58: resultados modelo de Variables de Riesgo-SMOTE

Cluster Centroids

Modelo	Accuracy	Precision	Recall	AUC	F1	Log-loss
LOGIT	0.872	0.79	0.486	0.601	0.835	0.341
KNN	0.868	0.73	0.475	0.577	0.810	0.367
TREE	0.840	0.84	0.423	0.563	0.840	0.434
NB	0.881	0.77	0.510	0.614	0.834	0.365
SVM	0.872	0.78	0.486	0.598	0.832	0.345
FOREST	0.870	0.77	0.481	0.592	0.828	0.363

Tabla A.59: resultados modelo de Variables de Riesgo-Clusters

C.2. Modelo basado determinantes sociales y psicológicos

SMOTE

Modelo	Accuracy	Precision	Recall	AUC	F1	Log-loss
LOGIT	0.775	0.710	0.314	0.436	0.747	0.480
KNN	0.775	0.679	0.363	0.473	0.757	0.442
TREE	0.775	0.626	0.432	0.511	0.756	0.499
NB	0.775	0.779	0.327	0.460	0.778	0.506
SVM	0.775	0.687	0.316	0.433	0.740	0.467
FOREST	0.775	0.718	0.364	0.483	0.772	0.442

Tabla A.60: resultados determinantes sociales y psicológicos-SMOTE

Cluster Centroids

Modelo	Accuracy	Precision	Recall	AUC	F1	Log-loss
LOGIT	0.814192	0.694656	0.364	0.47769	0.762754	0.465376
KNN	0.804855	0.679389	0.347656	0.459948	0.750865	0.497878
TREE	0.577031	0.877863	0.208333	0.33675	0.706484	0.634308
NB	0.817927	0.70229	0.370968	0.485488	0.768166	0.467903
SVM	0.821662	0.709924	0.378049	0.493369	0.773579	0.469969
FOREST	0.666667	0.839695	0.246637	0.381282	0.741124	0.633148

Tabla A.61: resultados determinantes sociales y psicológicos-Clusters

C.3. Modelo basado en estudio de varaibles de riesgo

SMOTE

Modelo	Accuracy	Precision	Recall	AUC	F1	Log-loss
LOGIT	0.835	0.855	0.415	0.559	0.843	0.369
KNN	0.858	0.840	0.456	0.591	0.850	0.338
TREE	0.838	0.817	0.416	0.552	0.829	0.396
NB	0.838	0.832	0.419	0.558	0.836	0.403
SVM	0.827	0.870	0.404	0.552	0.846	0.366
FOREST	0.844	0.847	0.430	0.571	0.845	0.350

Tabla A.62: resultados basado en estudio de varaibles de riesgo-SMOTE

Cluster Centroids

Modelo	Accuracy	Precision	Recall	AUC	F1	Log-loss
LOGIT	0.872	0.786	0.486	0.601	0.835	0.341
KNN	0.868	0.733	0.475	0.577	0.810	0.367
TREE	0.840	0.840	0.423	0.563	0.840	0.434
NB	0.881	0.771	0.510	0.614	0.834	0.365
SVM	0.872	0.779	0.486	0.598	0.832	0.345
FOREST	0.870	0.771	0.481	0.592	0.828	0.363

Tabla A.63: resultados basado en estudio de varaibles de riesgo-Clusters

D. Resultados Otros Modelos

D.1. Modelo con variables Socio-Demograficas

Posteriormente de la aplicación y evaluación de variables demográficas en la sección anterior además de diferente bibliografía que habla de ciertas variables socio-demográficas que son factores de riesgo[29] o tienen cierta correlación con la aparición de depresión en adultos mayores, por lo tanto se formulo un modelo con aquellas variables que según los resultados en las tablas (poner tablas sección anterior) tenían mayor relevancia.

EDAD	EDAD
NEDU	NEDU
ZONA	ZONA
SEXO	SEXO
nper	Número de persona en el hogar
ns12cod	¿Qué ocupación o tipo de trabajo desempeña actualmente o desempeñaba si está? (Codificación)
ns19_1	¿Tiene usted actualmente en uso y funcionamiento alguno de los siguientes bienes? : Calefont o sistema de calentamiento de aguas
ns19_3	¿Tiene usted actualmente en uso y funcionamiento alguno de los siguientes bienes? : Computador o notebook
ns19_5	¿Tiene usted actualmente en uso y funcionamiento alguno de los siguientes bienes? : Refrigerador
ns19_6	¿Tiene usted actualmente en uso y funcionamiento alguno de los siguientes bienes? : Automóvil de uso particular
ns25	¿Cuántas duchas tiene esta vivienda?
ns26	Tipo de vivienda:
ns27	En la cubierta del techo predomina:
ns28	El material predominante en las paredes exteriores es:
ns29	El material predominante en el piso es:

Tabla A.64: Variables de Modelo Socio-Demográficos

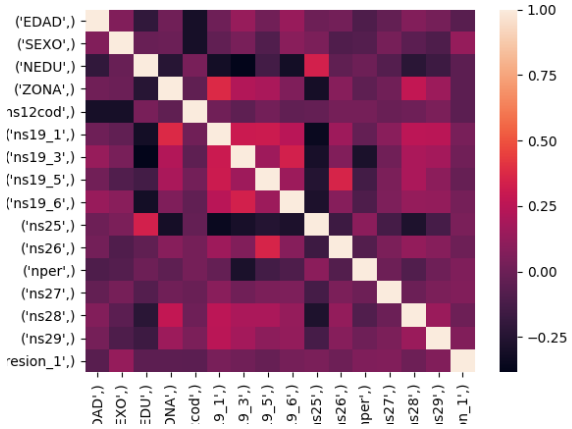
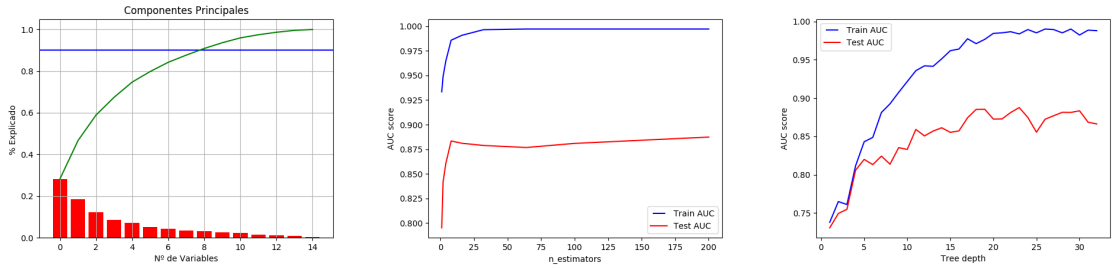


Figura A.15: Correlación de Mejores Variables

Fuente: Elaboración Propia



(a) Resultado Resultado PCA (b) Test N Estimadores (c) Test Máxima Profundidad

Figura A.16: Resultados Modelo Con Variables Riesgo Población.
Fuente: Elaboración Propia

Modelo	Accuracy	Precision	Recall	AUC	F1
LOGIT	0.7633	0.7420	0.8079	0.8433	0.7733
KNN	0.7660	0.7157	0.8856	0.8598	0.7907
TREE	0.7489	0.7636	0.7224	0.7840	0.7415
NB	0.7580	0.6994	0.9054	0.8502	0.7889
SVM	0.7574	0.7437	0.7851	0.8466	0.7635
FOREST	0.7649	0.7407	0.8596	0.8657	0.7897

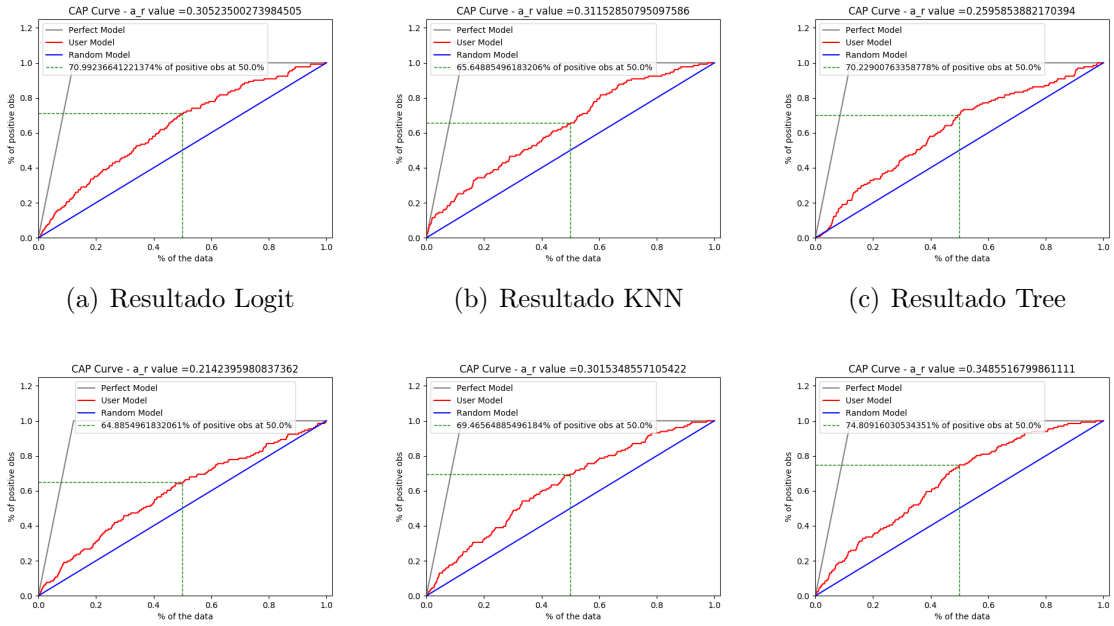
Tabla A.65: Cross validation modelo Variables de Socio-Demograficas

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.697	0.458	0.191	0.270	0.594	0.568
KNN	0.708	0.435	0.193	0.267	0.590	0.538
TREE	0.725	0.374	0.188	0.250	0.574	0.571
NB	0.635	0.496	0.167	0.250	0.575	0.600
SVM	0.702	0.412	0.182	0.253	0.577	0.577
FOREST	0.722	0.397	0.192	0.259	0.582	0.530

Tabla A.66: Validación modelo Variables de Socio-Demográfico ENS 2009-2010

D.2. Resultados modelo factores de riesgo adulto mayor

Posteriormente de la aplicación y evaluación de variables demográficas, de enfermedades crónicas y dolores crónicos en la sección anterior además de diferente bibliografía que habla de ciertas variables relacionadas a estas que son factores de riesgo[65] según un estudio español se diseño el siguiente modelo.



(a) Resultado Logit (b) Resultado KNN (c) Resultado Tree

(d) Resultado NB (e) Resultado SVM (f) Resultado Forest

Figura A.17: Resultados Modelo Socio-Demográfico.
Fuente: Elaboración Propia

Variable	Etiqueta
EDAD	EDAD
NEDU	NEDU
SEXO	SEXO
cd5	cd5 Durante las últimas 4 semanas, ¿con qué frecuencia ha tenido alguno de los siguientes problemas en su trabajo o en sus actividades cotidianas a causa de su SALUD FISICA... ¿Hizo menos de lo que hubiera querido hacer?
cd6	cd6 Durante las últimas 4 semanas, ¿con qué frecuencia ha tenido alguno de los siguientes problemas en su... a causa de su SALUD FISICA... ¿Tuvo que dejar de hacer algunas tareas en su trabajo o en sus actividades cotidianas?
cd7	cd7 ¿Con qué frecuencia ha tenido alguno de los siguientes problemas en su trabajo o en sus actividades cotidianas a causa de algún problema EMOCIONAL?
cd8	cd8 ¿Hizo su trabajo u otra actividad con menos cuidado que el de costumbre por algún problema emocional?
cd15	cd15 En general, durante los últimos 30 días, ¿qué grado de dificultad ha tenido para realizar las tareas del trabajo y del hogar?
sm7_1	sm7_1 ¿Alguna vez un médico o doctor le ha dicho que tiene o que padece de ... ? Artritis reumatoidea
o1	o1 Ahora le voy a hablar sobre sus huesos. Alguna vez en su vida, ¿ha tenido alguna fractura o se le ha roto algún hueso?

Tabla A.67: Variables modelo factores de riesgo adulto mayor

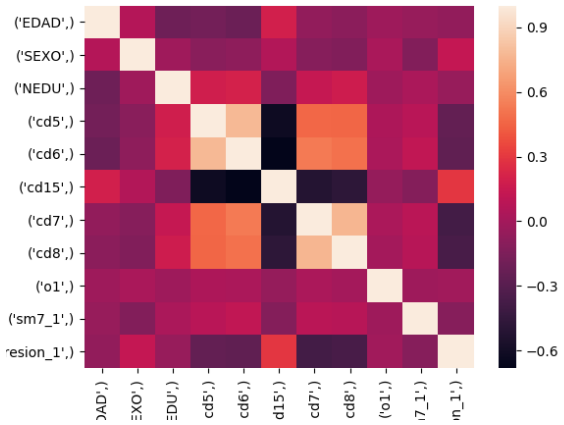
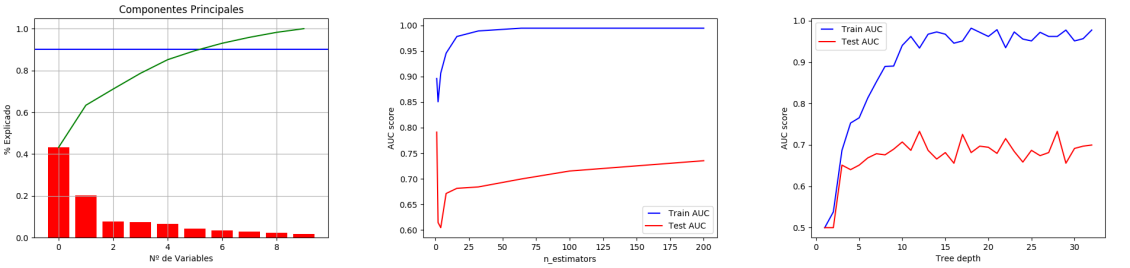


Figura A.18: Correlación de Mejores Variables

Fuente: Elaboración Propia



(a) Resultado PCA PCA (b) Test N Estimadores (c) Test Máxima Profundidad

Figura A.19: Resultados modelo factores de riesgo adulto mayor.

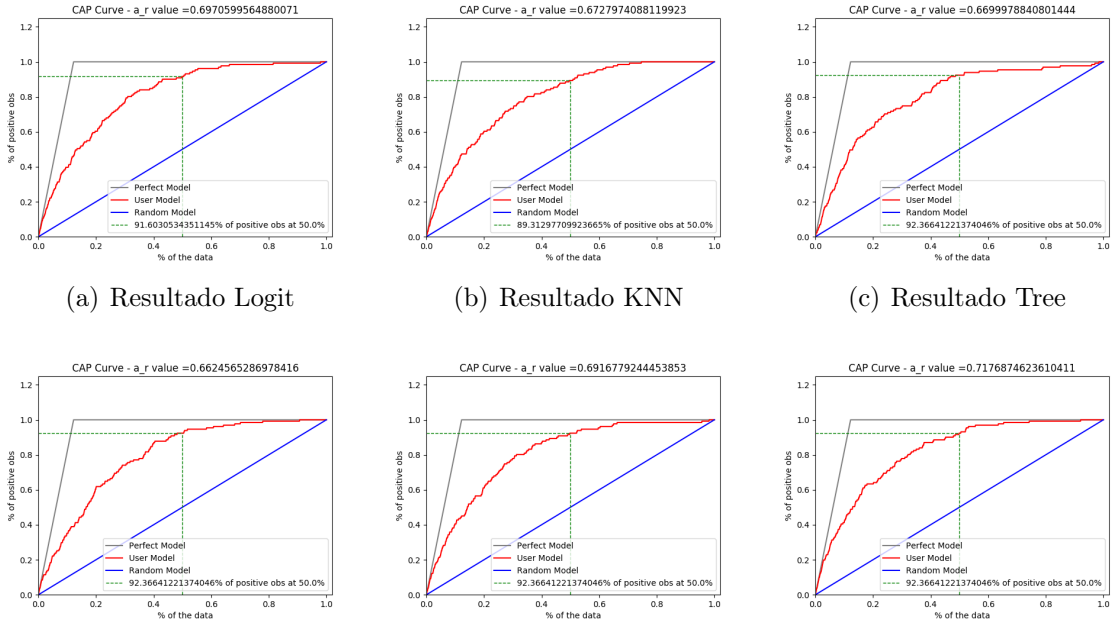
Fuente: Elaboración Propia

Modelo	Accuracy	Precision	Recall	F1	AUC
LOGIT	0.8948	0.8076	0.4160	0.8934	0.5350
KNN	0.8846	0.8047	0.3207	0.8433	0.4413
TREE	0.8835	0.7296	0.3959	0.8252	0.4931
NB	0.8394	0.4834	0.6819	0.8854	0.5569
SVM	0.8846	0.8621	0.3007	0.8909	0.4291
FOREST	0.8903	0.7627	0.3150	0.8792	0.4188

Tabla A.68: Cross Validation Modelo factores de riesgo adulto mayor

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.871	0.450	0.472	0.461	0.690	0.297
KNN	0.868	0.435	0.460	0.447	0.682	0.307
TREE	0.896	0.366	0.632	0.464	0.668	0.294
NB	0.835	0.496	0.369	0.423	0.689	0.319
SVM	0.866	0.427	0.448	0.438	0.677	0.302
FOREST	0.872	0.443	0.475	0.458	0.687	0.288

Tabla A.69: Validacion modelo factores de riesgo adulto mayor ENS 2009-2010



(a) Resultado Logit (b) Resultado KNN (c) Resultado Tree

(d) Resultado NB (e) Resultado SVM (f) Resultado Forest

Figura A.20: Resultados Modelo Socio-Demográfico.
Fuente: Elaboración Propia

D.3. Modelo basado en síntomas de enfermedades crónicas

Posteriormente de la aplicación y evaluación variables asociadas a enfermedades crónicas en la sección anteriores, además de diferente bibliografía que habla de estas variables relacionadas a estas que son factores de riesgo, junto con lo propagado de ciertas enfermedades entre los adultos mayores se diseño el siguiente modelo.

EDAD	EDAD
SEXO	SEXO
dataHR2p2	p2.¿El entrevistado es diabético?
cd31	cd31 En general, durante los últimos 30 días, ¿en qué grado se ha sentido triste, decaído o deprimido?
d1	d1 ¿Ha sentido alguna vez dolor o molestias en el pecho?
v2	v2 ¿Usted piensa que su vista es? (Si tiene lentes, aclarar que la pregunta se refiere a la visión con esos lentes)
au1	au1 ¿Considera que escucha en forma normal por los dos oídos?
r13	r13 ¿Ha sentido ud. ahogo o falta de aire cuando camina apurado o en una pequeña subida?
ts4	ts4 ¿Despierta usted sintiéndose cansado(a) o casi tan cansado como antes de dormir, por lo menos tres días a la semana?
m9p11	m9p11. En los últimos 3 meses ¿Ha notado cambios permanentes en la frecuencia o consistencia de sus deposiciones (fecas, heces, caca, al obrar o dar del cuerpo)?

Tabla A.70: Variables Modelo en síntomas de enfermedades crónicas

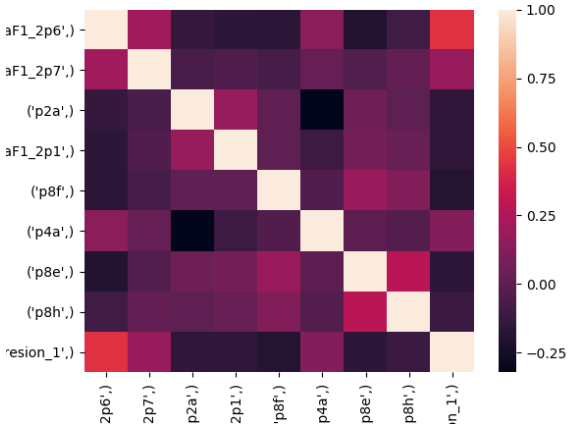
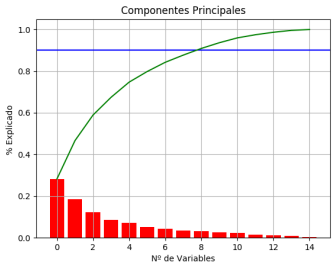


Figura A.21: Correlación de Variables Apreciativas

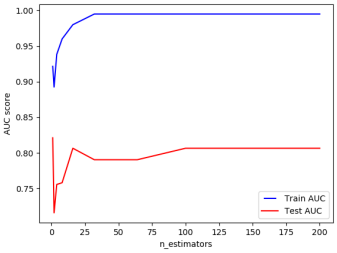
Fuente: Elaboración Propia

Modelo	Accuracy	Precision	Recall	F1	AUC
LOGIT	0.9348	0.9187	0.6044	0.9257	0.7215
KNN	0.9217	0.9173	0.5012	0.8825	0.6409
TREE	0.9315	0.8109	0.6840	0.8441	0.7336
NB	0.9087	0.6644	0.7576	0.9219	0.7020
SVM	0.9315	0.9798	0.5322	0.9244	0.6833
FOREST	0.9228	0.9200	0.4864	0.9127	0.5855

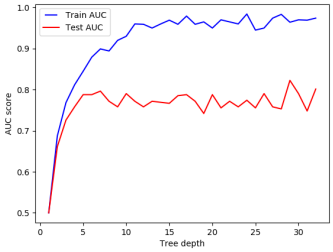
Tabla A.71: Cross Validation Modelo en síntomas de enfermedades crónicas



(a) Resultado PCA



(b) Test N Estimadores

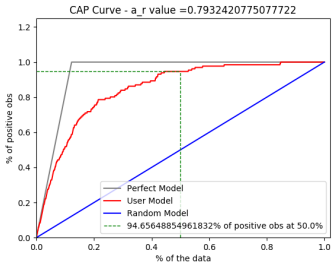


(c) Test Máxima Profundidad

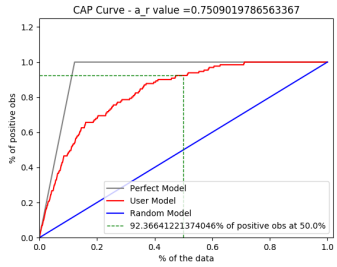
Figura A.22: Resultados modelo en síntomas de enfermedades crónicas
Fuente: Elaboración Propia

Modelo	Accura	Preci	Recall	F1	AUC	Log-loss
LOGIT	0.900	0.641	0.583	0.611	0.789	0.255
KNN	0.889	0.534	0.547	0.541	0.736	0.305
TREE	0.895	0.687	0.559	0.616	0.806	0.261
NB	0.882	0.695	0.514	0.591	0.802	0.272
SVM	0.888	0.588	0.538	0.562	0.759	0.269
FOREST	0.902	0.702	0.582	0.637	0.816	0.253

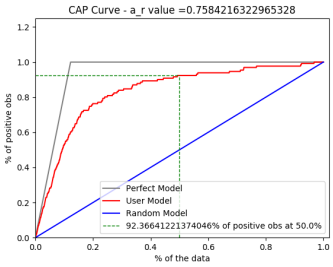
Tabla A.72: Validacion modelo en síntomas de enfermedades crónicas ENS 2009-2010



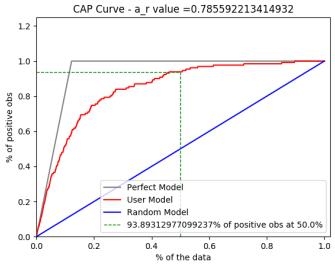
(a) Resultado Logit



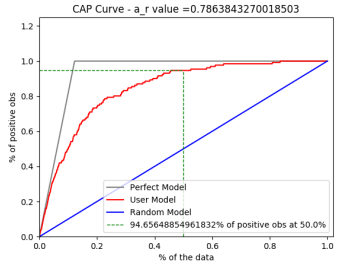
(b) Resultado KNN



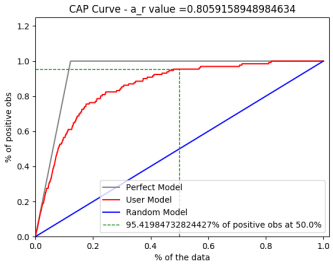
(c) Resultado Tree



(d) Resultado NB



(e) Resultado SVM



(f) Resultado Forest

Figura A.23: Resultados modelo en síntomas de enfermedades crónicas ENS 2009-2010
Fuente: Elaboración Propia