UNIVERSIDAD DE CHILE
FACULTAD DE ECONOMÍA Y NEGOCIOS
ESCUELA DE ECONOMÍA Y ADMINISTRACIÓN

Predicción del Potencial de Creación de Empleo en Planes de Negocio Mediante Herramientas de Data Analytics

Seminario para optar al Título de
Ingeniero Comercial Mención Administración

Participantes: Marco Zaror

Profesor Guía:Ph.DDavid Díaz Solís

Director de Escuela Economía y Administración: Claudio Bravo Ortega

Santiago, Julio de 2014

Contenido

1. Introducción 3

2. Revisión literaria 6

2.1 Teoría económica para la creación de empleos 6

2.1.1 Ontologías del emprendimiento 10

2.2 Data Mining 20

2.2.1 Definición 20

2.2.2 Aplicaciones Generales y Evolución del Data Mining 28

2.2.3 Text Mining 34

2.2.4 Aplicaciones de Text Mining en la Economía 36

2.3 Posicionamiento y Contribución 38

2.4 Preguntas de investigación 39

3. Metodología 41

3.1 Datos utilizados y preparación 41

3.2 Metodología CRISP DM 43

3.3 Metodología específica 52

3.4 Algoritmos a utilizar 56

3.4.1 Decision Trees 57

4. Resultados 63

4.1 Resultados obtenidos 63

4.2 Respuestas a preguntas de investigación 89

5. Conclusiones e investigación futura 93

6. Referencias 96

7. Anexos 99

1. Introducción

Al momento de evaluar el rendimiento de un país en particular, es normal evaluar una serie de factores. Entre estos, la creación de empleos es uno de los principales a considerar debido a que cuenta con una serie de ventajas dentro de las que se enumeran:

- Si aumentan los trabajadores, disminuye el número de personas que no recibe ingresos lo que aumenta la probabilidad de disminuir la pobreza.

- Una disminución en el desempleo implica que más personas se encuentran produciendo para el país lo que podría significar un aumento en la producción de bienes y servicios.

- Una mayor oferta de trabajo, estimula la competencia entre las personas que demandan empleo, lo que significa una mejor calidad de trabajadores para el futuro.

Presentada la importancia que reviste para un país la creación de empleos, se debe analizar quienes son los que están generando puestos de trabajo en nuestro país. Esto permitiría la creación de planes o estrategias para fomentar este tan complejo tópico. Una de las creadoras de puestos de trabajo por excelencia son las nuevas empresas, ya que debido a que están recién comenzando es bastante probable que comiencen a necesitar gente de manera exponencial. Con respecto a la relación existente entre el número de firmas y la creación de empleo, Ayyagari, Demirguc-Kunt&Maksimovic (2011) [4] documentan que las grandes empresas (> 99 empleados) son las que más contribuyen al empleo agregado, con un 54,6% del total de empresas según sus tamaños, mientras que las más pequeñas (< 19 empleados) sólo aportan un 16,5% en este ítem. Sin embargo, si analizamos el aporte que hacen las SMEs (pequeñas y medianas empresas con menos de 99 empleados) y lo comparamos con el aporte de las grandes, vemos que esta relación se compensa de manera bastante clara en donde las SMEs aportan un 45,6% y las grandes el ya mencionado 54,6%.

Existen algunos concursos, que preguntan o intentan determinar el potencial de empleo con que cuenta un plan de negocio, esto lo hacen por medio de jueces que leen cada documento y luego entregan un valor numérico que representa la capacidad de la nueva firma para contratar personas.

Esto último presenta variadas desventajas dentro de las que contamos:

- El tiempo que significa la lectura de cada plan de negocio uno a uno es enorme, generando grandes ineficiencias en el sistema.

- El juicio es subjetivo, debido a que puede depender del estado de ánimo de cada jurado, a experiencias pasadas con la industria a la cual el plan de negocio apunta o simplemente a la personalidad de la persona.

En la actualidad, no existe una literatura muy basta sobre la predicción de la potencial creación de empleos de nuevas empresas hacia el futuro, las investigaciones que más se relacionan a esto hacen relación al emprendimiento o a la creación de empresas. Un punto importante, es que estas relaciones son en retrospectiva, o sea que se analiza el tema hacia atrás (cuanto empleo se generó, no cuanto se generará). Esto presenta una serie de desventajas, debido a que existe una gran cantidad de empresas que pudieron haber generado una gran cantidad de empleos, pero que dado que no recibieron apoyo a tiempo, terminaron cerrando y obviamente sin entregar estos empleos potenciales. Por su parte, técnicas de Data y Text Mining han sido utilizadas para aplicaciones económicas, pero no en temas relacionados con la creación de empleo. Esta tesis busca aportar a cerrar esta brecha, buscando predecir el potencial que tiene un plan de negocio para crear empleo por medio de técnicas de Data y Text Mining. Además busca explorar sobre las razones por las cuales cada plan de negocio entregará empleo, generando un aporte no sólo nuevo sino que más amplio a la literatura existente.

Los resultados mostraron que el modelo obtuvo un rendimiento superior en casi un 20% al que podría haber obtenido un modelo aleatorio y que además, existían zonas en donde contaba con una alta seguridad de que su predicción era correcta, obteniendo resultados notablemente mejores en esas áreas (100% de precisión en algunos casos). Otro punto importante fue que se descubrieron las principales categorías o conceptos claves mencionados en los planes de negocios que permiten predecir un alto potencial de creación de empleo. Aquí resaltan términos

como “*Reduce Unemployment*” y “*Many Job Opportunity*”. Una categoría que ayudó a discriminar bastante entre los planes fue la relacionada al concepto de comunidad, en donde una gran cantidad de planes de negocio se enfocaban en entregar empleo para su zona o ciudad. Finalmente se testeó la capacidad de este modelo para adaptarse a otro concurso de planes de negocio, el concurso elegido fue el de YouWin. El experimento demostró que la extrapolación disminuye los niveles de precisión, pero que siguen existiendo zonas en donde el modelo tiene una gran seguridad de su predicción y en las que el porcentaje de error es muy bajo.

De esta forma, obtuvimos que el uso de un modelo de Text Mining para predecir el potencial de empleo de planes de negocio es posible y escalable a otros concursos con niveles de precisión bastante razonables. Sin embargo, es importante destacar esto significa el primer paso y que aún se puede mejorar bastante más este análisis desarrollando modelos como una mayor cantidad de recursos lingüísticos. Un punto importante a destacar se basa en la evaluación de los modelos, en donde no sólo es debido fijarse en la precisión promedio del modelo, sino que en como esta se comporta en cada clase y en cada predicción con el objetivo de aplicar más detalle al análisis y poder obtener resultados más confiables. Además, analizar la confianza del modelo en cada predicción es de vital importancia debido a que una posible aplicación podría basarse sólo en las predicciones con alta seguridad, lo que mejoraría notablemente los resultados.

La tesis se organiza como sigue: En el capítulo 2 se entrega una revisión literaria de las principales corrientes económicas relacionadas con la creación de empleos y se introduce el concepto de Data Mining, junto con su evolución y potenciales aplicaciones, dentro de las cuales se presenta el concepto de Text Mining. Además se muestra la brecha a la que se intentará aportar y se plantean las preguntas de investigación. En el capítulo 3 se habla acerca del marco metodológico y los algoritmos a utilizar, además de los datos disponibles. En el capítulo 4 se muestran los resultados obtenidos y las respuestas a las preguntas de investigación planteadas. En el capítulo 5 se muestran las conclusiones del trabajo y los potenciales gaps detectados. El capítulo 6 muestra la bibliografía usada y finalmente el capítulo 7 los anexos.

2. Revisión literaria

En este capítulo se revisaran los principales trabajos existentes en la literatura, donde se incluyan principalmente dos áreas de conocimiento. La primera, consiste en la teoría económica y su relación con la creación de empleo y las diferentes formas de medir la misma, además esta sección toca el tema del emprendimiento debido a su directa relación con la creación de puestos de trabajo. Dentro de esta sección se realiza una investigación sobre las diferentes ontologías existentes relacionadas con el emprendimiento, esto nos permite comprender la visión que se tiene en la literatura de los factores que afectan el emprendimiento y como se han abordado hasta ahora. La segunda, se enfoca en *Data Mining* y *Text Mining*, sus aplicaciones generales y específicas relacionadas con la economía en general y con la creación de empleos de manera más específica.

2.1 Teoría económica para la creación de empleos

La literatura existente no es muy extensa cuando se requiere observar las posibles relaciones que existen entre la creación de empresas o el número de emprendedores y la potencial cantidad de empleos que podrían generarse. Pese a esto, existen algunos autores que si han publicado al respecto, considerando la gran importancia que reviste para un país el disminuir su tasa de desempleo, más que nada por dos principales razones. Primero, mientras mayor sea el número de gente empleada, probablemente mayor será el producto interno bruto del país en cuestión, lo que desencadenaría en mejores condiciones para el país en general. Segundo, mientras mayor sea la cantidad de gente que se encuentre obteniendo algún ingreso, mayor será la contribución a disminuir la pobreza y a aumentar el estándar de vida promedio de la personas dentro del país respectivo.

Para poder tocar el tema del emprendimiento de una forma más general, se buscaron diferentes visiones en las que este tema fue investigado y documentado. Dentro de estas últimas destaca la visión de Cantillion (1730) [2]quien define el emprendimiento como una actividad de auto empleo, en donde el individuo en cuestión compra a un precio dado en el presente para vender a

un precio mayor en el futuro. Otra visión que vale la pena considerar es la entregada por Schumpeter (1942) [3] quien dice que el emprendimiento se relaciona al crecimiento por la forma en que convierte ideas en innovaciones exitosas, generando una “creación destructiva” en donde crean nuevos productos eliminando antiguos.

Con respecto a la relación existente entre el número de firmas y la creación de empleo, se observa que Ayyagari, Demirguc-Kunt&Maksimovic (2011) [4] documentan que las grandes empresas (> 99 empleados) son las que más contribuyen al empleo agregado, con un 54,6% del total de empresas según sus tamaños, mientras que las más pequeñas (< 19 empleados) sólo aportan un 16,5% en este ítem. Sin embargo, si analizamos el aporte que hacen las SMEs (pequeñas y medianas empresas con menos de 99 empleados) y lo comparamos con el aporte de las grandes, vemos que esta relación se compensa de manera bastante clara en donde las SMEs aportan un 45,6% y las grandes el ya mencionado 54,6%. Dados estos resultados, los autores plantean que para analizar el empleo agregado es necesario analizar todo el universo de empresas y no sólolas grandes firmas, que a priori se podría haber pensado que aportaban una gran parte del total de empleo agregado. Para complementar este análisis los autores decidieron separar los datos según países, dividiendo a estos según los ingresos que presentaban. Se observa que la diferencia entre las empresas chicas y grandes es menor en los países de bajos ingresos (22% aproximadamente) que en los países de mayores ingresos (30% aproximadamente). La razón podría estar basada en que la relación entre tamaño de la empresa e ingresos es en promedio positiva, por lo que es de esperarse que donde existan mayores ingresos, el aporte de las empresas más grandes será mayor.

Otro foco que plantean los autores es sobre la creación de empleos, la cual se midió como la variación de estos mismos cada 2 años. Para su análisis, separaron la muestra entre los países que tenían variaciones positivas y negativas de empleo. Un punto importante es que solamente el 10% de la muestra correspondió a países con variaciones negativas. Analizando la primera muestra, vemos que las firmas pequeñas aportan un 46,7% a la creación de nuevos empleos, las medianas (entre 20 y 99 empleados) un 30% y las grandes sólo un 16,9%. Este análisis también fue presentado según los ingresos de los países, donde se observó que en las zonas de bajos ingresos las empresas pequeñas son las que más empleos nuevos generan, mientras que en las zonas de altos ingresos son las empresas medianas las que toman este papel. Cabe destacar que

en ambos casos, las firmas grandes tienen valores extremadamente bajos, siendo un 4,5% y un 6,3% respectivamente. Dentro de la segunda muestra se observa que las empresas pequeñas mantienen una variación positiva del 37% aproximadamente, mientras que las medianas también lo hacen con un porcentaje muy cercano al 15%. A diferencia de las dos anteriores, las empresas grandes presentan una disminución en promedio de 157% lo que explica en gran parte por que al analizar en datos agregados da una variación negativa.

Dado esto, es posible decir que incluso en los países que presentan variaciones negativas de empleo, las empresas pequeñas tienen variaciones positivas por lo que podrían ser potencialmente las impulsoras de un cambio en la tendencia existente. Rehaciendo este análisis por zonas, vemos que la diferencia entre empresas pequeñas y grandes vuelve a ser mayor en las zonas de menores ingresos, en donde las empresas pequeñas crean un 66.9% de nuevos empleos y las grandes destruyen un 180%. Por su parte, en las zonas de altos ingresos vemos que las empresas pequeñas aportan un 22.3% y las grandes destruyen un 144%. Es interesante notar que en las zonas de bajos ingresos las compañías medianas se mantienen cerca de un 0%, mientras que en las zonas de altos ingresos aportan casi tanto como las empresas pequeñas.

Estas conclusiones se condicen con los resultados obtenidos por Haltinwanger, Jarmin& Miranda (2009) [5]que documentaron que en Estados Unidos las firmas jóvenes mostraron ser una importante fuente de creación de trabajos netos en comparación con las firmas más grandes.

Luego de esto, la pregunta que se plantean los autores es por qué este aumento en la creación de nuevos empleos no provoca un aumento en el crecimiento delpaís. La respuesta encontrada la basan en la productividad, documentaron que las empresas pequeñas cuentan con una tasa de productividad notoriamente menor que las más grandes, lo que genera que la relación entre el aumento de los nuevos empleos por parte de las empresas pequeñas y la tasa de crecimiento del país no sea lineal. Esto último replantea la discusión sobre si el foco debe ser el de crear empleos, o si además considerar la productividad como un factor preponderante, lo que probablemente será motivo de otras investigaciones en el futuro.

Por otro lado, Klappler&Love (2010) [6] centran la discusión en la creación de nuevas firmas en función de los ingresos del país en cuestión y como la tasa de creación anual se ve afectada por diversos factores. Realizando una comparación entre las diversas zonas divididas según ingresos, los autores documentaron que la tasa de creación de nuevas empresas por cada 1000 habitantes

en edad trabajadora es de 4 en las zonas de mayores ingresos, y de menos de uno en las zonas de menores ingresos. Dentro de las zonas llamadas “en desarrollo” existe una alta heterogeneidad en los valores, en donde Europa y Asia central cuenta con un valor de 2,26 y el África Subsahariana un valor de 0,58. Llevando estos valores a indicadores anuales por año, se llega a la conclusión de que las zonas de mayores ingresos crean 55000 nuevas firmas cada año, mientras que las de bajos ingresos sólo 15000 en promedio. Por su parte, dentro de las economías en desarrollo tenemos que Latinoamérica cuenta con un indicador de 35000 mientras que el África Subsahariana sólo crea 9000 firmas al año. Para concluir este análisis, los autores testearon la relación existente entre el PIB y la creación de nuevas firmas, obteniendo una significativa relación lineal positiva que confirma estudios anteriores que ya habían postulado esta tendencia.

El siguiente paso del análisis fue observar los factores que más determinaban esta tasa de creación dentro de cada país. Los resultados mostraron que tasas más altas de se dan en países que cuentan con un mejor gobierno, un entorno de negocios fuertemente regulado, una baja tasa de impuestos corporativos y un bajo nivel de burocracia. El ambiente de negocios fue medido por medio de dos indicadores, el primero consideraba todos los costos oficiales de crear una firma y los costos legales en que se debe incurrir (notarios, documentos, etc.), este valor se expresaba como un porcentaje del ingreso nacional bruto. El segundo indicador medía el número de procesos necesarios para comenzar un negocio. La relación entre estos dos indicadores y la tasa de creación es negativa y significativa, lo que hace perfecto sentido. Dentro del nivel de gobierno se usaron bastantes indicadores, sin embargo los más significativos fueron la cantidad y fuerza de regulaciones y la percepción de efectividad del gobierno en este punto en particular. Para finalizar su análisis, demostraron que la relación entre la tasa de impuestos corporativos y la creación de nuevas empresas es negativa y significativa.

Por su parte, Ghani&KerrO´Connell (2011) [7] plantean que la relación entre un “ambiente de emprendimiento” y la creación de empleo no es automática pero si existe, ya que reconocen que ciudades que fomentan el emprendimiento experimentan mayores niveles de creación de empleo. Dado esto, los autores intentan descubrir los factores que atraen de manera más efectiva a los emprendedores, encontrando 2 tópicos principales; los niveles de educación a nivel local y la calidad de la infraestructura física existente. El primer punto lo explican desde el punto de vista de que la educación permite mejorar las habilidades y de esta forma ayudar a la transferencia de

ideas de manera más rápida e informada. El segundo se basa en que, pese a que exista un alto nivel de educación, es necesario un nivel de infraestructura (carreteras, bodegas, etc.) para poder mover bienes y servicios de manera oportuna.

Para finalizar, es importante mencionar que no es el objetivo de esta sección profundizar de sobremanera en las causas o factores que subyacen la tasa de creación de empleos, sino entregar una visión general de los principales estudios que se han realizado al respecto.

2.1.1 Ontologías del emprendimiento

Para poder analizar las diferentes ontologías existentes en la literatura es primordial definir este concepto para asegurarnos de que el lector comprenda a cabalidad su significado y posibles implicancias. Si se toma en consideración a Osterwalder (2004) [8], vemos que el autor explica que este concepto tiene su origen en la filosofía antigua y consiste en una disciplina que convive con la organización de la realidad, contraria a la epistemología que se relaciona, más que nada con el conocimiento, finalmente lo define como un “modelo de referencia”. Yee-YeenChu y Wen-ChungHsu (2006) [9] la describen como un amplio concepto que dependerá del objetivo para el que quiera ser usado, el cual consiste en la percepción de la realidad por parte de uno mismo. En otras palabras, se podría definir como la concepción del mundo con que cuenta una persona en particular. Sousa, Manso, Costa y Almeyda (2012) [10] documentan que los sistemas ontológicos se encuentran en auge dentro de la industria de las tecnologías de información, además introducen el término OWL (*Web OntologyLanguaje*) que definen como uno de los enfoques principales para relacionar este concepto con datos. Los autores, basados en el trabajo de Lystras y Garcia (2008) [11] ilustran además el concepto de Servicios web de información semántica, el cual definen como el lugar donde construir o desarrollar sofisticadas bases de datos y procesarlas usando sistemas como OWL. En la siguiente figura se observa la visión de los autores del lugar donde se encuentra la ontología en el enfoque semántico.

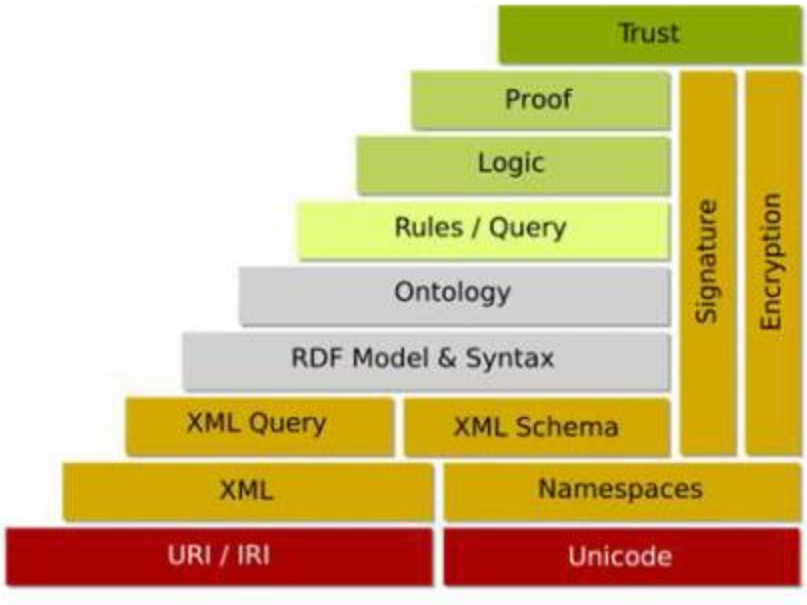


Figura1: Niveles de la Web Semántic, Lystras y Garcia (2008) [11].

Luego de haber definido este importante concepto, y antes de relacionarlo con el emprendimiento, se abarcara un concepto más general, el cual nos permitirá no sólo tener una visión más amplia del enfoque sino que más útil para evaluarlo desde una perspectiva de negocios. Este concepto se refiere al de modelo de negocio. Un trabajo muy aceptado y citado por distintos autores es el de Osterwalder (2004) [8], el cual desarrolla una ontología para los modelos de negocio. A su vez, es necesario definir el concepto de modelo de negocio para asegurarnos de que el lector lo comprende de la misma manera y con el mismo enfoque. Paul timmers (1998) [12] fue uno de los pioneros para definirlo, el entendió un modelo de negocios como la arquitectura completa de un producto o servicio, sus flujos de información y una descripción de todos los actores del negocio con sus respectivos roles. Por otro lado, Weill and Vitale (2001) [13] lo definen como una descripción de los roles y relaciones entre un proveedor y un consumidor, identificando además los flujos de información y de dinero. La definición más clara encontrada en la literatura es la que provee Osterwalder, Pigneur&tucci (2005) [14] donde lo definen como una herramienta conceptual que contiene todos los objetos y relaciones para expresar la lógica de una firma, lo que debe ser entendido como la entrega de una descripción simplificada del valor que se le entrega a los consumidores, como se entrega y sus consecuencias financieras.

Un aspecto importante sobre estos últimos es que el autor enfatiza el carácter de dinamismo que presentan los mismos, en base al constante cambio que sufren debido principalmente a su entorno. Esto continúa la línea de Linder&Cantrell (2000) [15] quienes documentan que los modelos de negocios son una foto en un momento determinado de tiempo, pero que la mayoría de estos se encuentran bajo una constante presión para cambiar debido a diversos factores en el entorno de la firma. Dentro de esta misma línea, Markides&Onyon (2010) [16] señalan que los modelos de negocios siempre estarán en constante cambio, debido a que los mercados siempre lo están. Por su parte, Saab y Fonseca (2008) [17]definen el entendimiento del emprendimiento como un proceso inductivo y extremadamente sensible al entorno.

Un punto importante es que el uso de ontologías no es un proceso antojadizo, sino que necesario a la hora de analizar el concepto de modelo de negocio y de emprendimiento. Morecroft (1994) [18]documenta que el proceso de modelar sistemas sociales o una ontología (como lo es un modelo de negocio) ayuda a identificar y comprender los elementos relevantes en un dominio específico y como estos se relacionan entre sí.

Tal como se explicó anteriormente, se comenzara con las ontologías existentes relacionadas con modelos de negocios para terminar con las relacionadas al concepto específico de emprendimiento. Un buen punto de partida es el desarrollado por Osterwalder (2004) [8]quien creo una ontología basada en 9 bloques sobre modelos de negocio con el objetivo de poder describirlos de manera precisa ya la vez creando un estándar para futuros trabajos. Estos fueron desarrollados sobre una base de 4 bloques, los cuales se presentan a continuación:

1) Producto: Se define no sólo como el producto o servicio, sino también como la propuesta de valor que ofrece la compañía al mercado.
2) Interfaz de cliente: Clientes objetivo de la empresa, como entregan su producto y servicio y como desarrollan relaciones fuertes con el cliente.
3) Manejo de la infraestructura: Como la firma maneja su logística, de qué manera y cuan eficiente es.
4) Aspectos financieros: Cual es el modelo de ingresos, la estructura de costos y como el modelo se hace sustentable.

Estas 4 áreas son comparables con las 4 perspectivas desarrolladas por Norton y Kaplan (1992) [20] para su cuadro de mando integral y se podría decir que su influencia en el trabajo del autor

fue significativa. Siguiendo una línea similar, Markides (1999) [21] desarrollo una ontología desde una perspectiva de la estrategia del negocio, en la que definió 3 bloques. Estos eran el “que” ofrece la compañía, a “quien” apuntaba y “como” lo lograría. Es posible notar que estos 3 enfoques se encuentran contemplados dentro del enfoque de Ostenwalder (2004) [8], donde el “que” se relaciona directamente con el producto, el “quien” con la interfaz de cliente y el “como” con el manejo de la infraestructura.

Además, Ostelwalder (2004) [8]agregó más detalle al análisis y desarrollo 9 sub áreas que buscan complementar las 4 ya implantadas. A continuación se muestra lo descrito anteriormente:

Pillar	Building Block of Business Model	Description
Product	**Value Proposition**	A Value Proposition is an overall view of a company's bundle of products and services that are of value to the customer.
Customer Interface	**Target Customer**	The Target Customer is a segment of customers a company wants to offer value to.
	Distribution Channel	A Distribution Channel is a means of getting in touch with the customer.
	Relationship	The Relationship describes the kind of link a company establishes between itself and the customer.
Infrastructure Management	**Value Configuration**	The Value Configuration describes the arrangement of activities and resources that are necessary to create value for the customer.
	Capability	A capability is the ability to execute a repeatable pattern of actions that is necessary in order to create value for the customer.
	Partnership	A Partnership is a voluntarily initiated cooperative agreement between two or more companies in order to create value for the customer.
Financial Aspects	**Cost Structure**	The Cost Structure is the representation in money of all the means employed in the business model.
	Revenue Model	The Revenue Model describes the way a company makes money through a variety of revenue flows.

Figura2: Ontología estándar para modelos de negocios, Ostelwalder (2004) [8].

Es posible observar que 4 los factores antes desarrollados, son desagregados con el objetivo de aumentar el detalle y la especificidad del análisis. Esto permite que cada firma pueda adaptarse mejor al modelo y a la ontología creada. Si se compara esta ontología con la propuesta por Markides 1999 [21] se observa que el nivel de especificidad es mucho mayor, y que no se ha perdido estandarización en las definiciones. Esto último es importante debido a que una

descripción muy específica podría quedar obsoleta rápidamente dado los constantes cambios que sufren los modelos de negocio.

Además, una segunda tabla fue creada para caracterizar a cada elemento de la ontología, la que cuenta con varios puntos que se muestran a continuación.

Name of BM-Element	NAME
Definition	Gives a precise description of the business model element.
Part of	Defines to which pillar of the ontology the element belongs to or of which element it is a sub-element
Related to	Describes to which other elements of the ontology an element is related to.
Set of	Indicates into which sub-elements an element can be decomposed.
Cardinality	Defines the number of allowed occurrences of an element or sub-element inside the ontology.
Attributes	Lists the attributes of the element or sub-element. The allowed values of an attribute are indicated between accolades {VALUE1, VALUE2}. Their occurrences are indicated in brackets (e.g. 1-n). Each element and sub-element has two standard attributes which are NAME and DESCRIPTION that contain a chain of characters {abc}.
References	Indicates the main references related to the business model element.

Figura3: Caracterización de cada elemento de la Ontología estándar para modelos de negocio, Ostelwalder (2004) [8].

Como se observa, cada elemento o bloque de la ontología cuenta con una serie de características adicionales. Esto permite analizar un modelo de negocios con diferentes niveles de granularidad, con mayor o menor detalle dependiendo de las necesidades específicas del usuario.

Sousa, Manso, Costa &Almeyda (2010) [10] usan esta metodología, pero documentan que dado el cada vez mayor nivel de interconexión presente en el mundo actual y el desarrollo de nuevas tecnologías que hasta hace poco eran impensadas, era necesario ampliar esta ontología con el objetivo de mantenerla valida y actualizada. La siguiente figura muestra la nueva ontología realizada por los autores.

Pillar	Building block	Description
Product	Value Proposition	A Value Proposition is an overall view of a company's bundle of products and services that are of value to the customer.
Customer Interface	Target Customer	The Target Customer is a segment of customers a company wants to offer value to.
	Distribution Channel	A Distribution Channel is a means of getting in touch with the customer.
	Relationship	The Relationship describes the kind of link a company establishes between itself and the customer.
Infrastructure Management	Value Configuration	The Value Configuration describes the arrangement of activities and resources that are necessary to create value for the customer.
	Capability	A Capability is the ability to execute a repeatable pattern of actions that is necessary in order to create value for the customer.
	Partnership	A Partnership is a voluntarily initiated cooperative agreement between two or more companies in order to create value for the customer.
Financial Aspects	Cost Structures	The Cost Structure is the representation in money of all the means employed in the business model.
	Revenue Model	The Revenue Model describes the way a company makes money through a variety of revenue flows.
Supplier Interface	Supplier	Supplier connections are a key aspect.
	Supply Chain	Company's needs good supply chains in order to reduce costs and improve quality in products and services.
	Relationship	A good supplier is a good partner for business and some type of relationships are a key advantage for success.
External Aspects	Political	All business is influenced by political changes in the country and globally.
	Economics	Macroeconomics systems have indirect influence, but relevant in business success.
	Law	Law changes can have a strong impact in tech based companies and other kind of businesses.
	Money	All business need money and some companies have a big relation with money exchange market.
	Society	Society and social change have a strong connection with business.

Figura4: Ontología estándar para modelos de negocio, Sousa, Manso, Costa &Almeyda (2010) [10].

Tal como se puede observar, 2 bloques principales y 8 sub bloques fueron agregados. La interfaz del proveedor nos muestra de cierta manera como el entorno de los nuevos modelos de negocio se ha hecho cada vez más complejo, en donde la relación con los proveedores juega un rol vital dado el aumento en la competencia. Esta relación no sólo se refiere a la humana, sino además a como las cadenas de valor de ambos se organizan de manera de ser lo más eficientes posible. En el bloque de aspectos externos se observa que aspectos como las leyes o la sociedad, que hasta hace un tiempo no eran tan mencionados al momento de hablar de modelos de negocio han aumentado su importancia debido a factores como el aumento de empresas, la mayor demanda o la creación de nuevas licencias.

Esto último no fue el único aporte de los autores, debido a que además agregaron 2 variables a la caracterización específica de cada elemento o bloque de la nueva ontología. Los cambios se definen a continuación:

- Se creó la característica Status, que apunta a saber si el elemento es provechoso o no para el modelo de negocios desde un punto de vista general.
- Se creó además una escala que busca reflejar en un valor entre 1 y 5 el ítem anterior para poder efectuar análisis basados en valores.}

La nueva tabla se puede observar en la siguiente figura:

Name of BM-Element	Description
Definition	Gives a precise description of the business model element.
Part of	Defines to which pillar of the ontology the element belongs to or of which element it is a sub-element.
Related to	Describes to which other elements of the ontology an element id related to.
Set of	Indicates into which sub-elements an element can be decomposed.
Cardinality	Defines the number of allowed occurrences of an element or sub-element inside the ontology.
Attributes	List the attributes of the element or sub-element. The allowed values of an attribute are indicated between accolades {Value1, Value2}. Their occurrences are indicated in brackets (e.g. 1-n). Each element and sub-element has two standard attributes which are Name and Description that contains a chain of characters {abc}.
References	Indicates the main references related to the business model element.
Status	Successful or Unsuccessful
Scale	1 to 5

Figura5: Caracterización para la Ontología estándar, Sousa, Manso, Costa &Almeyda (2010) [10].

Centrándose más en el concepto de emprendimiento, se observa que Chu&Hsu (2006) [9] documentan que en Taiwan, la creación de valor proveniente desde el emprendimiento se genera al aprovechar oportunidades mediante la generación y aplicación de nuevo conocimiento, soportando las condiciones de mercado. Este modelo se muestra en la siguiente figura.

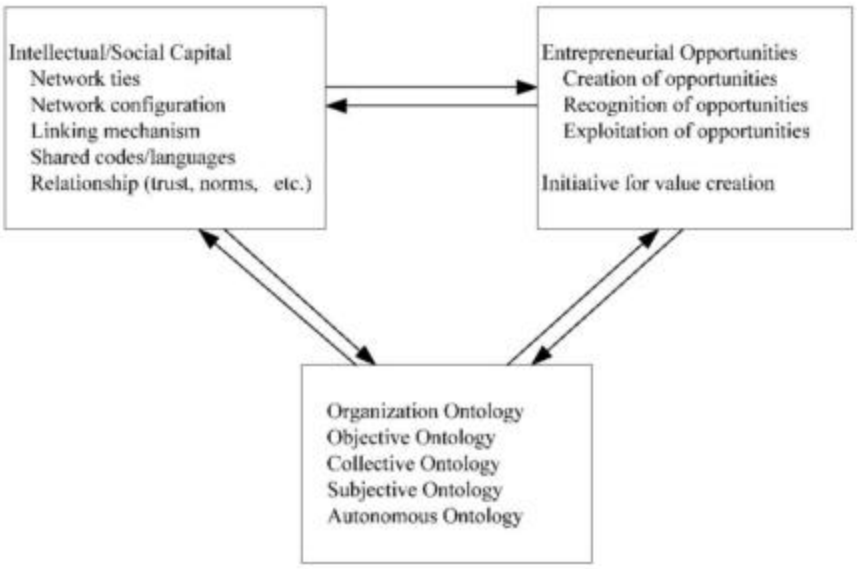


Figura6: Ontología estándar del emprendimiento, Chu&Hsu (2006) [9].

Es la interacción de estas 3 áreas la que permite aprovechar de mejor manera las oportunidades y con esto resultar exitoso al emprender. A continuación se explica cada una de las áreas mostradas anteriormente:

- *OrganizationOntology*: Este enfoque se basa principalmente en el conocimiento que tiene la firma tanto de sí misma como de sus competidores. Los autores explican que teniendo un conocimiento superior es posible explotar y desarrollar sus recursos de mejor manera que la industria, obteniendo de esta forma una ventaja competitiva.
- *Intellectual / Social Capital*: Cada vez más, el capital intelectual es considerado un activo intangible en la creación de oportunidades, sobre todo desde el punto de vista de lasstart-ups, debido a que quien posee intelecto humano, es capaz de estructurar rutinas más eficientes y permite una mejor relación dentro de la compañía y hacia fuera, lo que podría traducirse en una ventaja competitiva.
- *EntrepreneurialOpportunities*: El objetivo se basa en identificar oportunidades rentables que otros no hayan descubierto o explotado aún. Stevenson, Roberts &Grousbeck (1999) [35] indican que identificar y seleccionar las oportunidades correctas es la capacidad más importante de las start-ups.

Las restricciones, tanto geográficas como de industria, de la investigación anterior son bastante fuertes por lo que si se desea extrapolar resultados o concluir de manera más segura, se vuelve necesario analizar otras corrientes más generalizables. Jones, Coviello&Tang (2011) [22] analizan de manera agregada el fenómeno del emprendimiento a nivel global con una revisión literaria de casi 20 años y generan una ontología que resume todas las características del emprendimiento, la que se muestra a continuación.

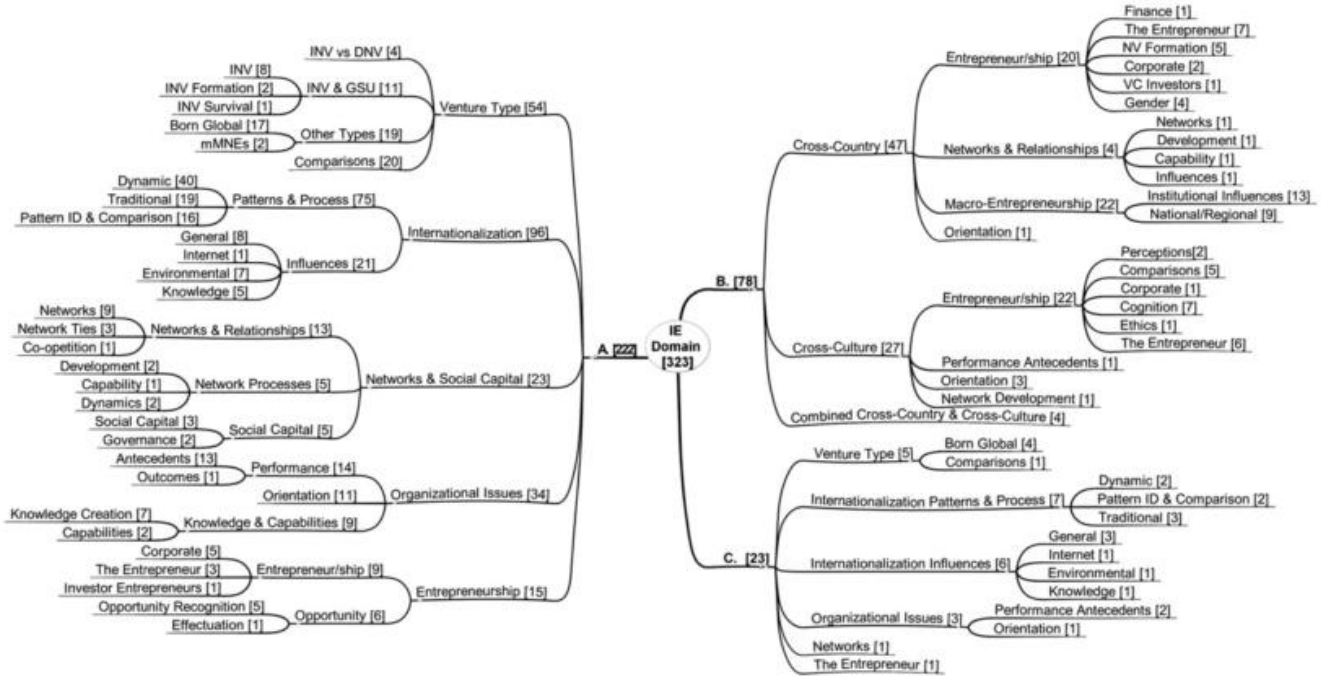


Figura7: Ontología estándar del emprendimiento, Coviello&Tang (2011).

Como se puede ver, la estructura cuenta con 13 ramificaciones que permiten explicar las distintas variantes, las cuales se definirán brevemente a continuación:

1. *Venture type*: Se centra en las características o antecedentes de cada tipo de organización que puede llegar a ser global. Ejemplos de esto podrían ser INV (*International New Venture)* o BG (*Born Global),* este último concepto se refiere a firmas que se internacionalizan de manera rápida y temprana.
2. *Internationalization:* Se enfoca más que nada en los patrones y procesos necesarios para el proceso de internacionalización, además de las influencias que se generan sobre él.

3. *Networks and Social Capital*: Se basa principalmente en las redes que se forman o que potencialmente se podrían formar, en los procesos para comprender estas redes y en el capital social, el cual se define como el valor creado fruto de la interacción de dos o más partes interesadas.
4. *OrganizationalIssues*: Consiste en el estudio delrendimiento, la orientación, conocimiento y capacidades de una firma para expandirse al extranjero.
5. *Entrepreneurship*: Se enfoca principalmente en el reconocimiento de potenciales oportunidades y en el emprendimiento corporativo, el cual se define como la capacidad de una firma para internacionalizarse de manera responsable y organizada.
6. *Cross Country Research*: Consiste principalmente en las diferencias que pueden existir en los comportamientos y percepciones de los emprendedores. También se incluyen datos de las sociedades en las que se desenvuelve cada emprendimiento (Un ejemplo de esto, es que la clase más educada de Japón rara vez deja un cargo en una compañía para emprender por su cuenta).
7. *Cross Culture Research*: Se basa en el análisis de como el comportamiento de los emprendedores se ve afectado por la cultura del país de origen y el destino.
8. *Combined Cross Country and Cross Culture*: Son más que nada comparaciones entre algunas zonas, que permiten caracterizar de mejor forma a algunos tipos de emprendimientos.

El último grupo de elementos trata sobre comparaciones entre cada tópico y permite obtener conclusiones más aisladas sobre la caracterización ontológica de un emprendimiento determinado. De esta forma se realizó un recorrido por las distintas ontologías existentes que se asocian al emprendimiento, comenzando con las más generales que hablaban más de modelos de negocios o start-ups y terminando con el emprendimiento como tal.

2.2 Data Mining

Hoy en día, el ritmo exponencial con el que se crean datos sobrepasa la capacidad que tenemos de manejarlos o trabajar con ellos, incluso de almacenarlos. Cada dos días se crea la misma cantidad de datos que se creó desde el inicio de los tiempos hasta el año 2003. Dado este escenario, es que surge la urgente necesidad de generar herramientas o metodologías que permitan extraer información de estas gigantescas bases y transformarlas en conocimiento. Como bien se podría pensar, esto no sólo atañe al plano de la administración, sino que la gama de campos de investigación que se enfrentan a este tema es cada vez mayor. Esta tarea se vuelve especialmente compleja en el campo de los datos no estructurados, como podría serlo una imagen o un documento de texto, debido a que los datos no siguen una estructura establecida lo que dificulta aún más la extracción de información útil de ellos.

2.2.1 Definición

Antes de definir Data Mining como tal, se debe realizar una definición intermedia para poder comprender a cabalidad su funcionamiento e implicancias. Esta definición intermedia consiste en el concepto de *Knowledge Discovery in Databases* ("KDD").

Margaret Dunham (2012) [23] entrega una definición que será usada como punto de partida para la definición de este concepto:

a) **KDD:** Es el proceso de encontrar patrones en los datos, y transformarlos en conocimiento/información útil para los objetivos de investigación respectivos.

Para complementar esta definición, usaremos la que entrega Fayyad, Piatetsky-Shapiro y Smyth (1996) [24], en donde definen el término "KDD" como el proceso de identificar nueva y no trivial información, potencialmente útil y que además contenga patrones comprensibles dentro de una base de datos. El término "*no trivial"* hace referencia a que la información no puede ser directamente o sencillamente extraíble de los datos mediante la mera visualización de los mismos. A su vez, "*nueva"* hace referencia a que la información debe contener un valor

agregado o a lo menos, generar discusión. La información o los datos se refieren, más que nada, a un grupo de casos y el "patrón" a encontrar se refiere a un subgrupo de los mismos con características similares. Existen otras definiciones de patrones, las cuales comprenden el hecho de encontrar una estructura en los datos, ajustar un modelo a los mismos o hacer una descripción con un alto nivel de especificidad. Un aspecto importante es que se plantea el "KDD" como proceso, lo que implica que comprende varios pasos, que incluyen la preparación de los datos, la búsqueda de información o patrones dentro de ellos, la evaluación del conocimiento obtenido y el perfeccionamiento de la metodología, todo esto en constantes iteraciones. A modo de conclusión, es posible decir que esto no se trata de una aplicación directa de alguna fórmula o algoritmo sobre los datos, sino que implica una serie de pasos que generan que la información obtenida sea nueva (al menos para los propósitos de la investigación), validable con algún grado de certeza y potencialmente útil en el problema de investigación. Además es importante que los patrones sean entendibles para el investigador (en el caso de que no lo sean, se puede aplicar algún tipo de proceso para estos efectos).

Un punto importante a considerar, el cual es discutido por los autores en el capítulo, trata sobre la evaluación de los modelos y la posibilidad de asignar medidas cuantitativas a sus resultados. Existen actualmente medidas de certeza (como podría serlo el ajuste del modelo a los datos, una correlación o la precisión al momento de clasificar nuevos datos), y existen también medidas de utilidad (como por ejemplo ganancias en términos de unidades monetarias cuando se intenta predecir el precio futuro de una acción). Sin embargo, existen otros criterios como la novedad o la "parsimonia" que son bastante más subjetivos, ya que dependen entre otras cosas del conocimiento previo del investigador. De esta discusión surge el concepto de "nivel de interés" (*interestingness*), el cual se propone como un valor del patrón o una medida global, que combina validez, novedad, utilidad y simplicidad. Las funciones de este indicador pueden ser definidas de manera explícita o bien implícitamente a través del ordenamiento sobre el proceso KDD en cada modelo en específico. Además, los autores plantean que es posible catalogar un patrón como información útil o conocimiento, en la medida en que supere un "umbral de *interestingness*".

Tanto Dunham (2012) [23] como Fayyad, Piatetsky-Shapiro&Smyht (1996) [24] concuerdan en que Data Mining es un paso dentro del proceso KDD que consiste en la aplicación de análisis de datos y "algoritmos de descubrimiento" que, bajo aceptables limitaciones de eficiencia

tecnológica y computacional, producen modelos sobre los datos. Estos modelos consisten en patrones que podrían ser infinitos, por lo que una investigación previa del espacio muestral suele ser requerida. En conclusión, tenemos que el proceso KDD implica la utilización de bases de datos, y el Data Mining es una etapa donde aplican algoritmos con el fin de enumerar o encontrar potenciales patrones provenientes de los datos que entreguen respuesta a un problema planteado.

Ambos investigadores vuelven a coincidir al momento de definir las etapas del proceso KDD y el proceso mismo como uno iterativo e interactivo que envuelve numerosos pasos y en donde el usuario cuenta con un rol activo mediante la toma de decisiones. Resaltan 9 etapas, las cuales se describen a continuación:

1) La primera etapa consiste en comprender el dominio de la aplicación y que conocimientos previos son relevantes. Además se debe identificar el objetivo del proceso KDD desde una perspectiva “de cliente”.

2) La segunda etapa consiste en la creación o selección de un grupo de datos en específico, enfocándose en un subconjunto de estos desde donde se intentara obtener información.

3) La tercera etapa consiste en la limpieza de los datos mediante algunos pre procesos, dentro de los cuales se incluyen; remover el ruido o recolectar información para modelarlo e incluso usarlo como input, tomar decisiones en base a datos faltantes, secuencias temporales o cambios conocidos.

4) La cuarta etapa implica la reducción de datos y la proyección de los mismos, esto se logra encontrando características útiles para representar los datos de acuerdo al objetivo de la investigación. Además se pueden usar disminuciones de la dimensionalidad de los datos, además de posibles transformaciones para reducir el número efectivo de variables en consideración.

5) La quinta etapa trata sobre alinear los objetivos del proceso KDD (primera etapa) con un método particular de Data Mining. Ejemplos de estos métodos pueden ser clasificación, reducción, regresión, clustering, etc.

6) La sexta etapa corresponde a la elección del algoritmo que se usará para buscar patrones en los datos. Esto implica decidir que modelos y parámetros son apropiados (por ejemplo modelos que funcionan sobre datos categóricos, no lo hacen sobre vectores reales o a veces funcionan pero de manera diferente), y alinearlos con el criterio usado en el proceso KDD (por ejemplo, un usuario podría estar más interesado en entender las razones de una predicción más que el resultado de esta misma).

7) La séptima etapa consiste en la tarea de Data Mining como tal. En otras palabras, implica la búsqueda de patrones de interés en un forma de representación en particular, en donde se incluyen reglas de asociación, arboles de decisión, clustering, entre otros. El investigador puede ayudar notoriamente al proceso ejecutando de manera correcta los pasos anteriores.

8) La octava etapa consiste en la interpretación de los patrones obtenidos (es posible que se deba retornar a algunas de las etapas anteriores para posibles iteraciones). Además, incluye la visualizaciónde los patrones obtenidos o de los datos luego de la aplicación del modelo.

9) La novena y última etapa consiste en consolidar el conocimiento obtenido, incorporándolo en un nuevo sistema para una investigación posterior o documentándolo y reportándolo a las partes interesadas. Esta parte del proceso implica además el chequeo y la resolución de potenciales problemas con creencias anteriores.

Es normal que el proceso KDD implique variadas iteraciones o loops entre 2 etapas cualesquiera. Debido a que un buen resultado en una etapa previa puede suponer una mejora en etapas posteriores lo que alienta a rehacer o mejorar iterativamente cada parte del proceso. Es variados artículos se encuentra la siguiente diagramación del proceso:

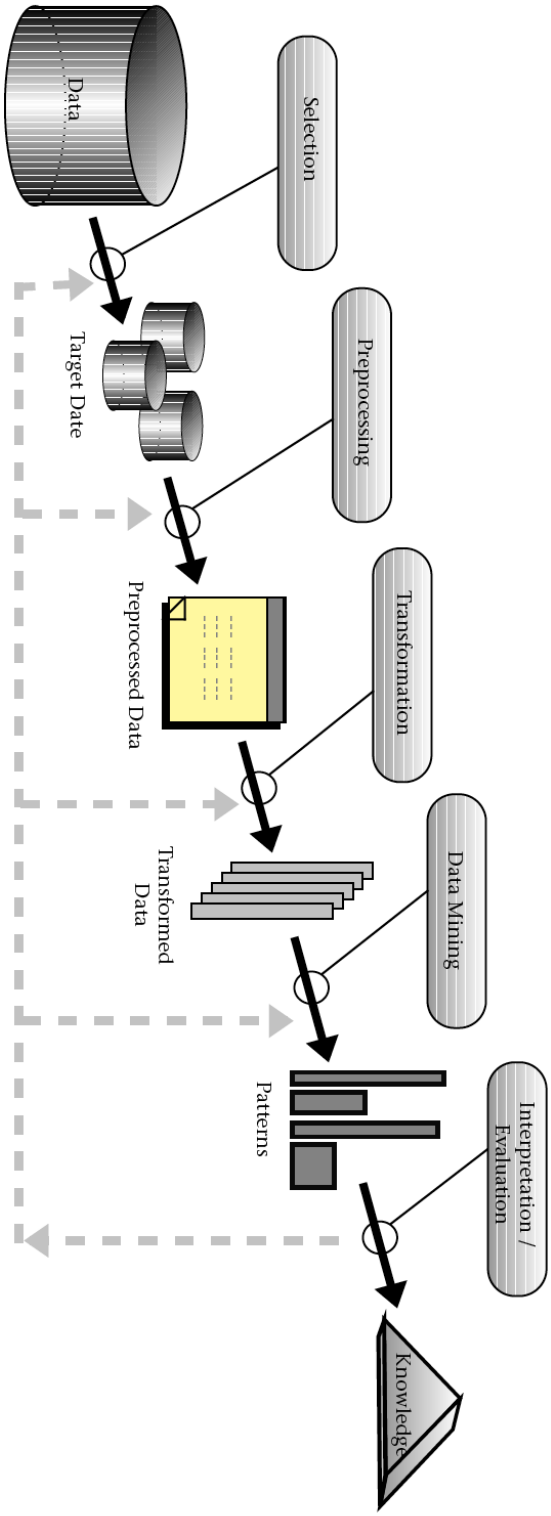


Figura8: Vision General de los Pasos que Componen el Proceso de KDD, Fayyad, Piatestsky-Shapiro&Smyth (1996).

Un gran número de investigadores se centran en la séptima etapa del proceso descrito anteriormente (pese a que, muchas veces una rigurosa preparación de los datos puede reportar beneficios mayores que la misma elección del algoritmo), que comprende ajustar modelos para buscar patrones desde los datos objetivos. Los modelos ajustados se consideran de “conocimiento inferido”, pese a que usualmente el juicio humano es requerido en todo el proceso de KDD. En este proceso de ajuste de modelos se usan principalmente 2 formalismos matemáticos:

1) Estadísticas: Este enfoque permite efectos no determinísticos en el modelo, es usualmente utilizado dada la incertidumbre presente en los procesos de generación de datos.

2) Lógicas: Este enfoque permite efectos puramente determinísticos.

La mayoría de los algoritmos de Data Mining se encuentran basados en métodos de *machine learningypatternrecognition*, dentro de este último destacanalgoritmos de clasificación, clustering, regresiones y más. Pese a que existe un amplio rango de algoritmos descritos en la literatura existente, es importante destacar que estos muchos métodos sólo subyacen algunas pocas técnicas fundamentales. La representación real que usa un modelo, proviene normalmente de la composición de unos pocos y bien conocidos modelos tales como polinomios, kernels, splines, funciones de base, etc. Estos algoritmos tienden a diferir en el criterio utilizado para medir la bondad del ajuste.

Basado en la literatura, es posible señalar que las tareas de predicción y descripción son, en la práctica, los dos objetivos principales dentro de los métodos de Data Mining. Predicción implica el uso de algunas variables de la base para predecir valores desconocidos o futuros de otras variables de interés, mientras que descripción se centra en buscar patrones que sean fácilmente interpretables para describir los datos de manera no trivial. La importancia de cada tarea va a depender del objetivo del problema a tratar. Las tareas básicas de Data Mining son las siguientes:

1) **Clasificación:** Trata acerca de la categorización de datos en clases predefinidas o grupos. También se le suele llamar aprendizaje supervisado en los casos en que estas clases son creadas después de examinar los datos. Un ejemplo de esto es la clasificación de si un plan de negocio tendrá una alta, media o baja potencial creación de empleo (siendo estos 3 valores las clases). Según lo expresado por Dunham (2012) [23], los algoritmos de

clasificación requieren que las clases sean creadas en base a los valores de ciertos atributos, los que fueron seleccionados luego de observar las características de los datos y habiendo determinado su pertenencia a una determinada clase.

2) **Regresión:** Se usa para clasificar datos en variables de predicción de valor real. La regresión implica normalmente el aprendizaje de la función que ejecuta esta clasificación. Toma como supuesto que los datos objetivos son ajustables por medio de la utilización de algún tipo de función conocida y con eso, se busca la función que pueda modelar los datos de la mejor manera posible. Otro análisis que se usa para determinar que función tiene el mejor ajuste, es el análisis de errores o de residuos.

3) **Análisis de series de tiempo:** Comprende el análisis de la variación en el valor de un atributo en el tiempo. Un pre procesamiento que suele desarrollarse para este tipo de tareas es dejar los datos espaciados por el mismo espacio de tiempo (días, meses, años, etc.). Este análisis cuenta con 3 funciones básicas; una es usar medidas de distancia para representar la similitud entre diferentes series; otra es examinar la estructura de la tendencia con el objetivo de modelar su comportamiento; y la tercera se trata de usar series de tiempo históricas para predecir valores futuros.

4) **Predicción:** Esta tarea se encuentra en parte de la literatura dentro de clasificación, sin embargo se decidió dejar fuera debido a sus marcadas diferencias, la cual consiste en anticipar un futuro estado de la naturaleza, en vez del actual. Existen casos en que esto se puede observar como una aplicación más que un método de Data Mining como tal.

5) **Clustering:** Tiene cierto parecido con la clasificación, sin embargo la diferencia radica en que los grupos no son predefinidos por el investigador sino que en los mismos datos se buscan patrones que permitan agruparlos. Alternativamente se podría considerar un proceso no supervisado. El objetivo es conseguir segmentar los datos en grupos que tengan una alta homogeneidad interna y una alta heterogeneidad externa, o sea con los distintos grupos. En otras palabras, lo que se hace es agrupar los datos que son más similares entre ellos en "*clusters"*. Dado el hecho de que estos datos no han sido

predefinidos, a menudo se requiere el conocimiento de un experto para interpretar el significado de los *clusters*creados.

6) **Summarization:** Implica mapear los datos en submuestras, dejando sólo la información más importante. A esta tarea también se le llama caracterización. En otras palabras, deriva información representativa de la base de datos.

7) **Reglas de asociación:** También se le suele llamar análisis de enlace, afinidad o asociación. Busca establecer relaciones entre los datos. Una aplicación clásica es la que se aplica a las *ventas retail*en la que se buscar determinar que productos son comprados juntos. Un aspecto a considerar para el investigador es que se debe descartar la posibilidad de que las relaciones establecidas por el algoritmo sean relaciones causales.

8) **Descubrimiento secuencial:** Busca encontrar patrones secuenciales en los datos, es bastante similar a la asociación entre eventos relacionados, aunque en este caso la relación se basa en el factor tiempo. A diferencia del ejemplo anterior, que requiere que los ítems sean comprados al mismo tiempo, aquí los productos son comprados en momentos distintos siguiendo un determinado orden.

Con todo lo expuesto anteriormente, se puede concluir que tanto Data Mining como Knowledge Discovery in Databases corresponden a la evolución e integración de variadas disciplinas, tales como administración de las bases de datos, recuperación de información, visualización de datos, reconocimiento de patrones, estadística, algoritmos, computación e inteligencia artificial, entre muchos otros, agrupados para obtener el mayor y más útil conocimiento posible de las bases de datos.

2.2.2 Aplicaciones Generales y Evolución del Data Mining

Tal como se comentó en el ítem anterior, Data Mining es el fruto de la evolución de distintas disciplinas. Sobre esto, Venkatadri y Lokanatha (2011) [25] señalan que en los inicios del Data Mining es posible apreciar 2 corrientes principales, la "Tendencia de los datos" y la "Tendencia computacional". En la primera, se consideraba que los algoritmos de DM trabajaban de mejor manera con datos numéricos recolectados de una base de datos única, con lo que sobre esto, variadas técnicas han evolucionado para archivos planos, bases de datos tradicionales y relacionales. Luego, con la influencia de la estadística y técnicas de *machine learning*, varios algoritmos evolucionaron para poder soportar datos no numéricos y bases relacionales. En la segunda, se destaca la influencia del DM en el desarrollo de lenguajes de programación de cuarta generación y varias técnicas de computación relacionadas. En los inicios del DM, la mayoría de los algoritmos se basaba sólo en estadística, luego evolucionaron con variadas técnicas computacionales tales como la "inteligencia artificial" o el "reconocimiento de patrones".

Debido en parte al gran éxito que ha obtenido esta corriente en áreas tales como el cuidado de la salud, retail, telecomunicaciones, detección de fraudes, análisis de riesgo, entre otros, y en parte al aumento en la complejidad de cada uno de los campos cubiertos, es que se han abierto nuevos desafíos sobre Data Mining. Estos desafíos implican diferentes formatos de datos, avances en computación, mayor complejidad de los negocios, etc. Venkatadri y Lokanatha (2011)[25] presentan como desafíos actuales del Data Mining los siguientes:

1) **Minería de Datos Heterogeneos:**

   1.1) **Hipertexto e Hipermedia DM:** Es una base de datos desde catálogos en línea, librerías digitales y bases de datos de información en línea que incluyen *hyperlinks*, marcas de grupo de texto, y otros tipos de datos. El Web Mining es la aplicación que trata sobre buscar patrones de datos obtenidos desde la web. Las técnicas de DM que se utilizan para los datos de Hypertexto e Hipermedia son clasificación y clustering.

1.2) DM "ubicuo": La palabra "ubicuo" significa tener la capacidad para estar en diversos lugares al mismo tiempo, en el contexto de DM, se relaciona con la generación de datos que provienen de datos móviles. El desarrollo de laptops, celulares y artefactos computacionales con gran capacidad y acceso a la red están llevando al "paradigma de la computación ubicua". En esta corriente se utilizan variadas técnicas tradicionales de DM, tales como una combinación entre aprendizaje de máquinas y estadística.

1.3) DM Multimedia: Los datos multimedia soportan una gran variedad de formatos, tales como imágenes, video, audio y animación. Las técnicas de DM que se aplican a datos multimedia, usualmente son árboles de decisión, *Support Vector Machines*, métodos de reglas de asociación y métodos de *clustering*.

1.4) DM Espacial: Los Datos espaciales incluyen datos astronómicos y satelitales. Las técnicas de DM que suelen utilizarse incluyen sistemas OLAP (*on-line analyticalprocessing)* y métodos de *clustering*espaciales.

1.5) DM de series de tiempo: Se trata de técnicas aplicadas a secuencias de puntos o datos medidos de manera sucesiva y en espacios de tiempo uniformes. Ejemplos de esto podrían ser, series de precios de acciones, tipos de cambio, volúmenes de ventas, datos de clima, etc. Para este tipo de problemas se suelen usar algoritmos de reglas de inducción o árboles de decisión.

2) Utilización de los recursos computacionales

Una de las principales características del Data Mining, es la utilización de computación avanzada y recursos de red, donde destacan computación paralela, de distribución y de redes. Algoritmos que suelen usarse en este campo son los denominados "a priori". La computación paralela y de distribución se integra con avanzadas tecnologías de red generando métodos como SVM, el que a su vez se usa en DM de distribución o distribuida. Los autores muestran que recientemente se han aplicado metodologías

pertenecientes al denominado "*softcomputing",* tales como lógica difusa, Redes Neuronales y *SupportVector Machines* para el análisis de variados formatos de datos almacenados en bases de datos distribuidas, lo que genera sistemas más robustos y que entregan soluciones interpretables y de bajo costo comparado con otras técnicas tradicionales para el análisis sistémico, sistema de pre procesamiento, procesamiento de información flexible, análisis de datos y de toma de decisiones.

3) Tendencias de investigación y computación científica

Las grandes cantidades de datos que se crean en variadas disciplinas científicas tales como astronomía, bioinformática o imagenologíaha llevado a que las técnicas de Data Mining descritas anteriormente deban ser refinadas. A modo de ejemplo, tenemos que las técnicas basadas en *kernels* directos son herramientas que se han vuelto muy potentes a la hora de tratar con problemas de predicción o de selección de características.

4) Tendencias de negocios

La mayoría de las técnicas de DM se usan en este ámbito, sin embargo las técnicas de clasificación y de predicción son las más usadas con el fin de dar soporte a otras decisiones de negocios. Sin embargo otras técnicas también se usan a menudo, tales como regresiones, *clustering,* etc. En este tipo de ambientes, el DM ha evolucionado hasta convertirse en lo que se denomina Sistemas de Soporte Decisional ("DSS") y ha migrado hacia la denominada "Inteligencia de Negocios".

Data mining type	Application Areas	Data Formats	Data mining Techniques/Algorithms
Hypermedia data mining	Internet and Intranet Applications.	Hyper Text Data	Classification and Clustering Techniques
Ubiquitous data mining	Applications of Mobile phones, PDA, Digital Cam etc.	Ubiquitous Data	Traditional data mining techniques drawn from the Statistics and Machine Learning
Multimedia data mining	Audio/Video Applications	Multimedia Data	Rule based decision tree classification algorithms
Spatial Data mining	Network, Remote Sensing and GIS applications.	Spatial Data	Spatial Clustering Techniques, Spatial OLAP
Time series Data mining	Business and Financial applications.	Time series Data	Rule Induction algorithms.

Figura9: Descripción de las Áreas de DM actuales y Técnicas para minar distintos tipos de datos, Venkatadri & Lokanatha (2011).

Para finalizar el análisis, Venkatadri&Lokanatha (2011) [25] identifican las futuras tendencias que enfrentara el campo del DM, destacando las siguientes:

1) **Estandarización de lenguajes de DM:** Existen variadas herramientas de Data Mining con diferentes sintaxis, lo que hace complejo usar distintas herramientas en el mismo problema. Es por esto que se plantea la conveniencia de estandarizar los lenguajes de manera que no sólo revista menor dificultad aplicarlos, sino que también aprenderlos.

2) **Pre procesamiento de datos:** Como ya se dijo anteriormente, las técnicas de Data Mining han ido evolucionando con el objetivo de poder encontrar patrones útiles en bases de datos que cada vez son más complejas y de mayor tamaño. Los algoritmos enfocados al pre procesamiento de los datos no parecen estar a la altura de las necesidades del mercado en este ítem, por lo que se plantea la necesidad de desarrollar nuevas aplicaciones de DM que permitan mejorar la eficiencia en este punto.

3) **Objetos de datos complejos:** Una de las apuestas de los autores que el DM va a usarse en todos los campos de la vida humana, es por esto que preocupa el hecho de

que sus algoritmos se restringen a sólo formas tradicionales de datos. Surge la necesidad para tratar con datos de alta velocidad, secuencialidad, ruido en las series de tiempo y objetos multi-representados.

4) **Recursos computacionales:** Mejoras en las áreas de alta velocidad, paralelismo, redes y *cloudcomputing* han creado nuevos desafíos para el DM. La principal es la mejora en la alta velocidad, lo que ha impuesto una gran demanda por nuevas y eficientes técnicas de DM para analizar datos masivos. El DM basado en redes debe poner atención en el tema de la privacidad y seguridad de los datos, mientras que el *cloudcomputing*está penetrando cada vez en más áreas, lo que hace que demande más y mejores técnicas.

5) ***Web mining:*** El desarrollo de la web y la gran cantidad de usos que posee hace que genere cada vez más contenidos, lo que representa más datos, por lo que el *Web Mining* probablemente seguirá creciendo. Existen diversas áreas que requieren mayor desarrollo, como lo son procedimientos de predicción, modelos de extracción de uso de datos y en general hace falta un mayor entendimiento de como las diferentes partes del proceso podrían impactar los resultados de los respectivos problemas de investigación.

6) **Computación científica:** En los últimos años, el DM ha atraído a la investigación en variadas aplicaciones científicas y computacionales, como lo son el descubrimiento de correlaciones significativas o patrones con una eficiencia asombrosa. A su vez, se destaca la necesidad de un mayor desarrollo de investigación en el campo de información ambiental, recursos naturales y problemas de biología molecular.

7) **Tendencias de negocios:** La principal necesidad que se vaticina en el ambiente de negocios, es la relacionada al E-Business. Las técnicas de DM son prometedoras en el desarrollo de nuevas herramientas que pueden ser usadas para proveer mayor privacidad al cliente, satisfacción y mejores productos.

Pese a que el número de posibles aplicacionesde Data Mining es gigantesco, a continuación se describen 3 aplicaciones clásicas con el fin de ilustrar la flexibilidad de los enfoques desarrollados:

1) **Servicios de salud:** Existen herramientas de DM que han sido utilizadas en variadas ocasiones por los servicios de salud, debido principalmente a que puede beneficiar significativamente a todas las partes interesadas. Un ejemplo de esto, es que herramientas de DM pueden ayudar a las compañías aseguradoras a detectar fraudes o a realizar segmentaciones de mercado.

2) **Aplicaciones del *Retail*:** Una de las aplicaciones más exitosas es el análisis más conocido como análisis de canastas, por ejemplo si un supermercado amacena las compras de sus clientes, el DM podría utilizarse para identificar aquellos clientes que consumen chocolate sobre la mantequilla, y así con una amplia gama de productos.

3) **Banca:** En la literatura se muestran diversas aplicaciones en actividades tales comoel delineamiento de procedimientos de líneas de negocios, CRM, detectar fraudes, clasificaciones de niveles de crédito, etc.

Hoy en día, la utilización de las distintas herramientas de DM ha llegado a tal nivel de especialización que algunas de sus ramas han tomado importancia por sí solas. Este es el caso de la denominada “inteligencia de negocios” o “Business intelligence” (BI), término que es definido por Rud (2009) [26] como el conjunto de teorías, metodologías, procesos, arquitecturas y tecnologías que transforman los datos en información útil y relevante para propósitos del negocio. BI es capaz de manejar grandes volúmenes de información, ayudando a identificar y desarrollar oportunidades que pueden convertirse en una clara ventaja competitiva.

Se ha mostrado en esta sección lo que significa Data Mining, su evolución en el tiempo, los futuros desafíos y la flexibilidad de las metodologías empleadas que permiten enfrentar estos mismos. Esto genera que los 3 ejemplos anteriores, puedan ser mucho más diversos en áreas como juegos, ingeniería, investigaciones espaciales, música, etc.

2.2.3 Text Mining

La proliferación de medios digitales en la actualidad ha resultado en una necesidad para desarrollar nuevos mecanismos para que los humanos busquen, procesen y analicen cada vez mayores cantidades de datos desde múltiples fuentes de información. Un ejemplo directo de esto son los planes de negocio en formato digital que son generados diariamente por emprendedores para presentarlos a distintos *stakeholders,* incluyendo concursos tanto privados como públicos. Este problema se ve exacerbado dado que el texto es un tipo de dato no estructurado y por lo tanto, no puede ser procesado y analizado por las analíticas tradicionales de manera directa.

2.2.3.1 Definición

Recopilando algunas definiciones existentes en la literatura, se obtiene que Tan (1999) [27]define Text Mining como el proceso de extraer interesante y no trivial conocimiento desde documentos de texto. Como se puede ver, esta definición se asemeja bastante a la de Data Mining, pero se enfoca en datos no estructurados como lo es el texto. Por otro lado, Feldman&Sanger (2004) [28] lo definen comouna nueva y excitante área de la ciencias de la computación que intenta solucionar la crisis existente de información combinando técnicas de data mining, machine learning, procesamiento de lenguaje natural y manejo del conocimiento. Siendo más específicos lo expresan como un enfoque que provee las herramientas necesarias para encontrar conocimiento en bases de datos de texto. Hearst (2003) [29] lo define como el proceso automático para descubrir información desconocida y generar conocimiento.

Basados en Diaz(2013) [1], es posible decir que la Minería de Texto o *Text Mining,* consiste en una tarea de pre-procesamiento de documentos o grupos de estos (corpus) por medio de diversas técnicas, dentro de las que se cuentan extracción de información, extracción de términos especiales, categorización automática de texto y almacenamiento de representaciones intermedias de los datos en formatos estructurados, como podrían serlo los análisis de conglomerados (*clusters*), de distribución, de tendencias, modelos de clasificación, reglas de asociación y además la visualización de estos resultados.

Dentro de las técnicas existentes para el pre-procesamiento descrito, existen 2 categorías principales: *lingüísticas y no lingüísticas.*Las técnicas lingüísticas consideran las características del lenguaje natural del texto en los documentos, como por ejemplo la sintaxis, agrupación de conceptos, reglas gramaticales, sinónimos, etc. En otras palabras intentan comprender el significado humano que posee el texto. Por otro lado, las técnicas no lingüísticas consideran a los documentos como una colección de caracteres, palabras, frases, párrafos, etc. sin comprender realmente lo que significan para los seres humanos. Básicamente, basan todo su análisis en la frecuencia de símbolos en cada documento o corpus. De esta manera, calculan la proximidad existente entre palabras o grupos de palabras, y son capaces de hacer este análisis relativo a otros documentos.

Debido a que las técnicas lingüísticas buscan darle sentido humano al texto, normalmente suelen hacer uso de recursos externos al software de minería de texto para ayudar en el análisis. Estos recursos suelen hacer referencia a como reglas del lenguaje natural se aplican al corpus de documentos. Por ejemplo, los significados y categorías gramaticales dependen directamente del idioma de los documentos que serán analizados, y por lo tanto, estos deben seguir las reglas impuestas por el idioma respectivo. De la misma manera, el usuario puede querer hacer uso de reglas del lenguaje natural que son aún más específicas que el idioma, como podría serlo un contexto o dominio en el cual los documentos se encuentran enmarcados. Ejemplo de estos dominios pueden ser el financiero, musical, histórico, comercial, etc. Algunos dominios incluso cuentan con subdominios, como la contabilidad, el rock, el derecho penal, finanzas corporativas, etc. En estos casos el usuario puede hacer uso de recursos específicos asociados a dichos dominios como taxonomías y ontologías.

Dado esto último, es que la minería de texto también puede definirse como dependiente del dominio o independiente del dominio. Es importante destacar que aunque se use una minería de texto independiente del dominio, igualmente se pueden usar recursos del lenguaje natural dado que muchos de éstos son independiente del cuerpo de conocimiento específico en estudio. Normalmente, los softwares de minería de texto son más efectivos si es que algún nivel de dependencia del dominio es incorporado en el análisis, pese a que en la actualidad los softwares se encuentran bastante lejos de cubrir una amplia gama de dominios en nuestro idioma.

Las herramientas de minería de texto, suelen traer un número bastante limitado de recursos asociados a los distintos idiomas que son capaces de soportar. Dentro de estos recursos se cuentan diccionarios, sinónimos, extractores automáticos de entidades, eventos, etc. Dado que el esfuerzo de desarrollar estos componentes de software es una tarea compleja y larga, es común que los paquetes de componentes se encuentren sujetos a protección de propiedad intelectual de sus desarrolladores y que no se encuentren disponibles fácil o gratuitamente.

2.2.4 Aplicaciones de Text Mining en la Economía

Una interesante aplicación es la realizada por Hendry&Madeley (2010) [30] en la cual usan técnicas no lingüísticas detextmining para extraer mensajes de las comunicaciones del banco de Canadá e investigar si esos mensajes tenían un impacto significativo en las tasas de interés. Los datos utilizados provenían principalmente de dos fuentes, reportes de política monetaria y PressRelease de tasas de interés.

Varios modelos fueron creados, cada uno con alrededor de 10 categorías o grupos de conceptos que fueron capaces de predecir alrededor del 40% de la varianza en la tasa de interés. Cada categoría se relaciona con un aspecto del problema, a modo de ejemplo tenemos que la primera categoría se enfocaba en predecir el crecimiento del PIB y contenía palabras como por ejemplo *labour*, *real* y *financecredit*. La segunda se enfocaba en las decisiones sobre el riesgo e incluían términos como *risk, upside, downside*y *uncertain* y la tercera se relacionaba con descensos en la tasa de interés y contiene palabras tales como *credit, slowdown*y *condition*.

Sus conclusiones arrojaron que mediante los análisis de textmining fueron capaces de encontrar términos que permitían explicar la varianza delos retornos de mercado y de las volatilidades de la tasa de interés. Además, gracias a las categorías que crearon fueron capaces de profundizar en la razones de tales cambios. Como discusión propuesta se propone aumentar el número de documentos (para el estudio se hizo sólo con 95) y explorar otros métodos de *textmining* como técnicas lingüísticas por ejemplo.

Otra aplicación es la conocida como análisis de sentimiento, la cual busca interpretar el enfoque de un texto, por ejemplo si este representa un hecho positivo o negativo, o incluso si representa

rabia, agradecimiento, felicidad, etc. Dentro de esto se enmarca el artículo de Li, Xie, Chen, Wang &Deng (2014) [31] en donde buscan demostrar que existe una relación entre las noticias del proveedor FINET y la tasa de interés en la bolsa de Hong Kong.

Los autores dividieron las acciones en 4 grupos; Comercio, Finanzas, Propiedades y Utilidades. Además trabajaron con las variaciones en el precio de cada acción dentro de este grupo y utilizaron *Support Vector Machines* para las tareas de clasificación de las mismas. Con respecto a la manera de aplicar el análisis de sentimiento, se aplicaron 4 enfoques. Los dos primeros usaban un enfoque de polaridad de las palabras (buenas o malas), otro enfoque usaba un *textmining no linguistico* (frecuencia de palabras) y el cuarto usaba “diccionarios de sentimiento”, el cual usaba bases de palabras creadas manualmente con objetivos similares a los de la investigación. Con esto realizado dividieron las muestras y crearon el modelo con una muestra de entrenamiento para testearlo en otra muestra, llamada de prueba.

Los resultados mostraron que el análisis de sentimiento como tal obtuvo el mejor rendimiento, sobretodo en el área de finanzas. Es importante destacar que esta área fue la que represento los mejores resultados de precisión en casi todos los enfoques, lo que se explica posiblemente porque existen palabras que sólo atañen a esta área o que sólo explican variaciones de las tasas de interés representativas de esta. Estos resultados fueron probados en la muestra de entrenamiento y de prueba, obteniéndose, en términos relativos, resultados bastante congruentes.

Una conclusión importante que obtuvieron los autores, fue que el enfoque de polaridad no fue suficiente para efectuar predicciones correctas, debido a que sólo dos posibles valores de sentimiento no alcanzaban a interpretar la cantidad de posibles estados existentes en el contexto. Como trabajo futuro, los autores proponen seguir el análisis aumentando el número de diccionarios en el modelo creado para el análisis de sentimiento.

2.3 Posicionamiento y Contribución

Como se ha podido observar a lo largo de este capítulo, se ha revisado la literatura existente relacionada con 2 tópicos; La creación de empleos, en donde se incluye los esfuerzos realizados por medirla, por predecirla y por identificar los factores que permitirían impulsarla y las técnicas de Data Mining, focalizándose también en el Text Mining.

Tal como se expresó en la introducción, el objetivo de esta tesis es predecir la potencial creación de empleos que un plan de negocio pueda entregar, por medio de herramientas de Data y Text Mining. Es por eso que se buscaron aplicaciones económicas que tocarán el tema de la creación de empleos, donde destaca Ayyagari, Demirguc-Kunt&Maksimovic (2011) [4] quienes documentan la relación existente entre el número de firmas y los empleos creados. Como este análisis existen otros similares, pero que tocan el tema en retrospectiva. En otras palabras, basan su investigación en los empleos que ya fueron creados, no en los que podrán serlo. Por otro lado, se encuentra el trabajo de Osterwalder (2004) [8] quien relaciona el tema con el emprendimiento, creando una ontología para este último, sin embargo sólo toca el tema de la creación de empleos de forma tangente, debido a que agrega al análisis el hecho de si el plan de negocio fue exitoso o no, lo que podría relacionarse con una probable creación de empleo. Sin embargo, vemos que este análisis vuelve a ser en retrospectiva.

Por otro lado, vemos que técnicas de Data y Text Mining se han usado para aplicaciones económicas, en donde resaltan principalmente análisis de sentimiento para predecir fluctuaciones de la tasa de interés, como el realizado por Li, Xie, Chen, Wang &Deng (2014) [31], obteniendo resultados prometedores. Esto, representa el comienzo de una visión hacia el futuro pero que no se centra aún en la creación de empleos. Es por esto que se observa una brecha en este aspecto, por lo que la contribución de esta tesis se basa en analizar el tema de la creación de empleos hacia adelante, prediciendo si un plan de negocio dará o no empleo en el futuro, intentando acortar esta brecha. Las aplicaciones de esto son variadas, por ejemplo la posibilidad de incentivar la creación de empleo por medio de concursos enfocados solamente en este aspecto. Además, dado al carácter lingüístico del modelo de Text Mining, se creará una ontología relacionada directamente con la creación de empleos, la cual no existe actualmente en

la literatura. El principal aporte de este, es que se generarán nuevos recursos para crear una ontología relacionada con el tema que permita crear modelos cada vez mejores y más confiables.

A modo de conclusión, se obtiene que la relación existente entre el área de creación de empleos y Data Mining es extremadamente pequeña y difusa, donde lo más relacionado a esto han sido aplicaciones económicas que poco se relacionan con el objetivo del problema.

2.3 Preguntas de investigación

A continuación se muestran las preguntas de investigación de esta tesis:

1) Es posible generar un modelo que en base al análisis de un plan de negocio prediga el potencial de empleo con una precisión superior a un modelo aleatorio.

H1: No es posible generar un modelo que realice la tarea.

H2: Es posible generar un modelo que realice la tarea con un 100% de precisión.

H3: Es posible generar un modelo que realice la tarea con una precisión superior a la obtenida por un modelo aleatorio pero menor al 100%.

H4: Es posible generar un modelo que realice la tarea con un mejor rendimiento que un modelo no lingüístico.

Esta pregunta es importante debido a que la existencia de un modelo que pueda realizar esta tarea de mejor manera que un modelo aleatorio es la base de esta tesis, por lo que es la primera pregunta que debe ser realizada y sus conclusiones serán la base de los resultados obtenidos. Además, plantea dos comparaciones obligatorias a la hora de evaluar un análisis de este tipo, las cuales son un modelo aleatorio y otro no lingüístico.

2) Es posible encontrar ciertos factores claves que explican el nivel de empleo generado

H1: el factor más importante es la industria del plan de negocios.

H2: el factor más importante es aportar a la reducción del desempleo.

H3: el factor más importante es entregar ayudar a la comunidad.

Esta pregunta es importante debido a que ahonda en las razones por la cuales un plan de negocio busca crear empleo. La creencia popular apunta a que sólo se buscan beneficios económicos y se ve el empleo como un medio para ello, por su lado otros apuntan a que se buscar ayudar a la comunidad. Saber las razones por la cuales se está entregando empleo es igual o más importante que saber si está entregando en primer lugar, debido a que permite comprender el perfil de los distintos emprendedores y de esta manera desarrollar soluciones mucho más específicas y personalizadas.

3) Es posible predecir el potencial de empleo con una seguridad del 100%

H1: Es posible predecir el potencial de empleo de la totalidad de planes de negocio con una seguridad del 100%.

H2: Es posible predecir el potencial de empleo de al menos un 20% de planes de negocio, con un seguridad del 100%.

Esta pregunta reviste una gran importancia debido a que cuando un modelo este tipo se aplica en la vida real se evalúa en función de sus resultados, y cada predicción suele llevar asociado un desembolso económico o una decisión. Es por esto que un 100% de seguridad, aunque no sea en el total de la muestra le entrega un gran valor al modelo que no había sido considerado en las dos preguntas anteriores. Finalmente en la hipótesis 2 se usa un 20% dado que se toma este valor como un mínimo umbral de predicciones que cuenten con una máxima seguridad, tomando el supuesto de que un valor menor de predicciones no permitirían obtener resultados contundentes en algún tipo de aplicación.

4) Es posible aplicar el modelo a un concurso distinto del que se usó para entrenarlo.

H1: Es posible aplicar el modelo a otro concurso y obtener una precisión mayor a la obtenida con un modelo aleatorio.

H2: Es posible aplicar el modelo a otro concurso y obtener al menos un 20% de predicciones con un 100% se seguridad.

Esta pregunta busca medir el grado de extrapolación que tiene el modelo, lo que se puede traducir como el grado de aplicabilidad que tiene debido a que el objetivo es que este modelo se aplique a diferentes concursos y permita obtener resultados. En otras palabras, esta pregunta se relaciona con el grado de potencial uso que tendrá este modelo.

3. Metodología

Esta sección comenzara con una descripción de los datos a utilizar, para luego pasar a la metodología como tal. Seusará como guía el libro “CRISP DM 1.0 *Stepbystep data miningguide”*[32], el cual fue creado por la empresa IBM, la cual es pionera no sólo en la aplicación de modelos de data mining, sino que también en todo lo relativo a tecnologías de la información.CRISP DM son las iniciales en inglés de proceso estándar inter industrias para Data Mining (Cross Industry Standard Processfor Data Mining) y se define como una guía para modelar proyectos por medio de esta tecnología, independientemente de la industria o de la tecnología que se ocupe. El objetivo de este proceso es el de estandarizar la manera en que se realizan proyectos de Data Mining y generar con esto, que los mismo se vuelvan más confiables y precisos.

3.1 Datos utilizados y preparación

Para esta investigación, se utilizaron 615 planes de negocio en formato digital provenientes del concurso InfoDev y 6820 provenientes del concurso YouWin. Los primeros se encontraban en un formato de archivo de texto (.pdf o .txt) mientras que los segundos se encontraban juntos en un archivo de Microsoft Excel. Ambos grupos de documentos fueron facilitados por el Banco Mundial, bajo el marco del Proyecto “Future Jobs” dirigido e implementado por los Profesores de la Universidad de Chile Dr. Rodrigo Wagner y Dr. David Díaz.

Con los planes del concurso YouWin no se efectuó preparación alguna debido a que se usaron como muestra de prueba para comprobar la capacidad de extrapolación del modelo, o sea como este era capaz de ajustarse a datos nuevos. Por otro lado con los planes del concurso InfoDev si se realizó una preparación, la cual se detalla a continuación:

1) Lectura: Lo primero fue leer cada plan de manera minuciosa para ser capaz de “predecir” su potencial creación de empleo.
2) Clasificación: Luego de leerlo, se procedió a clasificarlo en alto, medio o bajo en función de los descrito anteriormente.
3) Extracción de ejemplos de texto: El último paso consistió en extraer desde el texto ejemplos que explicasen la clasificación realizada en el paso anterior.

Esto permitió obtener una base de datos en donde se contaban 4 variables:

1) ID: Un indicador único de cada plan, creado con fines de identificación del mismo.
2) Nombre del archivo: Fue creado con el objetivo de realizar algunas tareas automáticas en la base que permitieran su llenado de forma más rápida. Además, se consideró útil teniendo en cuenta que el nombre de cada plan da una idea general del mismo.
3) N-Potentialworkers: Esta fila contiene los valores fruto de la clasificación realizada por el analista, a cerca de la potencial creación de empleo que un plan tiene (alto, medio o bajo).
4) Selectioncriteria “N-Potentialworkers”: Esta fila contiene los ejemplos de texto que explican la clasificación expresada en la variable anterior.

Con esta base se consideró que los datos eran apropiados para continuar con las siguientes fases descritas anteriormente.

3.2 Metodología CRISP DM

La metodología CRISP DM se define como un proceso jerárquico, que cuenta con cuatro niveles desde lo más general a lo más específico. Los 4 niveles son: Fases principales, tareas genéricas, tareas específicas y procesos.

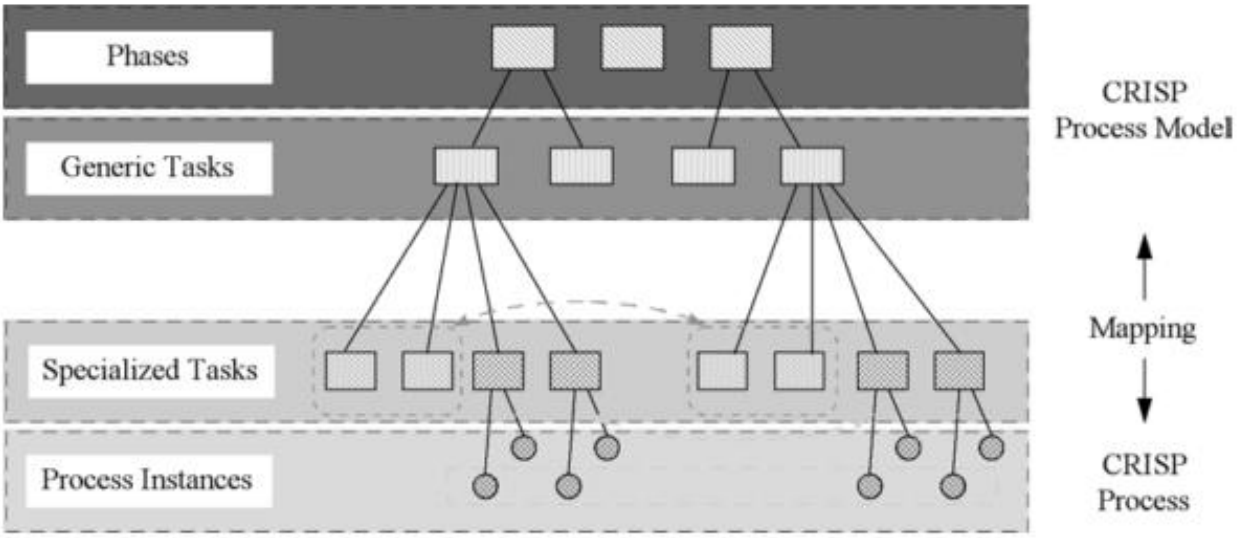


Figura100: Niveles definidos de la metodología CRISP-DM, “CRISP DM 1.0 Stepbystep data miningguide” *[32].*

En el nivel superior, el proyecto de *Data Mining* se fragmenta en ciertas fases. Cada una de estas a su vez, se vuelve a fragmentar en una o más tareas genéricas de segundo nivel. Este nivel cuenta con el carácter de genérico debido a que busca cubrir todas las aplicaciones de DM. Estas tareas genéricas se encuentran diseñadas para que sean lo más estables y completas que se pueda. Al usar la palabra completo, se entiende que lo que se busca es que el modelo pueda usarse en proyectos futuros.

El tercer nivel de tareas especializadas, fragmenta cada actividad genérica en actividades más específicas. Es importante destacar que el orden y la forma de realizar estas tareas específicas pueden variar dependiendo de la situación y/o contexto en la que se encuentre el proyecto. Un ejemplo para esto podría ser que en el segundo nivel exista una tarea llamada “construcción del modelo”. Esta tarea, se separaría en diversas tareas en el nivel 3 en donde una sería

“construcción del modelo de respuesta”, el cual contiene actividades específicas con respecto al problema y a la herramienta de DM escogida.

Cada tarea y fase desempeñada de forma discreta en un orden específico representa una secuencia de eventos que no siempre se cumplirá. En la práctica, el orden de estas tareas podría variar y en muchas ocasiones, la consecución de una tarea no implica que no se pueda rehacer la tarea anterior, lo que implicaría repetir la cadena. El proceso CRISP DM no busca capturar todas las posibles rutas a través del proceso de DM, debido a que esto supondría un mayor nivel de complejidad que limitaría los beneficios potenciales. El cuarto nivel, se define como una grabación de acciones, decisiones y resultados de una acción de DM concluyente. Cada proceso se organiza y se define en base a las tareas definidas en los niveles superiores (tanto genéricas como específicas), pero representa una acción en particular.

Esta metodología presenta 2 guías, las cuales son el “Modelo de Referencia” y la “Guía de usuario”. El primeroentrega una visión más bien general de las fases, tareas y resultados y principalmente describe las vías de acción a seguir en un proyecto de DM. Por otro lado, la segunda es bastante más detallada y entrega tips específicos para cada fase del proceso, también busca describir que hacer en un proyecto de DM pero de manera más minuciosa. Esta última no será reseñada.

3.2.1 El modelo de referencia genérico CRISP DM

Este modelo entrega una visión del ciclo de vida para un proyecto de DM. Contiene las etapas del proyecto, las tareas dentro de cada una y sus resultados. En la siguiente figura se muestra el ciclo de vida de un proyecto según este modelo.

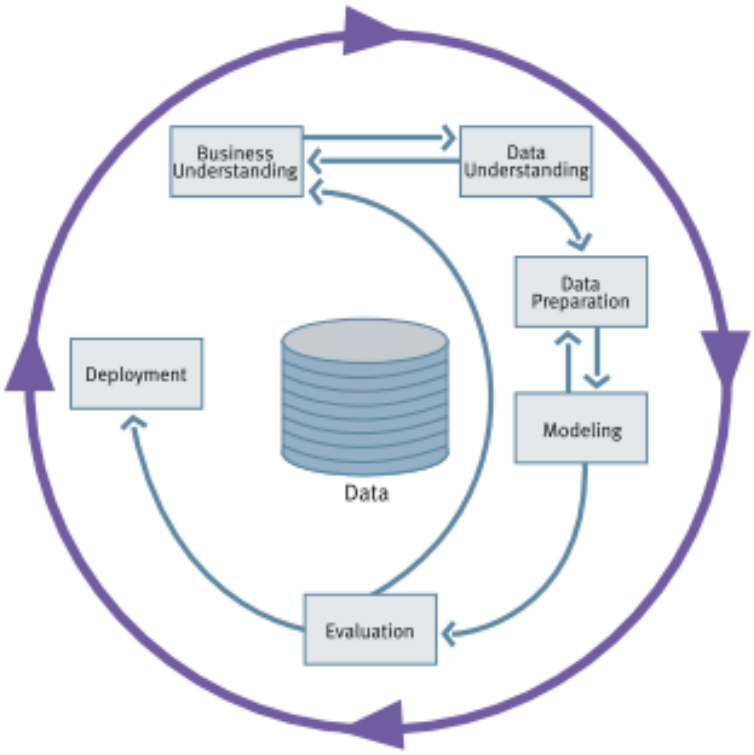


Figura11: Fases del modelo de referencia CRISP-DM, libro “CRISP DM 1.0 Stepbystep data miningguide” *[32].*

La secuencia que se propone entre las distintas fases no es de carácter estricto. El resultado de cada una determina la siguiente tarea a realizar. Las flechas indican, más que nada, las dependencias más importantes y las fases que comúnmente poseen interdependencia.

El circuito exterior representa el ciclo natural de DM. Su forma circular y sin final se debe a que un proyecto de DM no se ha terminado una vez que se llega a una conclusión. Lo aprendido en el proceso y las soluciones encontradas pueden generar nuevas preguntas del negocio, normalmente más específicas. A continuación se reseña cada una de las fases:

1) Entendimiento del negocio (*Business understanding*)

   El principal objetivo de esta etapa, es lograr un conocimiento del negocio y comprender los requerimientos del proyecto desde esa perspectiva, luego estos requerimientos se convertirán en la definición del problema de DM y en un plan, desarrollado a priori, para lograr los objetivos. Esta etapa incluye las siguientes fases:

a) Definir los objetivos del negocio: Consiste en la descripción de los objetivos primarios, además de otras posibles interrogantes que pudiesen existir, o en las que se desea profundizar. Se debe describir además la utilidad del resultado del proyecto, desde una perspectiva de negocios.

b) Evaluación de la situación: Esta tarea implica aplicar más detalle al análisis, pudiendo encontrar potenciales restricciones (aspectos legales, por ejemplo), recursos (tecnológicos, humanos, herramientas de DM, etc.), contingencias, supuestos, entre otros hechos que se consideren determinantes para el proyecto.

c) Determinar metas de DM: Implica definir los objetivos del proyecto pero desde una perspectiva técnica, define los criterios para el éxito (lograr un porcentaje mínimo de precisión en un predicción, por ejemplo).

d) Realizar un plan del proyecto: Implica una descripción de los pasos a seguir para la consecución del proyecto junto con la duración de cada uno, los recursos que serán requeridos, inputs, outputs y las posibles iteraciones que se realizaran. Se deben incluir las posibles revisiones que se realizan al final de cada fase y la evaluación inicial de las técnicas a usar.

2) Entendimiento de los datos (*Data understanding*)

Esta etapa comienza con la recolección de los datos iniciales y con las actividades que buscan una familiarización con los datos, implica además identificar sus posibles problemas de calidad, descubrir los primeros indicios y detectar potenciales subconjuntos que podrían resultar interesantes para desarrollar hipótesis con respecto a información que no se puede observar de manera directa. Las tareas a desarrollar en esta fase son las siguientes:

a) Recolección inicial de los datos: Implica la adquisición de los datos listados en los recursos del proyecto, incluye también el proceso de carga de estos.

b) Descripción de los datos: Examinar las propiedades de la información que se obtuvo (metadata), incluyendo su formato, cantidad, etc. Además se debe evaluar si esos datos satisfacen los requerimientos iniciales.

c) Exploración de los datos: Implica la realización de consultas a los datos, visualización y técnicas de reporte. Incluye análisis simples, los que pueden ser advirtiendo relaciones o estadísticos. Se deben describir o graficar los resultados de esta tarea, incluyendo los primeros descubrimientos o hipótesis iniciales.

d) Verificar la calidad de la información: Consiste en revisar si los datos son capaces de cumplir con los requerimientos del proyecto, si son correctos o si faltan valores y por qué ocurre esto. Luego se debe realizar una lista de estos y proponer posibles soluciones.

3) Preparación de los datos (*Data preparation*)

Esta fase contempla todas las actividades necesarias para desarrollar el conjunto final de los datos. Las tareas que esta fase incluye no siguen un orden particular, e incluso pueden realizarse más de una vez, dentro de las más comunes tenemos la selección de registros, limpieza de datos y transformaciones. Más formalmente las tareas que esta fase incluye son:

a) Seleccionar los datos: Esta tarea implica decidir qué datos se usaran en el análisis, que atributos y que columnas. El criterio para esta selección implica analizar la relevancia respecto a los objetivos de DM, la calidad y las potenciales restricciones técnicas que pudiesen ocurrir. Es importante que las razones para la inclusión o exclusión de información se expongan en alguna lista, debido a que estas pueden cambiar conforme avanza el proyecto.

b) Limpieza: Implica seleccionar subconjuntos de datos limpios y decidir qué hacer con los que no estén (faltantes). Se deben enumerar y describir las acciones que se desarrollaron para solucionar los problemas de calidad de información, además se debe tener en cuenta si se hizo alguna transformación a los datos que implique algún cambio para efectos de limpieza.

c) Construcción: Implica la creación de nuevos atributos o registros, y transformaciones de valores para atributos existentes.

d) Integración de nuevos datos: Existen métodos, mediante los cuales es posible juntar información desde diferentes tablas, registros o incluso bases de datos. En otras palabras, esta tarea consiste en la fusión de tablas que incluyen diferente información, pero que se usarán para el mismo objetivo.

e) Formato: Consiste en distintos tipos de modificaciones que no implican un cambio en el significado de los datos, pero que pueden ser necesarios para alguna herramienta de modelación o incluso para efectos de una mejor visualización.

4) Modelamiento (*Modelling*)

En esta etapa, se seleccionan y aplican varias técnicas que permiten realizar el modelamiento, además sus parámetros son calibrados con el objetivo de obtener resultados óptimos. Normalmente existe más de una técnica para el mismo problema de DM, algunas cuentan con requerimientos específicos sobre la forma de los datos, es por esto que a veces se debe volver a la etapa de preparación de los mismos. Las tareas que incluye esta fase son las siguientes:

a) Seleccionar la técnica de modelamiento: Cabe destacar que es posible que se haya determinado la herramienta a utilizar en la fase de “Entendimiento del Negocio”. Esta fase supone meramente la elección de la técnica a utilizar (por ejemplo un árbol de

decisión C5.0), además se debe especificar los supuestos que tiene esta técnica sobre los datos.

b) Generar una prueba de diseño: Antes de construir un modelo, es necesario generar un procedimiento para probar la calidad del modelo y validarlo. Es por esto que una técnica comúnmente usada separa los datos en conjuntos de entrenamiento y prueba. Esto debe encontrarse documentado, describiendo el plan de entrenamiento y prueba para evaluar el modelo.

c) Construcción del modelo: Se deben documentar los parámetros ajustados y los valores escogidos, junto con la explicación de la elección. Se deben describir los modelos resultantes, su interpretación y las potenciales dificultades encontradas con sus significados.

d) Evaluación del modelo: Consiste en la interpretación de acuerdo al criterio del moderador sobre los criterios de éxito del proyecto. Se juzga el éxito de la aplicación de la metodología y se intenta rankear los modelos. Es importante considerar los objetivos de negocio creados debido a que de ellos depende el resultado del modelo. Se debe revisar la selección de los parámetros definidos y ajustarlos para una próxima iteración, a su vez estas iteraciones deben realizarse hasta que se considere que se ha encontrado el mejor modelo. Todas las iteraciones deben documentarse.

5) Evaluación (*Evaluation*)

Es importante destacar que antes de continuar con la evaluación del modelo, se deben revisar exhaustivamente los pasos ejecutados para crearlo, estar seguro que cumple con los supuestos y los objetivos del negocio. Un objetivo muy importante, se basa en determinar si existe algún aspecto importante del negocio que no haya sido suficientemente considerado. Al final de esta fase, se debe decidir sobre los resultados del uso del DM. Las tareas a realizar son las siguientes:

a) Evaluar los resultados: Se debe evaluar el grado en que el modelo cumple los objetivos del negocio y busca determinar si existe alguna razón por la cual el modelo podría ser considerado deficiente. Además se evalúan otros resultados generados, debido a que estos pueden revelar información importante que no ha sido considerada o futuras necesidades. Si un modelo cumple con los criterios tanto técnicos como del negocio, se encuentra aprobado.

b) Proceso de revisión: Se realiza una revisión más exhaustiva de la tarea de *data mining*realizada con el objetivo de determinar si existe algún paso o factor importante que no se haya considerado o haya sido pasado por alto. Esta revisión también se puede ampliar al modelo, donde destacan atributos como la calidad del mismo, los parámetros seleccionados, etc.

c) Determinar los próximos pasos: Consiste en tomar la decisión de si cerrar el proyecto y proceder a la etapa de implementación o si definir un nuevo proyecto de DM. Se deben describir las decisiones tomadas, su racionalidad y como se procederá.

6) Implementación (*Deployment*)

Normalmente, la fase de creación del modelo no es la fase final. Incluso si el objetivo del modelo fuera aumentar el conocimiento de los datos, este se debe organizar y presentar en una forma que el cliente pueda aplicarlo. Suele involucrar la aplicación de modelos “en vivo” en el proceso de la toma de decisiones de la organización. Un ejemplo de esto podría ser el de la personalización en tiempo real de páginas web. Las tareas en esta fase son las siguientes:

a) Plan de implementación: Se toman los resultados y se decide qué estrategia usar para su implementación.

b) Plan de monitoreo y mantención: Una estrategia de mantención sirve para prevenir el uso incorrecto de DM en periodos muy largos de tiempo. Para monitorear la implementación de los resultados, es necesario contar con un plan detallado.

c) Producción de un reporte final

d) Revisión del proyecto: Se debe evaluar lo que ha resultado adecuado y los ítems que necesitan mejoras.

A continuación se muestra un esquema con las fases del proceso CRISP DM, con las etapas involucradas comentadas anteriormente.

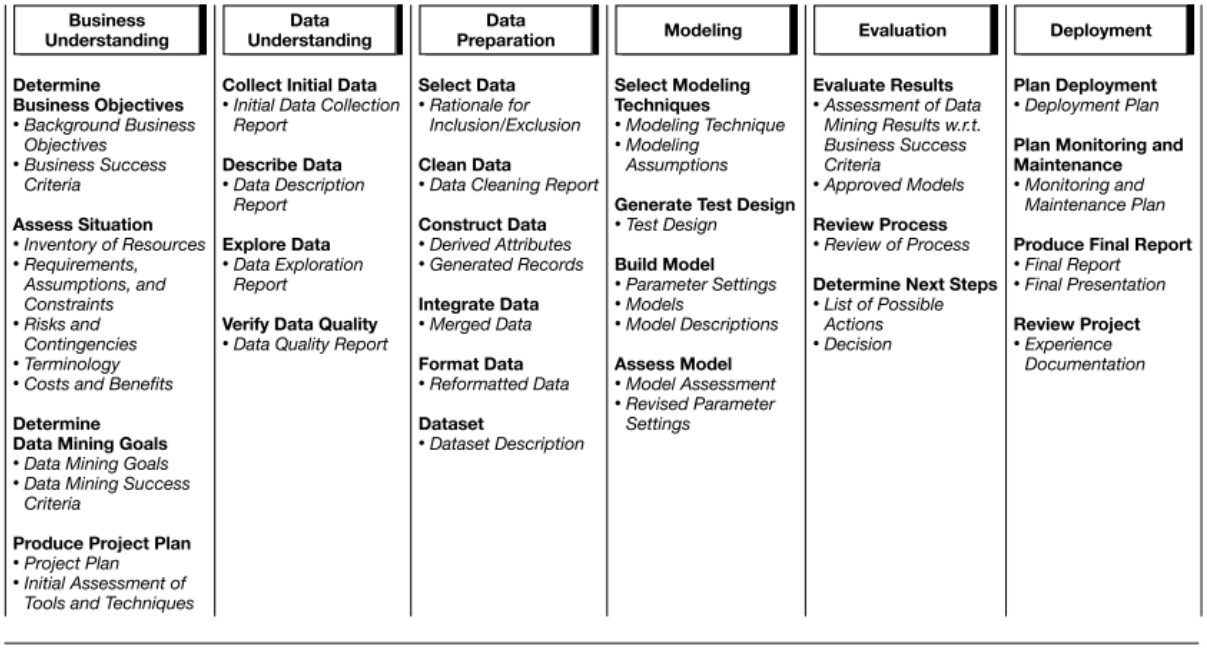


Figura12: Tareas Genéricas y resultados del modelo de referencia CRISP-DM, libro “CRISP DM 1.0 Stepbystep data miningguide” *[32].*

3.3 Metodología específica

Usando como base el modelo de referencia CRISP-DM, se desarrolló una metodología específica para el análisis y la clasificación de los distintos contenidos que son mencionados por los emprendedores cuando presentan sus planes de negocio. Esto con el objetivo de predecir la potencial generación de empleos que cada uno tendrá. Esta metodología se detalla a continuación.

Los objetivos de esta etapa a nivel general son desarrollar las herramientas para poder analizar una gran cantidad de planes de negocio en formato digital. Esto mediante diferentes pre-procesos que serán realizados sobre los documentos con el objetivo de darle una estructura que permita analizarlos de una manera correcta. A nivel específico, tenemos que los objetivos son:

- La transformación de archivos de texto (datos no estructurados), o sea los planes de negocios, a datos estructurados en donde cada documento puede ser descrito como un grupo de características, lo que permitirá su análisis de manera directa.
- La creación de recursos, como ontologías o diccionarios, que permitan darle sentido humano al texto con el objetivo de obtener mejores resultados.
- La creación de un modelo de textmining que permita extraer de manera automática los contenidos de los documentos para los que fue entrenado utilizando los recursos creados anteriormente.

Para cumplir estos objetivos, el trabajo se organizó en dos fases. Cada una contiene sub fases, aplicaciones y resultados específicos. En la primera fase, el objetivo principal es el de elaborar una metodología combinando tanto el análisis tanto de una máquina como de un ser humano en una muestra de ejemplos de texto extraídos de planes de negocio. El conocimiento generado en esta etapa será usado para analizar la muestra total de planes completos en donde los extractores creados serán testeados y adaptados, en caso de ser necesario. La segunda fase consistirá en la aplicación del modelo creado en la primera fase a una muestra de planes que el modelo no ha visto, lo que nos entregara información útil para elaborar el reporte final. A continuación se muestran las fases de manera ordenada y con una explicación más detallada:

Fase 1: Metodología y desarrollo de prueba de concepto

- **Sub-Fase 1.1: Clasificación humana y análisis de la muestra.**

  Una muestra de planes de negocio del concurso InfoDev fue clasificada de manera manual en alto, medio o bajo dependiendo de su potencial creación de empleo. Además, los analistas no sólo clasificaron cada documento sino que también extrajeron un ejemplo de texto que explicara la clasificación escogida. Esto, con el objetivo de generar una base de datos que tuviera ejemplos de texto, los que luego serán usados para entrenar el modelo por medio de la identificación de conceptos clave que ayudarán a discriminar entre los distintos valores.

- **Sub-Fase 1.2: Desarrollo y Refinamiento del modelo de Text Mining para la muestra.**

  Los ejemplos de texto entregados por el analista en la fase anterior son agrupados en categorías dependiendo de sus similitudes generando diccionarios y una ontología respecto de conceptos mencionados en los planes de negocios que se refieran al potencial de creación de empleo. Luego, se entrena un software de Text Mining (IBM SPSS Modeller v.15) para extraer de manera automática estos conceptos clave desde el texto.
  El resultado del paso anterior es una matriz de confusión en donde cada fila representa un plan de negocio y cada columna un concepto y/o categoría presente en los textos. En cada celda, un valor de 1 significa que el concepto se encuentra presente en el texto, mientras que un 0 significa lo contrario. Esto resulta bastante útil al momento de conocer las características que subyacen la creación de empleo. Un ejemplo de esto último, es que si la categoría "industria" se encuentra presente en el texto, es posible deducir que en el plan se mencionan oportunidades de empleo relacionadas con una industria en particular, como por ejemplo la industria de la agricultura o de la manufactura.

  Usando la matriz de confusión creada y los documentos manualmente clasificados, es posible enseñarle a la matriz a distinguir planes con alto, medio o bajo potencial de

creación de empleo basado en la presencia de conceptos y/o categorías. El paso final de esta etapa es la creación de un modelo, en este caso un árbol de decisión que aprenda a discriminar los planes y que pueda después ser aplicado a una muestra mayor de documentos. Las reglas de clasificación que usa el árbol de decisión pueden usarse para interpretar que conceptos y categorías son más importantes a la hora de discriminar entre los documentos. Esto puede ser bastante útil al momento de analizar y obtener un conocimiento más a fondo de las causas que subyacen la potencial creación de empleo.

Fase 2: Desarrollo y análisis automático de documentos

- **Sub-Fase 2.1: Aplicación del modelo refinado de Text Mining en la muestra total de documentos.**

  Sabiendo que palabras son las que más discriminan entre los distintos planes de negocio y teniendo el árbol de decisión generado, el siguiente paso es aplicarlo a una muestra de planes de negocio que el modelo no haya visto antes. Estos documentos serán escaneados en busca de conceptos clave y clasificados en alto, medio o bajo. Los resultados mostraran que tan bien o mal es capaz de discriminar el modelo y que categorías son las más importantes a la hora de crear empleo en este concurso.

  El modelo de Text Mining y el árbol de decisión pueden ser aplicados a otro concurso. Esto, permitiría al analista obtener un mayor conocimiento del nivel real de extrapolación que tiene el modelo. Incluso, el analista podría crear un segundo modelo de Text Mining para el segundo concurso y fundirlo con el primero. Probablemente, al aplicar los dos modelos la máquina será capaz de distinguir de mejor manera la potencial creación de empleo de un plan de negocio.

- **Sub-Fase 2.2: Análisis estadístico de los contenidos y conceptos extraídos.**

Con los valores predichos y los reales, el analista puede entender el rendimiento de la máquina discriminando entre los diferentes planes de negocio. Métricas como la precisión, Class Precision o Class Recall pueden ser realmente útiles para entender, no sólo si predije de manera correcta o incorrecta, sino que también cómo se comportan las predicciones en cada clase en particular. El modelo puede ser evaluado con un gráfico de ganancia en donde se le puede comparar con una predicción aleatoria y generar un indicador visual del rendimiento del modelo.

Una comparación bastante interesante es la que se puede hacer con un modelo de Text Mining no lingüístico, porque esto nos permitiría comprender el nivel de entendimiento que tenía el analista mientras desarrollaba el modelo y como se puede mejorar este entendimiento para mejorar los resultados. Además, cada predicción tiene un score asociado, el cual representa la probabilidad de estar correcto o la confianza de la predicción. Este indicador permite al analista testear como el modelo funciona en un grupo de planes de negocio (podría ser en la cola de una distribución de planes, por ejemplo).

Algunos concursos tienen sus propios clasificadores de creación de empleo, los resultados del análisis permiten al analista comparar los resultados de la máquina con los de los jueces de cada concurso. Esto permitiría comprender mejor como comprende cada concurso la creación de empleo en particular.

- **Sub-Fase 2.3: Elaboración del reporte final.**

La elaboración de un reporte final debe contener y explicar la predicción de los planes que el modelo no ha visto por medio de diferentes métricas como la precisión, class precisión y recall. Además, debe mostrar el rendimiento del modelo versus un modelo no lingüístico y otro aleatorio, con el objetivo de comprender su funcionamiento y de esta manera posibles maneras de mejorarlo. Por otro lado, también debe contener el árbol de decisión ya que este permite comprender cuales fueron las características más importantes a la hora de discriminar entre los diferentes planes de negocio, lo que es

crucial a la hora de concluir a cerca de un tema tan complejo como lo es la creación de empleo. Otro output que es muy importante es la librería de conceptos y reglas, los cuales son la base para el funcionamiento del modelo y para potenciales trabajos futuros.

Lo descrito anteriormente resume la metodología que se aplicara en esta tesis, dividida según sus fases y sub-fases. Además presenta una explicación de cada una.

3.4 Algoritmos a utilizar

Para esta tesis se utilizaron principalmente 2 algoritmos, los cuales son Text Mining y árboles de decisión. El primer algoritmo fue definido anteriormente en la revisión literaria, debido a que su definición era necesaria para explicar las distintas aplicaciones del mismo en la literatura existente. Por otro lado, los árboles de decisión cuentan con variadas ventajas, las cuales basaron la decisión de usar este método de clasificación en desmedro de otros. Las principales ventajas se enumeran a continuación:

- A distinción de modelos de "caja negra", como podrían serlo redes neuronales o *Support Vector Machines*, los árboles de decisión muestran el procedimiento realizado, explicando cómo se llegó al resultado. Esto es de vital importancia si buscamos obtener las razones por las cuales se obtuvo un determinado resultado.
- De todas las variables existentes, el árbol de decisión no usa las que no considera relevantes. Esto es ventajoso ya que nos permite darnos cuenta de cuales variables ayudan y cuáles no, pudiendo con esto reducir la base.
- Entrega un orden de las variables que si sirven, dada su forma de árbol en donde lo explica de manera visual.

De acuerdo de Berlanga, Rubio & Vilá (2013) [36], los árboles permiten examinar resultados y determinar visualmente como fluye el modelo. Estos resultados visuales ayudan a buscar subgrupos específicos y relaciones que probablemente no encontraríamos con estadísticos más tradicionales. A continuación se describe en más detalle esta técnica.

3.4.1 DecisionTrees

Se trata de una técnica predictiva que se utiliza en problemas de clasificación, *clustering* y predicción que utiliza el enfoque de “dividir y conquistar” para fragmentar el espacio de búsqueda en sub grupos o representarlo en una jerarquía condicional en un determinado sistema. Existen diversas maneras y algoritmos para construir los árboles como el denominado CART (*Classification and RegressionTrees*), Bosques Aleatorios, C4.5, ID3 entre otros. CART y Bosques Aleatorios son utilizados generalmente para regresión, mientras que C4.5 e ID3 para tareas de clasificación. Es posible describir genéricamente este algoritmo como sigue:

1) Supongamos que “N” es el número de casos de entrenamiento y “M” el número de variables en el clasificador.

2) Determinar “m”, el número de variables input a ser usadas para determinar la decisión en un nodo del árbol (m debe ser mucho menor que M).

3) Elegir un set de entrenamiento escogiendo n veces con reemplazo desde todas las N clases disponibles de entrenamiento (muestra por bootstrapping). Utilizar el resto de los casos para estimar el error del árbol por medio de la predicción de sus clases.

4) Por cada nodo del árbol, aleatoriamente, escoger m variables sobre las cuales basar la decisión en ese nodo. Calcular la mejor partición basado en esas m variables en el grupo de entrenamiento.

El proceso de construcción de un árbol de decisión implica enfrentar varios temas, dentro de los más importantes se cuenta el tamaño de la muestra de entrenamiento y como se escoge el mejor atributo de separación. Los siguientes temas son los enfrentados en la mayoría de los procesos de construcción de un árbol de decisión:

1) Escoger los atributos de separación: No todos los atributos son iguales, hay algunos que sirven de mejor manera que otros para la tarea descrita. La elección de uno en particular no sólo implica el examen de los datos sino que también el criterio de un experto.

2) Orden de los atributos de separación: El orden en que se escogen los atributos también es importante, por ejemplo elegir el atributo “industria” antes que “oportunidades de empleo” puede alterar el desempeño del árbol, pues puede evitar comparaciones innecesarias.

3) Separaciones asociadas con el orden de los atributos: Cuando el dominio es pequeño, por ejemplo para “genero”, no existen muchas separaciones. Sin embargo, si el dominio es contiguo o tiene un gran número de valores, el número de separaciones a usar no es trivial.

4) Estructura del árbol: En el mejor de los casos, lo que se desea es un árbol balanceado con la menor cantidad de niveles posibles.

5) Criterio de detención: La creación del árbol determina de manera definitiva cuando los datos se encuentran perfectamente clasificados, sin embargo pueden existir situaciones donde una detención temprana es deseable debido a que previene la creación de un árbol más grande y complejo.Básicamente, esto es un *trade-off* entre precisión de la clasificación y desempeño. Además, detenerse de manera temprana puede ser útil para prevenir el sobre ajuste.

6) Datos de entrenamiento: La estructura del árbol creado depende de los datos de entrenamiento. Si el número de datos de entrenamiento es muy pequeño, el árbol podría no ser lo suficientemente específico para trabajar de manera adecuada con bases más grandes. Por otro lado, si el número de datos de entrenamiento es muy grande, el árbol puede sobre ajustarse a los datos.

7) Poda: Una vez que el árbol ya fue construido, es posible que sean necesarias algunas modificaciones para mejorar s desempeño durante la fase de clasificación. Este proceso puede remover comparaciones redundantes o sub-árboles con el fin de lograr un mejor rendimiento.

Como fue señalado anteriormente, los árboles de decisión pueden ser ejecutados por medio de la utilización de distintas aproximaciones. A continuación se describen las más importantes:

1) **ID3:** Esta técnica se basa en la teoría de información y tiene como principio básico la minimización del número esperado de comparaciones. La estrategia básica para ello consiste en escoger atributos de división con la mayor ganancia de información primero. A su vez, la cantidad de información asociada con un valor de atributo está asociada con la probabilidad de ocurrencia.

   El concepto utilizado para cuantificar la información se denomina entropía. Esta es utilizada para medir la cantidad de aleatoriedad o incertidumbre en los datos. Cuando todos los datos pertenecen a una única clase no existe incertidumbre y por ende la entropía es cero. El objetivo de la clasificación en un árbol es dividir iterativamente los grupos de datos en subgrupos donde todos los elementos de cada subgrupo final pertenecen a la misma clase.

   Desde un enfoque más formal, dadas las probabilidades$p_1, p_2, \dots, p_s$, donde $\sum_{i=1}^{s} p_{i=1}$, la entropía se define como:

$$H(p_1, p_2,..p_s) = \sum_{i=1}^{s}(p_i \log(\frac{1}{p_i})) \quad (1)$$

   Dado el estado de una base de datos D, H(D) encuentra la cantidad de orden en ese estado. Cuando ese estado es dividido en s nuevos estados $S = D_1, D_2, \dots D_s$, es posible mirar la entropía existente en esos estados. Cada paso del ID3 escoge el estado que ordena de mejor manera la división. El estado de una base de datos está completamente ordenado si todas las tuplas dentro de él están en la misma clase. ID3 escoge el atributo de división con la mayor ganancia en información, donde ganancia es definida como la

diferencia entre cuanta información es necesaria para hacer una correcta clasificación antes de la división vs cuanta información se necesita después de la división. Claramente, la división debería reducir la cantidad de información necesaria por la mayor cantidad posible. Esto se calcula mediante la determinación de las diferencias entre las entropías del set de datos original y la suma ponderada de las entropías de cada una de las subdivisiones de datos. Las entropías de las subdivisiones de datos son ponderadas por la fracción del set de datos que es colocada en cada división. El algoritmo ID3 calcula la ganancia de una división particular por medio de la siguiente formula:

$$Ganancia\ D,S\ = H\ D\ -\ \sum_{i=1}^{s} P\ D_i\ H\ D_i \qquad (2)$$

Es importante destacar que es necesaria la división en rangos cuando el dominio de un atributo es continuo o contempla una gran cantidad de valores, división que tendrá algún grado de arbitrariedad, por lo que debe ser llevada a cabo de acuerdo al dominio experto.

2) **C4.5 y C5.0:** El algoritmo C4.5mejora al ID3 en la siguiente maneras:

   a) Datos faltantes: Cuando se construye el árbol de decisiones, los datos faltantes son ignorados. El ratio de ganancia se construye considerando, sólo por medio de observación, los registros que si tienen valores para ese atributo. Para clasificar un registro con un valor perdido el valor para ese ítem puede ser predicho basado en lo que se conoce sobre los valores del atributo para otros registros.

   b) Datos continuos: La idea básica es dividir los datos en rangos basado en los valores de atributos para ese ítem que fue encontrado en la muestra de entrenamiento.

   c) Poda: Básicamente, existen 2 estrategias de podado en C4.5:

      1) Mediante el uso de un sub-árbol de reemplazo, el cual es reemplazado por un "nodo hoja" si este reemplazo resulta en una tasa de error cercana a la del árbol original. El sub-árbol de reemplazo trabaja desde la parte baja hacia la raíz.

2) Otra estrategia de podado llamada "levantamiento de sub-árbol", reemplaza un sub-árbol por su más usado sub-árbol. Aquí un sub-árbol es levantado desde su ubicación actual a un nodo superior en el árbol. Nuevamente, se debe determinar el incremento en la tasa de error por este reemplazo.

d) Reglas: C4.5 permite la clasificación vía árboles de decisión o bien por las reglas generadas por estos mismos. Además, se proponen algunas técnicas para simplificar reglas complejas. Una aproximación es reemplazar el lado izquierdo de una regla por una versión más simple, si todos los registros en los datos de entrenamiento son trabajados idénticamente.

e) División: El enfoque ID3 favorece atributos con muchas divisiones pudiendo así conducir a una sobre especificación. En el extremo, un atributo que tiene un valor único para cada tupla en el set de entrenamiento podría generar una clase. Este enfoque usa el ratio de ganancia de modo opuesto a la ganancia. Este ratio se define como:

$$Ratio\ de\ Ganancia\ \ D,S\ \ = \frac{Ganancia(D,S)}{H(\frac{D_1}{D},\frac{D_2}{D},\ldots,\frac{D_S}{S})} \quad (3)$$

Con el objetivo de realizar la división, C4.5 utiliza el mayor ratio de ganancia que asegura un incremento en la información mayor que el promedio. Esto es para compensar por el hecho de que el ratio de ganancia es sesgado hacia las divisiones donde el tamaño de un subgrupo es más cercano al inicial.

Una de las mejoras que propone el 5.0 en precisión está basada en el *boosting*, el cuales un enfoque para combinar diferentes clasificadoras. Se ha demostrado que el error es menos que la mitad que se encuentra con un C4.5 en algunas bases de datos. *Boosting* no siempre ayuda cuando los datos de entrenamiento contienen mucho ruido, debido a que trabaja por medio de la creación de múltiples grupos de entrenamiento provenientes desde uno sólo; a cada ítem en el set de entrenamiento le es asignado un peso, el cual

indica la importancia del ítem en la clasificación. Un clasificador es construido para cada combinación de pesos utilizados, creando múltiples clasificadores en la práctica. Cuando C5.0 ejecuta una asignación a cada clasificador le es asignado un voto, la votación es ejecutada y la tupla target es asignada a la clase con mayor número de votos.

3) **CART (Árboles de clasificación y regresión):** Esta técnica genera un árbol de decisiones binarias. Tal como el ID3, la entropía es utilizada como una medida para escoger el mejor atributo de división y criterio. A diferencia del ID3, donde un hijo es creado para cada sub-categoría, esta técnica crea sólo 2 hijos, la división es ejecutada alrededor de lo que se determina ser el mejor punto de división. En cada paso se utiliza una búsqueda exhaustiva para determinar la mejor división donde "mejor" está definido por:

$$\theta\ \frac{s}{t}\ = 2P_l P_r\ \ \sum_{j=1}^{m}\ P\ \ {}_{t_l}^{C_j}\ - P\ \ {}_{t_r}^{C_j} \qquad (4)$$

Esta fórmula es evaluada en el nodo actual t, y para cada posible atributo de división y criterio s. En esta notación, L y R son usados para indicar los sub árboles de la izquierda y derecha del nodo actual del árbol. P_L yP_r son la probabilidad que una tupla en un set de entrenamiento esté en el lado derecho o izquierdo del árbol, lo cual se define como:

$$\frac{tuplas\ en\ el\ sub\ árbol}{tuplas\ en\ el\ set\ de\ entrenamiento} \qquad (5)$$

Se asume que las dos ramas son tomadas en igualdad. $P\ C_j\ t_L$ ó$P\ C_j\ t_R$ corresponde a la probabilidad que una tupla esté en la clase C_j y en el lado izquierdo o derecho del sub-árbol, lo cual está definido como:

$$\frac{tuplas\ de\ clase\ j\ en\ el\ sub\ árbol}{tuplas\ en\ el\ nodo\ objetivo} \qquad (6)$$

En cada paso, sólo un criterio es escogido como el mejor por sobre todos los otros posibles criterios. Finalmente, es importante señalar que el objetivo en este apartado no es profundizar en el algoritmo, sino más bien esbozar los principios básicos detrás del algoritmo a utilizar en esta tesis.

4. Resultados

A continuación se muestran los resultados obtenidos, divididos según las fases desarrolladas en la metodología específica.

4.1 Resultados obtenidos

Fase 1: Metodología y desarrollo de prueba de concepto

- **Sub-Fase 1.1: Clasificación humana y análisis de la muestra.**

La primera etapa de este análisis fue la de leer los planes de negocio uno a uno, categorizarlos en base a su potencial para crear empleo y extraer ejemplos de texto que pudieran explicar esta clasificación. De estos últimos, después se extraerán conceptos que serán agrupados en subcategorías y categorías, las cuales serán la base para analizar el potencial de empleo de futuros planes de negocio.

Esta clasificación manual fue aplicada a 615 documentos del concurso “InfoDev”, en donde lo primero fue clasificar manualmente cada plan de negocio en ‘Low’, ‘Medium‘o ‘High’ dependiendo de su potencial para generar empleo. Luego, se seleccionó una parte del texto (una oración, párrafo o palabras) que explicasen, según el criterio del analista, la clasificación entregada. Con esto, se obtuvo una base de oraciones, párrafos o palabras que representaban la potencial creación de empleo.

En la Figura se muestra la base de datos que se generó. Su estructura consiste en el ID del plan de negocio, el nombre del archivo, la clasificación que se le asigno y el ejemplo de texto que la explica, si aparece “No information” es porque no existían ejemplos de texto que pudieran explicar esta clasificación.

	A	B	C	D
1	ID	File name	N-potential workers	Selection criteria "N-potential workers" (can use part of the text / if it is necessary)
2	1	1 Swivel Displays Solutions	Low	No Information
3	2	2 MS ApeandSuperApe Entertainment & InfoTech	Medium	No Information
4	3	3. the tecchNology incubation center Maiduguri	Medium	"creation of jobs and with the ultimate objective of reducing poverty as well as make the community self reliant"
5	5	5. Community Engineering Programme	Medium	No Information
6	6	6. Splynx Jogos	High	"employ more workers", "increase employment for the people"
7	7	7. Makeys Computing	Medium	"create jobs", " increase productivity and employment opportunities"
8	8	8. iletken_TechNologies	Low	No Information
9	9	9. MiniMax	Medium	"Producing jobs"
10	10	10. FamilyClick	Low	No Information
11	11	11. t-mega strategy consulting	Low	No Information
12	12	12. Red ProBuy	Low	No Information
13	13	13. OffTrackPlanet	Low	No Information
14	14	14. Artin Dynamics	Medium	"The biggest development benefit would be the employment benefit"
15	15	15. Visual NACert	Medium	"requires the creation of a large number of jobs"
16	16	16. Mixed Dimension	Low	No Information
17	17	17. Trustmark Communications	Medium	"provide employment opportunities"
18	18	18. Kidbox	Low	No Information
19	19	19. developWay	Low	No Information
20	20	20. WANDYFOODS	Medium	"PROVIDE EMPLOYMENT OPPORTUNITIES", "EMPLOY MORE PEOPLE"
21	21	21. Xcode Life Sciences	Medium	No Information

Figura 13

- **Sub-Fase 1.2: Desarrollo y Refinamiento del modelo de Text Mining para la muestra.**

En esta etapa, la base de datos con los ejemplos de textos fue usada para crear un modelo lingüístico que permitiera predecir el potencial de empleo de un plan de negocio. Con el objetivo de entrenar el modelo, la base de datos fue dividida en una muestra de entrenamiento y otra de prueba. Para la primera se usaron 240 documentos aleatoriamente seleccionados, mientras que los 375 restantes se usaron para probarlo.

Usando los ejemplos de texto de los documentos de la muestra de entrenamiento, el analista buscó manualmente patrones que pudieran usarse para obtener palabras claves o conceptos que los representasen. Por ejemplo, el plan número 6 (Figura 13) contenía la frase “employ more workers”, mientras que el plan 20 contenía “EMPLOY MORE PEOPLE”. Es posible observar que estas 2 frases son bastante similares por lo que podrían ser agrupadas con una regla de extracción que estableciera que si aparecía el concepto “employ more” y luego la palabra “worker” o “people” se asociara directamente con un potencial de empleo “High” o “Medium”. Este proceso se repitió hasta que todos los ejemplos de texto y sus posibles combinaciones (en la

medida que lo ameritara) estuvieran agrupados en familias de conceptos que representasen ideas o significados similares.

Estos patrones fueron representados mediante reglas de extracción, las cuales son las que usará la máquina para diferenciar entre los distintos planes de negocios. Primero, el analista creó reglas de extracción que representasen conceptos fácilmente entendibles desde el texto, esto se refiere a que cada palabra clave fuera asociada directamente a una categoría, ejemplos de esto pueden ser: “employmentgeneration”, “jobopportunities” o “jobcreation”. Luego de esto, se crearon reglas más complejas, las que usualmente mezclan reglas de extracción más simples. Ejemplos de lo anterior, son conceptos o palabras clave que no se pueden relacionar directamente con creación de puestos de trabajo en el texto, como por ejemplo “manufacturingindustry” o “agriculturalindustry”, las cuales no mencionan que generan empleo pero que se podría inferir que si lo hacen.

En la siguiente etapa, reglas de extracción similares se agrupan para crear tipos. Un ejemplo de esto se puede observar en laFigura 1. Reglas de extracción para los conceptos “manufacturinginustry” and “grahaagriindustry” se clasificaron en el tipo “Industria”. Esto permite al analista crear reglas de extracción que usen definiciones genéricas de tipos, en vez de referirse a un concepto en específico.

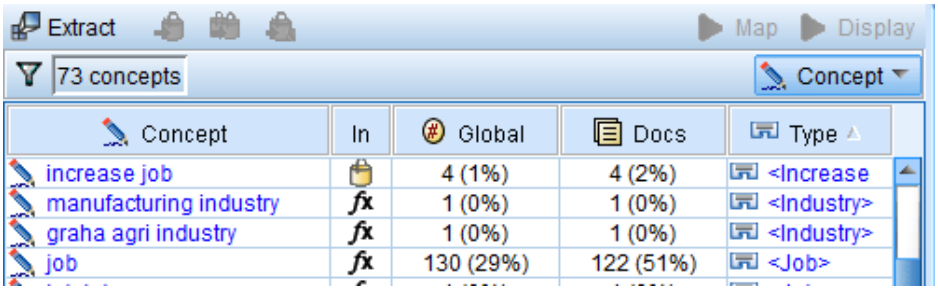
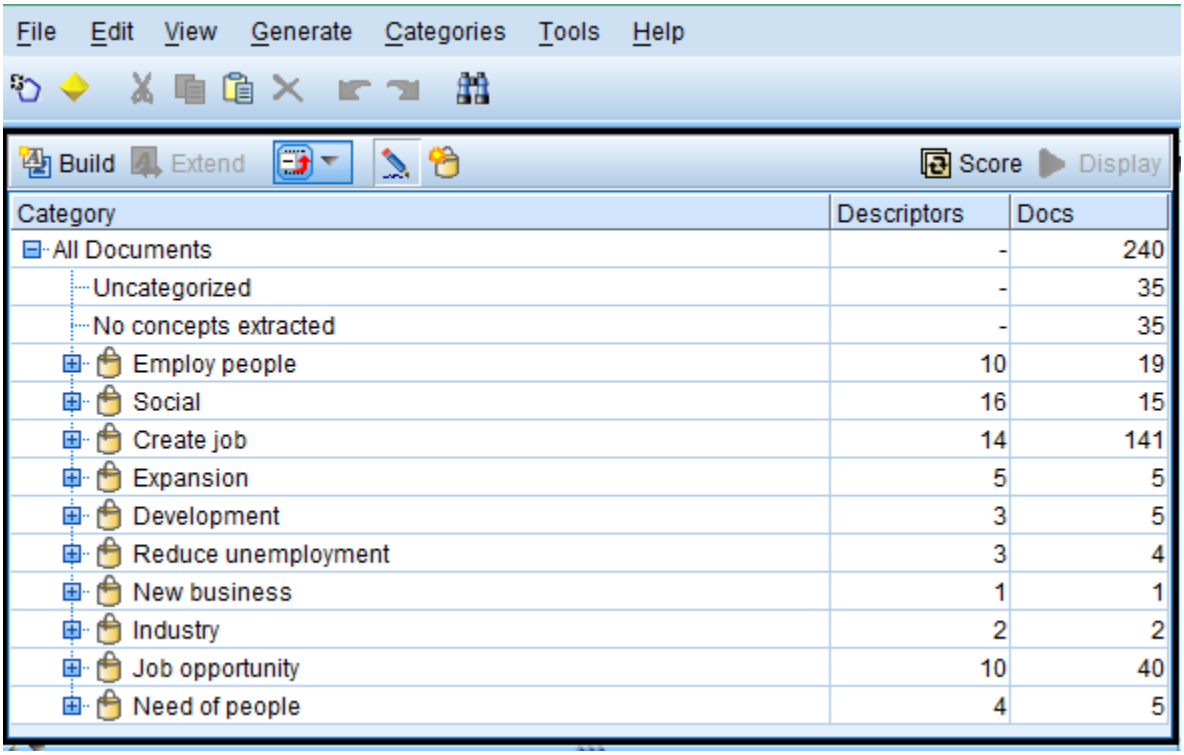


Figura 14

Además, una nueva clasificación fue realizada agrupando diferentes tipos creando subcategorías de conceptos. Por ejemplo, los tipos “job” y “increasejob” pertenecen a la subcategoría “jobopportunity”. Finalmente las subcategorías fueron nuevamente agrupadas en supra categorías, las que se relacionan directamente con características de la creación de empleo. En el apéndice 2 se pueden observar los conceptos y sus reglas de extracción respectivas. Como se puede ver en laFigura, el analista obtuvo 10 categorías (familias de conceptos). Es importante

notar que cada categoría, cuenta con reglas de extracción y conceptos relacionados con el mismo tópico.



Category	Descriptors	Docs
All Documents	-	240
Uncategorized	-	35
No concepts extracted	-	35
Employ people	10	19
Social	16	15
Create job	14	141
Expansion	5	5
Development	3	5
Reduce unemployment	3	4
New business	1	1
Industry	2	2
Job opportunity	10	40
Need of people	4	5

Figura 15

En la Figura2se encuentra un ejemplo de la categoría “Job opportunity”. Dentro de ella fueron creadas 6 subcategorías, las cuales son “New job”, “Variety”, “Rural job”, “Exciting”, “Offer” y “Create”. La figura muestra además el número de reglas de extracción que fueron creadas y el número de documentos en los que aplica.

Category	Descript...	Docs
All Documents	-	240
Uncategorized	-	35
No concepts extracted	-	35
Job opportunity	10	40
job opportunities		27
New job	1	6
new job opportunities		6
Variety	1	1
provide & variety job opportunities		1
Rural job	1	1
rural job opportunities		1
Exciting	1	1
create & exciting job opportunities		1
Offer	2	2
offers job opportunities		1
offer job		1
Create	3	22
create & job opportunities		10
create & opportunities		2
provide & job opportunities		10

Figura 16

Para aumentar la confiabilidad del estudio, 2 analistas crearon 2 modelos de manera independiente. Cada modelo tiene sus propias categorías, subcategorías y tipos. Ambos modelos fueron usados de manera conjunta para entrenar a la máquina con el objetivo dediscriminar entre los distintos planes de negocios. Las categorías creadas en ambos modelos se muestran en la Tabla1. El apéndice 1 muestra la taxonomía de ambos modelos.

Model 1	**Model 2**
Create jobs	Create job
Employ people	Development
Employment opportunity	Employ People
Human Resources	Expansion
Industry	Industry
Internationalization process	Job opportunity

Social	Need of people
	New business
	Reduce unemployment
	Social

Tabla1

En total, el modelo 1 tiene 25 tipos, 55 subcategorías y 7 categorías. El modelo 2 cuenta con 46 tipos, 34 subcategorías y 10 categorías.

Sorprendentemente, ambos analistas encontraron o nombraron categorías de manera bastante similar. De hecho, 5 categorías se encuentran presentes en ambos modelos, 4 incluso con el mismo nombre. El modelo 2 identifico más categorías, pero los conceptos de ellas se encuentran también en el modelo 1 como tipos o subcategorías. La gran concordancia entre los analistas fue usada como un criterio para la alta confiabilidad del análisis.

Estos modelos, fueron usados para leer los planes en la muestra total (615 documentos del concurso InfoDev) y buscar los conceptos con que habían sido entrenados. Es importante considerarque los modelos fueron entrenados solamente usando los ejemplos de texto de la muestra de entrenamiento (240 documentos) mientras que en esta etapa fueron aplicados a los documentos completos.

En particular, el concurso InfoDev consiste en un formulario donde los interesados responden preguntas sobre características específicas de sus negocios. De esta forma, estas preguntas representan ejemplos de texto que son comunes para todos los planes. Es por esto, que es esperable que los modelos extraigan estos conceptos pese a que no sean relevantes para determinar el potencial de empleo. Para resolver este tema, todo el texto que no correspondiera a una respuesta de los interesados fue removido (principalmente la introducción y las preguntas). Con este proceso realizado, el analista pudo estar seguro de que todos los conceptos reconocidos por el modelo representaban respuestas de los aplicantes.

El resultado de esta etapa es una matriz de dispersión, en donde cada fila representa las respuestas de los aplicantes en el plan de negocio y cada columna una categoría binaria. Un valor de 1 en una columna, indica la presencia del término de la categoría en ese documento y un valor de 0 indica lo contrario. La Figura 293muestra parte de la matriz de dispersión obtenida.

Business_Plan_Type	Doc ID	Create jobs	Create jobs/Many job opportunity	Create jobs/Many jobs	Create jobs/Produce jobs	Create jobs/job	Create jobs/job create	Create jobs/job creator	Create jobs/job opportunity	Create jobs/new job	Employ people
Low	133	1	0	1	0	1	1	0	0	0	1
Low	134	1	0	0	0	1	0	0	0	0	1
Low	135	1	0	0	1	1	0	0	0	0	1
Low	136	1	0	0	0	1	1	0	0	0	1
Low	137	0	0	0	0	0	0	0	0	0	0
Low	138	0	0	0	0	0	0	0	0	0	0
Low	139	1	0	1	0	1	1	0	0	0	0
Low	140	1	0	0	1	1	1	0	0	0	1
Low	141	1	0	1	0	1	1	0	0	0	1
Low	142	1	0	0	0	1	1	0	0	0	1
Low	143	1	0	0	1	1	0	0	0	0	0
Low	144	1	0	1	1	1	0	0	0	0	1
Low	145	1	0	1	1	1	1	0	0	0	1
Low	146	1	0	0	0	0	1	0	0	0	0
Low	147	1	0	1	0	1	1	0	0	0	0
Low	148	1	0	0	1	1	0	0	0	0	1
Low	149	1	0	0	0	1	1	0	0	0	0
High	150	1	0	0	1	1	1	0	0	0	0
High	151	1	0	1	1	1	1	0	0	0	1
High	152	1	0	0	0	0	1	0	0	0	1
High	153	1	0	1	0	1	0	0	0	0	0
High	154	1	0	0	0	0	0	0	0	1	1
High	155	1	0	0	0	1	0	0	0	0	1
High	156	0	0	0	0	0	0	0	0	0	0
High	157	1	0	1	1	1	1	0	0	0	1
High	158	1	0	1	0	1	1	0	0	0	1
High	159	1	0	0	0	1	1	0	0	0	1
High	160	1	0	0	0	1	0	0	0	0	0
High	161	1	0	0	0	0	0	0	0	0	1
High	162	1	0	0	1	1	1	0	0	0	1
High	163	0	0	0	0	0	0	0	0	0	0
High	164	0	0	0	0	0	0	0	0	0	0
High	165	1	0	0	0	1	0	0	0	0	0
High	166	1	0	0	1	1	1	0	0	0	1
High	167	1	0	1	1	1	1	0	0	0	0
High	168	1	0	1	1	1	1	0	0	0	1
High	169	1	0	1	0	1	1	0	0	0	1

Figura 17

Construyendo un clasificador automático mediante árboles de decisión

La matriz contiene 615 filas que fueron divididas en un 80% como una nueva muestra de entrenamiento (477 filas) y un 20% como nueva muestra de prueba (138 filas). Con esto realizado, el objetivo es crear un nuevo modelo que automáticamente clasifique cada plan de negocios como “Low” o “High-Medium”, en términos de la potencial creación de empleo, según la clasificación manual realizada en la **SubFase 1.1**.

Dado el hecho, de que en la muestra de prueba existe un número no balanceado de documentos en cada clase (véaseFigura) y de que el número de documentosen la clase “High-Medium” era bastante mayor al de la clase “Low”, se utilizó una técnica de sobre muestreo. Esto consiste en multiplicar el número de documentos en la clase “Low” por un factor que permita igualar el número de planes en ambas clases, el resultado se puede observar en laFigura 314.

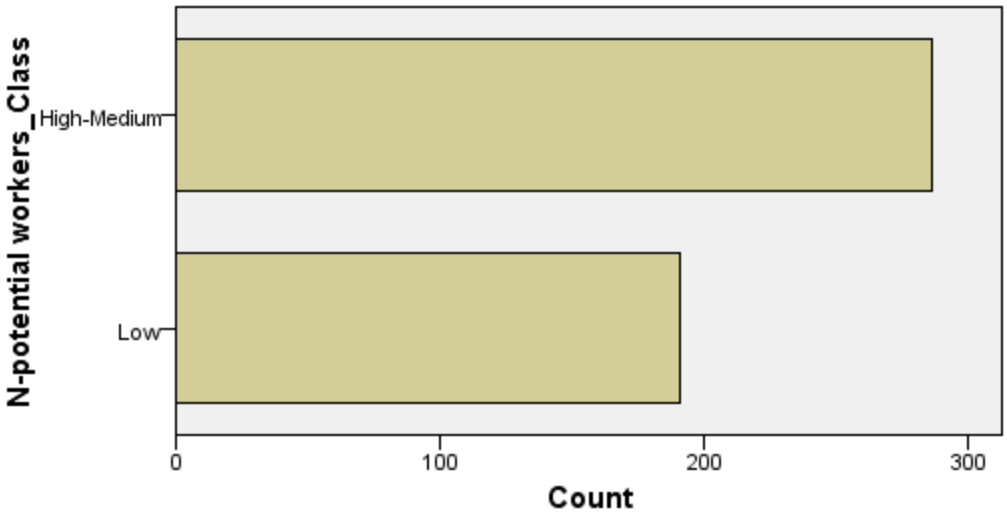


Figura18

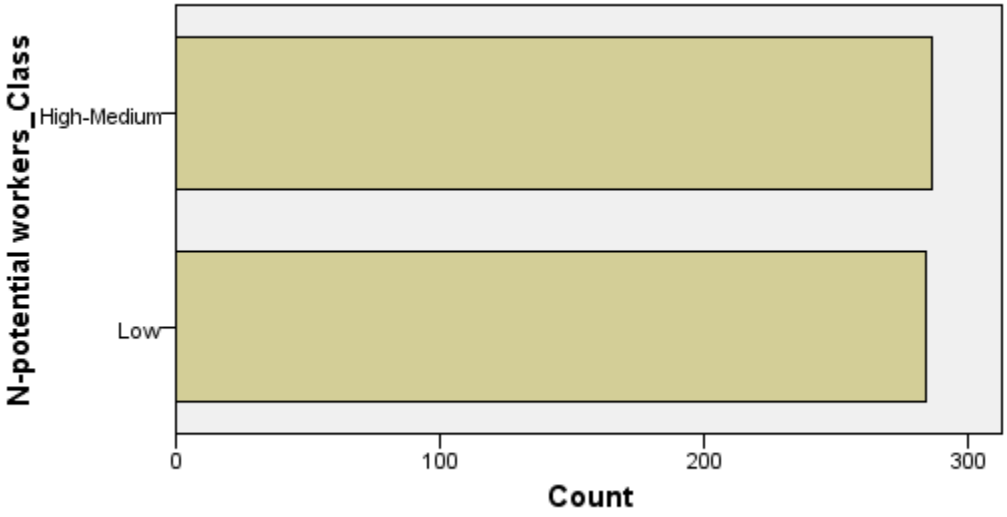


Figura 19

Distintos algoritmos de clasificación fueron entrenados usando esta muestra balanceada (Support Vector Machines, Árboles de Decisión, Redes Neuronalesy Regresiones, entre otros). Finalmente

se eligió el árbol de decisión debido a que es capaz de mostrar las categorías que son más importantesa la hora de predecir el potencial de empleo de un plan de negocio sin decaer en términos de desempeño. La Tabla2 muestra el rendimiento del modelo (árbol de decisión) en la muestra de entrenamiento balanceada.

La precisión de los modelos se mide como el número de veces que son capaces de clasificar un plan de negocio en la misma categoría que lo hizo el analista. En la Tabla2 se muestra una matriz de confusión en donde cada fila muestra el número de documentos clasificados en la misma categoría. Las filas muestran la clasificación que hicieron los analistas, mientras que las columnas muestran la clasificación realizada por la máquina. Si se observa la diagonal de izquierda a derecha, se puede apreciar el número de documentos clasificados correctamente. Class Recall es una medida que se usa bastante en tareas de clasificación e indica la proporción de casos correctamente clasificados sobre el total de posibles casos. A su vez, Class Precision representa el ratio entre los correctamente clasificados sobre el número de predicciones de cada clase hecha por la máquina. Ambas métricas son de vital importancia a la hora de evaluar un modelo de clasificación.

Por ejemplo, como se puede observar en la Tabla2, sobre 286 documentos que fueron manualmente clasificados como “High-Medium” (TRUE HIGH_MEDIUM) la máquina fue capaz de clasificar correctamente 219 o un 76.7% de los casos en esa clase. Para los documentos clasificados manualmente como “Low”, 271 (91,2%) fueron correctamente clasificados por la máquina.

	HIGH_MEDIUM	**LOW**	**RECALL**
TRUE HIGH_MEDIUM	219	67	76.6%
TRUE LOW	26	271	91.2%
CLASS PRECISION	89.4%	80.2%	

Tabla2

De la misma manera, la Tabla 3 muestra el rendimiento del modelo en la nueva muestra de prueba. Es importante tomar en consideración que en esta muestra no hubo proceso de balanceo

o nada similar, dado que cuando el modelo sea probado en la vida real, la base de datos no vendrá balanceada y además de esta forma se llegará a una mejor estimación del verdadero poder clasificador del modelo.

En la Tabla3, el Recall para el "High-Medium" alcanzo un 66.3% y un 69,2% para el "Low". Con respecto a la "Class precisión" es posible observar que la clase "High-Medium" fue clasificada de mejor manera que la clase "Low". El hecho de que el Recall de la clase "Low" sea bastante mayor al de la clase "High-Medium" nos indica que se necesitan varios falsos positivos para obtener un verdadero positivo.

	HIGH_MEDIUM	LOW	RECALL
TRUE HIGH_MEDIUM	63	32	66.3%
TRUE LOW	12	27	69.2%
CLASS PRECISION	84.0%	45.8%	

Tabla3

Analizando este resultado, se puede deducir que cuando el modelo predice la clase "Low" su rendimiento es peor que cuando predice "High-Medium". Esto proporciona varias e interesantes aplicaciones que se pueden discutir con mayor profundidad más adelante. Sin embargo, es posible argumentar que un modelo con estas características permite una clasificación rápida de los planes de negocio con la clase "High-Medium" en términos de la potencial creación de empleo. Además, si se compara con una predicción aleatoria, se observa que el modelo es en promedio 16.3% mejor que el 50% que se podría esperar de un clasificador naive.

Para establecer otro benchmark para esta clasificación, el analista entrenó además un modelo *no linguistico* de textmining y usó los conceptos automáticamente extraídos para replicar la tarea de clasificación. Como se muestra en la Tabla4, la predicción de este modelo alternativo es ligeramente peor que la del modelo *linguistico* cuando clasifico la clase "High-Medium", pero es casi igual que una muestra aleatoria cuando clasifico la clase "Low". Una posible explicación para la caída en la precisión ocurre porque este modelo basa sus predicciones sólo en conceptos que fueron seleccionados por la máquina en base a la frecuencia con que aparecían. No existe

una lógica humana en sus reglas de extracción lo que claramente lo limita a la hora de discriminar entre los documentos.

	HIGH_MEDIUM	‘LOW’	RECALL
TRUE HIGH_MEDIUM	63	32	66.3%
TRUE ‘LOW’	19	20	51.3%
CLASS PRECISION	76.8%	38.5%	

Tabla4

Interpretación del árbol de decisión en base a sus reglas.

Como se dijo anteriormente, una ventaja del árbol de decisión es que permite al analista entender en mayor profundidad las características de la creación de empleo. En particular, un árbol de decisión permite al investigador desarrollar un modelo de clasificación que pueda predecir o clasificar futuras observaciones en base a reglas. Si se tiene una base de datos dividida en clases (como por ejemplo “Low” y “High-Medium”), el algoritmo puede usar los datos previamente clasificados para crear reglas que se pueden usar para clasificar nuevos datos con un mejor rendimiento.

Este enfoque tiene variadas ventajas. Primero, el proceso de razonamiento detrás del modelo es evidente, lo que es muy bueno en contraste con modelos de caja negra, como podrían serlo *Support Vector Machines* o Redes Neuronales donde su lógica interna es bastante compleja. Segundo, el algoritmo escoge automáticamente los atributos que de verdad importan cuando se debe clasificar entre una clase y otra, en otras palabras, el árbol se construye de tal manera que los atributos que no contribuyen a mejorar el rendimiento del modelo son omitidos. Esto puede contener información muy importante sobre los patrones en los datos, dando la oportunidad de disminuir la base para terminar sólo con aquellos que si son relevantes a la hora de clasificar.

Los árboles de decisión pueden ser convertidos en una colección de reglas “si ocurre x entonces y”, este conjunto de reglas en muchos casos muestran la información de una manera mucho más comprensible y facilitan la toma de decisiones. Esta representación, es extremadamente útil cuando se estudia como grupos específicos de atributos o conceptos se relacionan con una conclusión en particular. Por ejemplo, la siguiente regla presenta un patrón en la base de datos

que entrega un perfil para un sub grupo de planes de negocio que contienen la clase "High-Medium".

> Si el concepto **Reduce Unemployment** no está presente y el concepto **Many Jobs Opportunities** si, entonces el potencial de empleo del plan de negocio es **'High-Medium'**.

Lo anterior, es una regla humana de lenguaje. Para una máquina, puede expresarse de la siguiente manera:

> IF Reduce_Unemployment=FALSE and Many_Jobs_Opportunities=TRUE THEN
>
> N-potential worker Class= 'High-Medium'

En este sentido, un árbol de decisión puede entenderse como un conjunto de reglas en donde los enunciados lógicos y sus resultados se presentan en una figura con forma de árbol. En la Figura , cada nodo terminal representa el resultado finalde un enunciado lógico o de una regla. Para leer los resultados de este árbol se debe avanzar desde arriba hacia abajo. Por ejemplo, en la Figura un nodo terminal se encuentra encerrado en un círculo azul y representa la siguiente regla.

> *"Si los conceptos Reduce Unemployment y ManyjobOpportunity no están presentes y el concepto local si, entonces el plan tendrá una clase "High-Medium" con un 92.86% de probabilidad".*

En este nodo terminal, el valor Total muestra cuantos casos clasificó la regla (14 planes de negocio, 2,45% de muestra de entrenamiento). Además, una característica interesante de los árboles de decisión es que las reglas están creadas de tal manera que los atributos más importantes para la tarea de clasificación se encuentran más arriba en la estructura. Siguiendo esta lógica, como se muestra en laFigura , la categoría más importante para clasificar los distintos planes fue **Reduce Unemployment**. La segunda categoría más importante fue

Manyjobopportunity, seguida de los conceptos **Local** y **Create Jobs.** La Figura no muestra todas las reglas que encontró el algoritmo. La lista completa de reglas de puede encontrar en el anexo 2.

En relación con la interpretación de los resultados de los planes de negocio, es posible obtener algunas conclusiones preliminares: El hecho de que el concepto **Reduce Unemployment** se presente arriba en el árbol, indica que existen planes de negocio que no mencionan directamente que crearan empleo. Además,el concepto **ManyjobOpportunity** indica que la frase "jobopportunity" se vuelve más relevante al momento de usar un modificador de texto tal como "many" o "lot". Con respecto al concepto **Local**, es posible mencionar que existe un gran cantidad de planes enfocada en ayudar a su comunidad, por lo que este concepto se vuelve relevante. La siguiente categoría es **Create Jobs**, la cual guarda directa relación con el problema de investigación, lo que hace evidente que la máquina tienda a rankearla dentro de las categorías más relevantes a la hora de discriminar entre los distintos planes de negocios.

Existen otros conceptos que el árbol también considera importantes, dentro de los que se cuentan los siguientes:

- **Community**: Existen documentos que se encuentran totalmente enfocados en generar beneficios para su comunidad, es por esto que no es poco común encontrar planes que generen empleos para su propia área o ciudad.
- **Employopportunity**y**farmergroups**: El primero se encuentra directamente relacionado con la creación de empleos mientras que el segundo se relaciona fuertemente con el concepto de comunidad visto anteriormente.
- **Reducingpoverty**: Cuando un Proyecto genero empleo, indirectamente se encuentra reduciendo la pobreza por lo que la mayoría de los aplicantes no pierde la oportunidad de decirlo.
- **Unemploymentyouths** y **Young**: Estos dos conceptos son importantes debido a que la mayoría de las compañías que se encuentran entregando trabajo se enfocan en las personas más jóvenes debido a que estos son más baratos y además pueden hacer trabajos no calificados sin problemas.

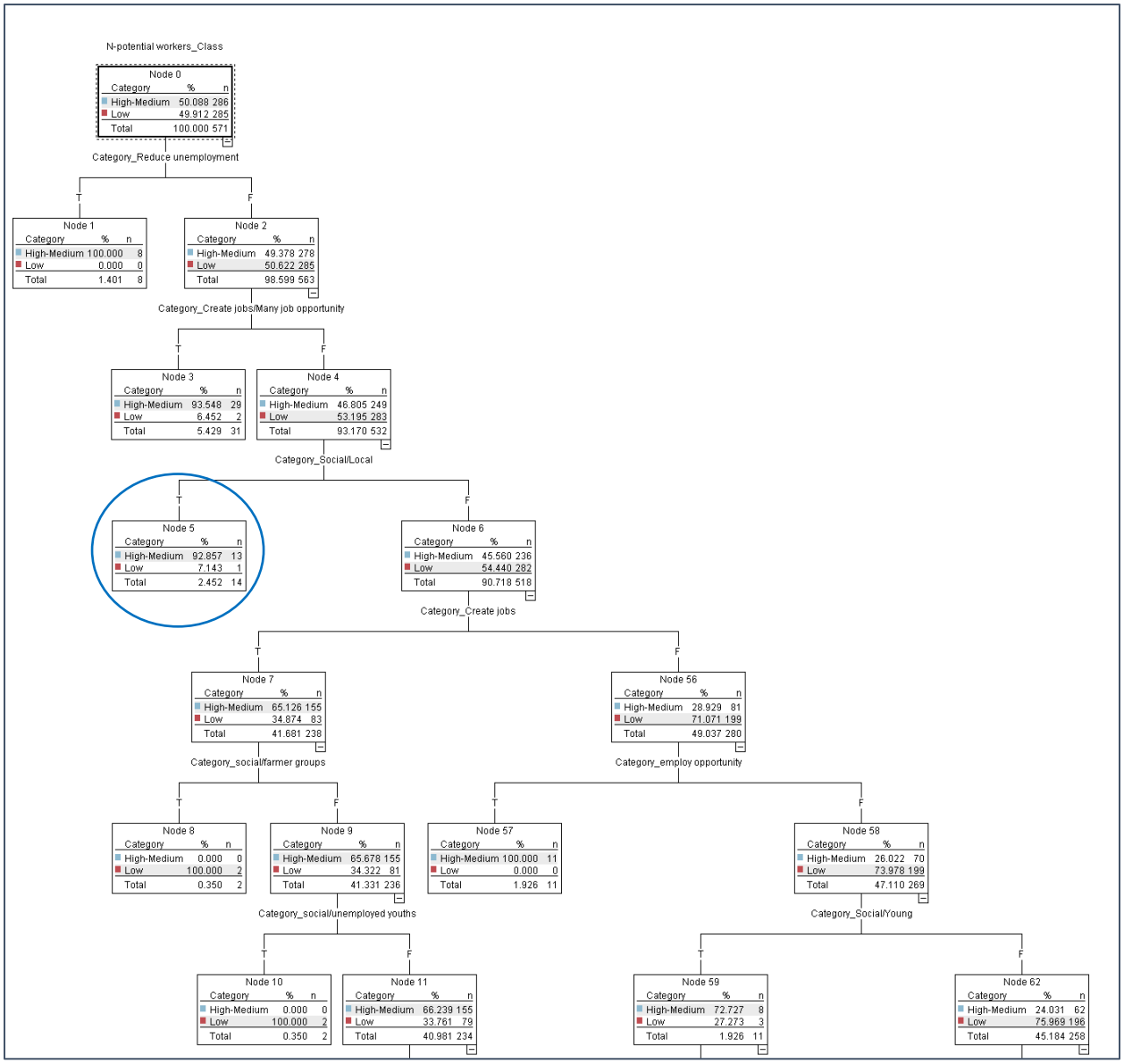


Figura 20

No todos los árboles de decisión son creados iguales

Otra característica importante de los árboles de decisión es que cada predicción se encuentra asociada a un score que representa la confianza de clasificar bien cada observación. Este valor permite al analista rankear las predicciones de acuerdo a la confianza, partiendo con las que tienen una mayor probabilidad de estar en lo correcto y terminando con las que tienen una menor probabilidad.

Extraordinariamente, un modelo con una precisión promedio puede volverse bastante bueno si sólo se consideran las predicciones con alto score. La Figura muestra la cuenta de predicciones en la muestra de prueba ordenadas por el score de la clase “High-Medium” ($-N-Potentialworkers Score). Como se puede observar en la figura, un alto número de predicciones correctas ocurren cuando se tiene una confianza mayor al 80%. De la misma manera, un alto número de predicciones correctas para la clase “Low” ocurren cuando el score para la clase “High-Medium” es menor al 30%.

Una aplicación interesante de esto, es el escaneo rápido de planes que puede hacerse y luego rankear estos mismos por su score. Por ejemplo, el 10% mejor según score puede ser automáticamente clasificado como “High-Medium” y ofrecerles fondos. O por el contrario, los documentos rankeados en el 10% más bajo pueden rechazarse sin otro análisis. En específico, el número de predicciones en el 20% mayor del “High-Medium” score en la muestra de prueba fue 55 y de esos, 34 se encontraban sobre el 90% de confianza (25.37%).

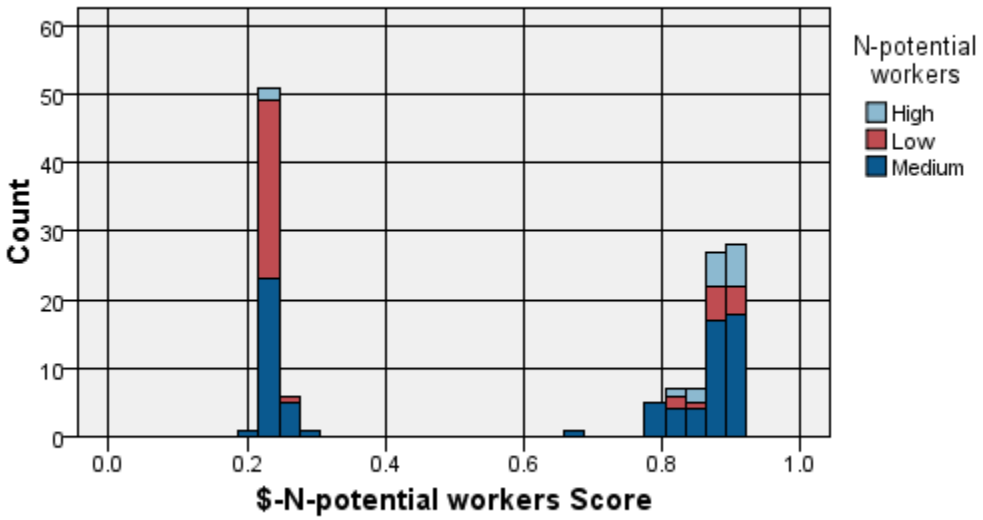


Figura 21

Para estudiar en mayor detalle el rendimiento del modelo en cada decil del score se generó un gráfico Lift. Este grafico compara para cada decil, el número de casos clasificados correctamente

con el número de casos incorrectos. El indicador Lift representa el ratio de los correctamente clasificados sobre los incorrectos en cada decil, además puede ser considerado un tipo especial de indicador de Class Recall. Un valor de 1 significa que el número de casos correctamente clasificados es igual que el de casos incorrectamente clasificados. Además, el grafico Lift muestra las variaciones en este indicador cuando se van añadiendo nuevos casos a la muestra y compara esas variaciones con la variación nula que se obtendría en un modelo aleatorio o naive. Este análisis entrega al analista una guía no sólo para evaluar los modelos, sino también para comprender en detalle la relación entre el score y el rendimiento del modelo.

En la Figura 34

5yFigura 23 se muestran los graficos Lift para las predicciones “Low” y “High-Medium” respectivamente. Es importante tomar en consideración que los deciles se encuentran ordenados desde el mejor al peor. El decil 1 tiene los scores mas altos, mientras que el 10 los mas bajos.

Como se puede observar en las dos figuras, el modelo es 2.25 veces mejor cuando predice la categoría “Low” que un modelo aleatorio hasta el cuarto decil. Por otro lado, es 1.8 veces mejor cuando predice la categoría “High-Medium” hasta el quinto decil.

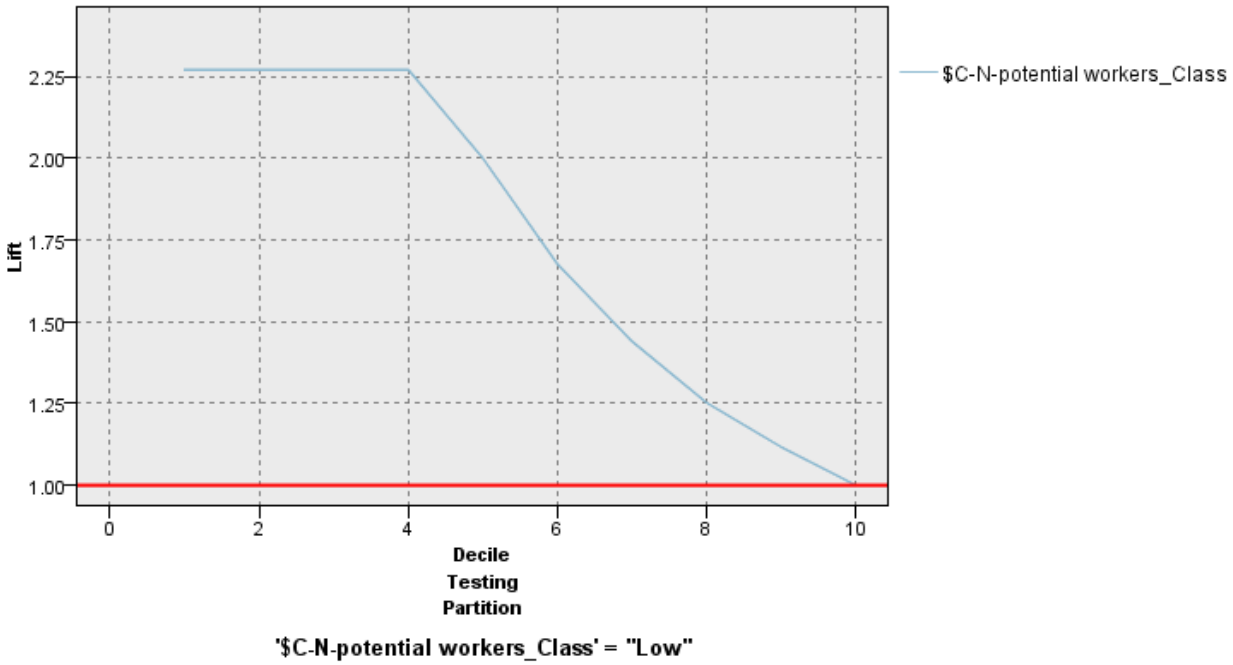


Figura 22

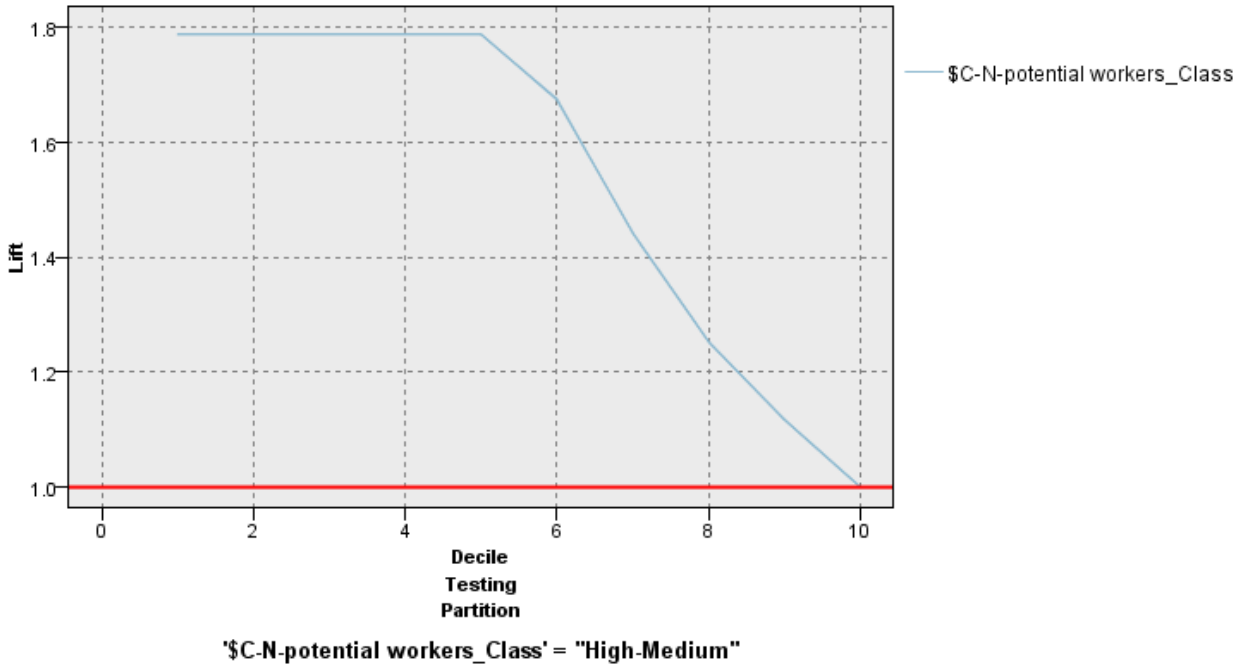


Figura 23

De la misma manera, se podría estudiar no sólo el comportamiento del indicador Lift, sino que además cómo evoluciona el indicador Class Recall desde el mejor decil al peor. La importancia de esto radica en el hecho de que el analista puede comprender hasta que decil el modelo funciona con un 100% de precisión. La Figura 24 y la Figura 25 muestran la precisión de cada predicción en cada decil. En el grafico asociado a la clase “High-Medium” se observa que el Class Recall en la muestra de prueba es de un 100% hasta el sexto decil y los scores en ese rango se encuentran sobre 80%. Esto muestra que la predicción de la clase “High-Medium” se encuentra cercana a la perfección cuando el modelo se encuentra por lo menos 80% seguro.

El score de la predicción de la clase “Low” se calculó como 1-“High-Medium”. Esto explica por qué en la Figura se muestra una relación inversa entre la precisión y el score. Una conclusión que se puede extraer de este gráfico es que cuando el score de la predicción de la clase “High-Medium” es menor a 30%, el Class Recall del modelo es 100% cuando predice la clase “Low”.

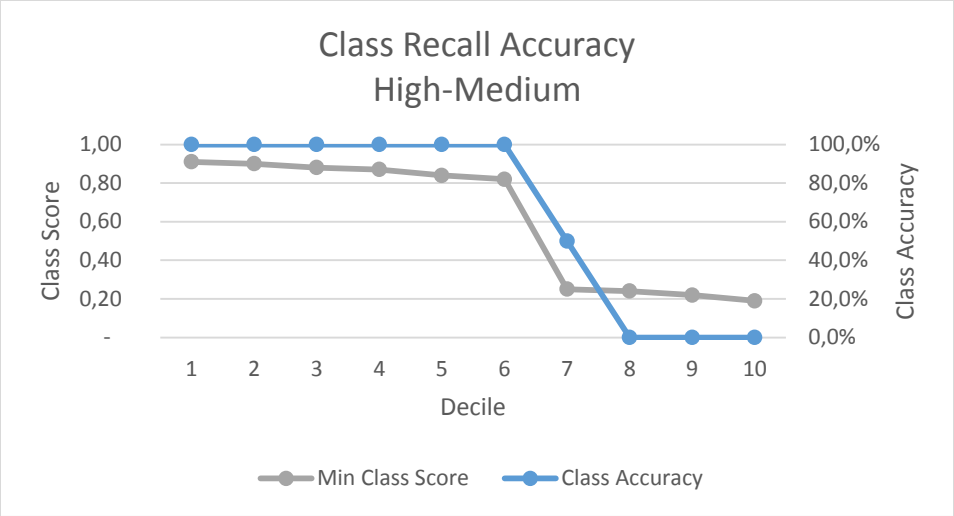


Figura24

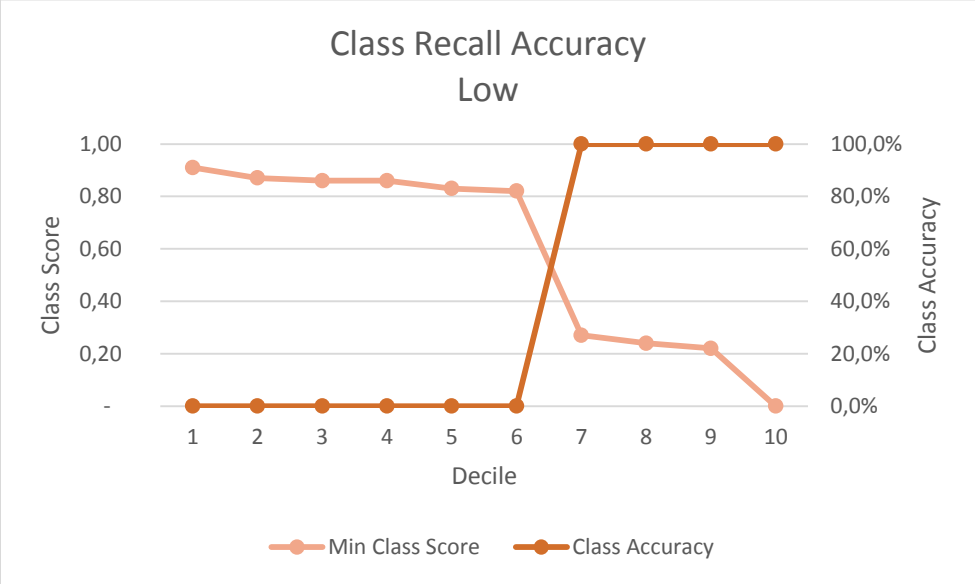


Figura 25

Finalmente, un interesante análisis a realizar es testear la correlación existente entre los valores de la clasificación manual de generación de empleo y los predichos. Como se puede observar en la Figura 1, la correlación entre estas dos variables en la muestra completa es positiva y estadísticamente significativa (r=0.561 p-value 0.000). A su vez el indicador R-cuadrado indica

que el modelo realiza como se esperaba la tarea de clasificación para la que fue creado. Esta correlación también es positiva y estadísticamente significativa (r=0.35 p-value 0.000).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.561(a)	.315	.314	.524

a. Predictors: (Constant), $-N-potential workers Score

ANOVA(a)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	76.876	1	76.876	279.708	.000(b)
	Residual	167.380	609	.275		
	Total	244.255	610			

a. Dependent Variable: N-potential workers score

b. Predictors: (Constant), $-N-potential workers Score

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.110	.042		26.212	.000
	$-N-potential workers Score	1.112	.067	.561	16.724	.000

a. Dependent Variable: N-potential workers score

Figura 1

Fase 2: Desarrollo y análisis automático de documentos

- **Sub-Fase 2.1: Aplicación del modelo refinado de Text Mining en la muestra total de documentos.**

Un tema importante es testear el modelo en un concurso que no sea en el que fue creado. Esto podría ayudar el investigador a conocer la escalabilidad del modelo en otros concursos. Existen diferentes factores que hacen que la tarea de clasificar planes de un concurso nuevo sea una tarea bastante desafiante:

- Las palabras o los idiomas que los aplicantes usan para expresar generación de empleo pueden variar dado el lenguaje local, la cultura, la audiencia a la que documento va dirigido y el contexto en el que el plan fue elaborado.

- Otro problema relacionado es el lenguaje como tal: si el modelo fue generado usando una muestra de planes escritos en inglés, es imposible aplicarlo automáticamente a una muestra de planes escritos en español.
- Normalmente cada concurso tiene su propio formato (la manera en que se hacen las preguntas por ejemplo; entrevistas estructuradas, formularios, distintos tipos de archivos, etc.). Esto puede confundir al modelo cuando busque los conceptos. Una manera de arreglarlo es pre procesando cada concurso, dejando solamente las respuestas en un archivo de texto.

Tomando estas consideraciones, el modelo que fue creado con los planes de negocio del concurso InfoDev fue aplicado a otro concurso llamado YouWin. Este concurso fue elegido principalmente por que los documentos se encontraban en un formato simple para trabajar (hojas de Microsoft Excel) y porque contaba con un considerable número de documentos que podían usarse para probar el modelo.

Además, cada documento de este concurso incluía un indicador llamado "markjobcreation", el cual entregaba una calificación de la potencial generación de empleo de un plan de negocio, realizada por jueces independientes.

Este indicador contenía valores que fluctuaban entre 1 y 25 (véase Figura 27), siendo 25 el valor más alto que podía entregar un juez al momento de evaluar la potencial generación de empleo. Como se dijo anteriormente, el modelo fue creado para discriminar entre 2 clases; "High-Medium" o "Low", es por esto que el indicador "markjobcreation" del concurso YouWin se transformó en una nueva variable con dos categorías. Todos los planes evaluados con un valor bajo 10.5 fueron clasificados como "Low" mientras que los otros fueron clasificados como "High-Medium". Este valor de 10.5 fue calculado como la media, dado el valor de los jueces, menos 1 desviación estándar. Siguiendo esa regla, el concurso quedo con 6045 planes de negocio en la clase "High-Medium" y 775 en la clase "Low".

- **Sub-Fase 2.2: Análisis estadístico de los contenidos y conceptos extraídos.**

La Tabla5 muestra los resultados al aplicar el modelo construido en la muestra de InfoDev a la muestra total de YouWin (6820 planes de negocio). Como se puede observar, el modelo

nuevamente tiene un mejor rendimiento en la clase “High-Medium” (55.1% Class Recall, 90% Class Precision). Comparando esto con el rendimiento en InfoDev, se observan menores valores en la precisión, lo que se explica dadas las razones expuestas anteriormente.

Se realizó un análisis de correlación con el objetivo de estudiar la relación entre el score del modelo y la clase predicha. Este análisis mostro que la correlación fue positiva y estadísticamente significativa, pero fue más bien pequeña (r=0.05 p-value 0.000).

	‘HIGH-MEDIUM’	**‘LOW’**	**RECALL**
TRUE ‘HIGH-MEDIUM’	3,333	2,712	55.1%
TRUE ‘LOW’	369	406	52.4%
CLASS PRECISION	90%	13%	

Tabla5

Pese a que estos resultados pueden no verse del todo buenos, es importante analizar el rendimiento de las predicciones que fueron realizadas con un alto valor de confianza. Para poder apreciar la verdadera usabilidad del modelo, el investigador estudio el rendimiento del modelo rankeando cada predicción con respecto al score en la predicción de la clase “High-Medium”. Se generó un gráfico Lift para las dos clases. La Figura 28 muestra como la precisión del modelo mejora cuando aumentan los deciles alcanzando su valor máximo en los deciles 3 y 4. De la misma manera, la Figura 29 muestra mejores resultados para la predicción de la clase “Low”, donde el modelo tiene un rendimiento 1.3 veces superior hasta el tercer decil.

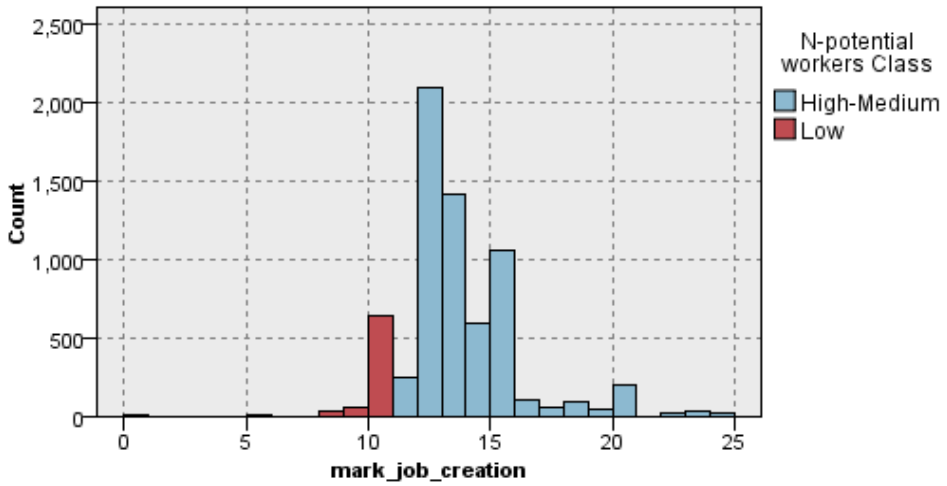


Figura27

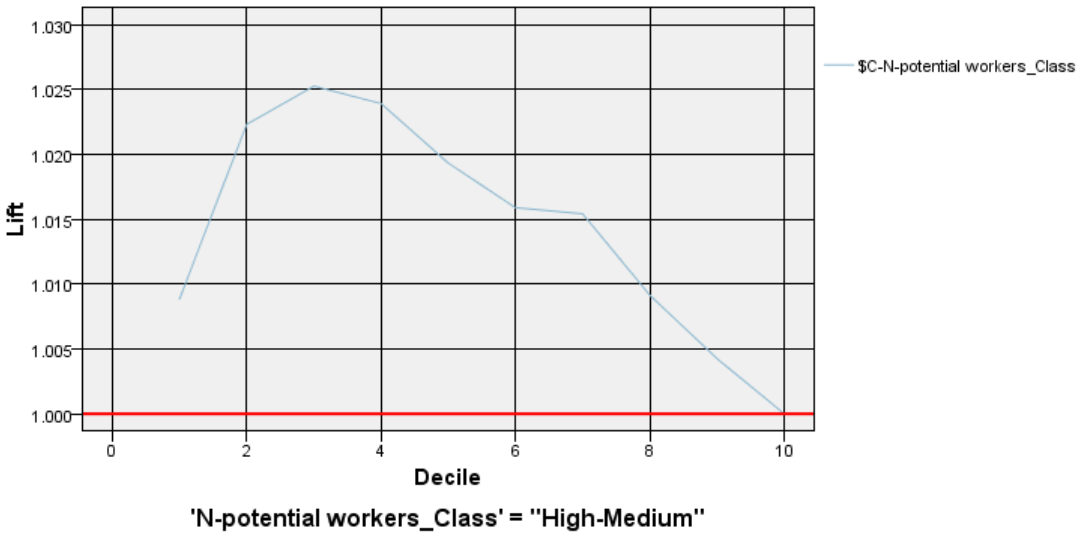


Figura2

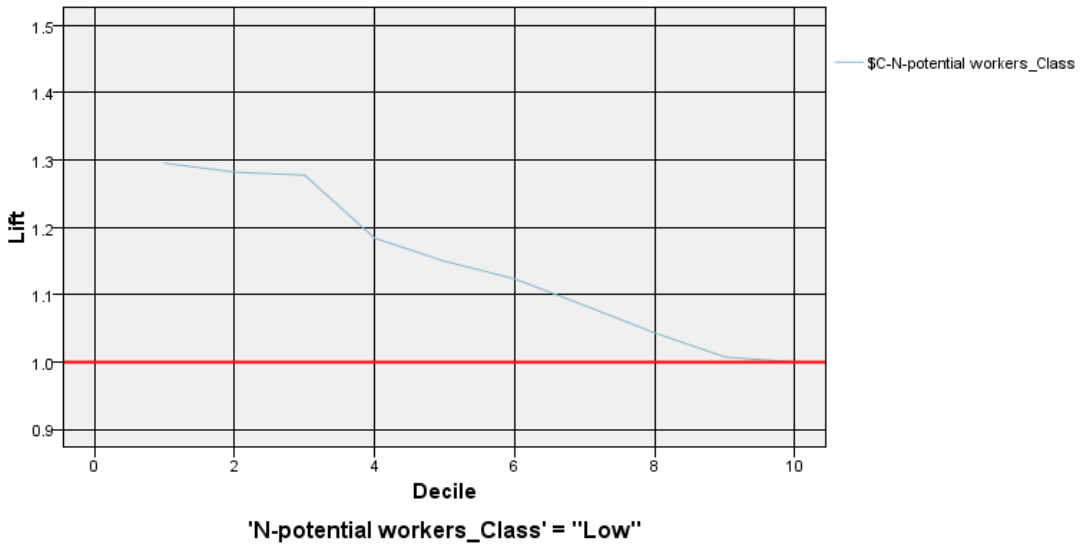


Figura 293

Como se puede observar en la Figura 30 y en la Figura 31, la predicción de la clase “High-Medium” tiene un 100% de concordancia en los primeros cinco deciles. Otra interesante conclusión es que cuando el modelo tiene un score mayor a 80%, la concordancia con los jueces es de 100%. Esto implica que cada vez que un documento es clasificado como “High-Medium” con un score mayor a 80%, el indicador “markjobcreation” es mayor a 10.5. De la misma manera, en el grafico que muestra la predicción de la clase “Low” es posible apreciar que cuando el score es menor a 30% la concordancia con los jueces es de un 100% también.

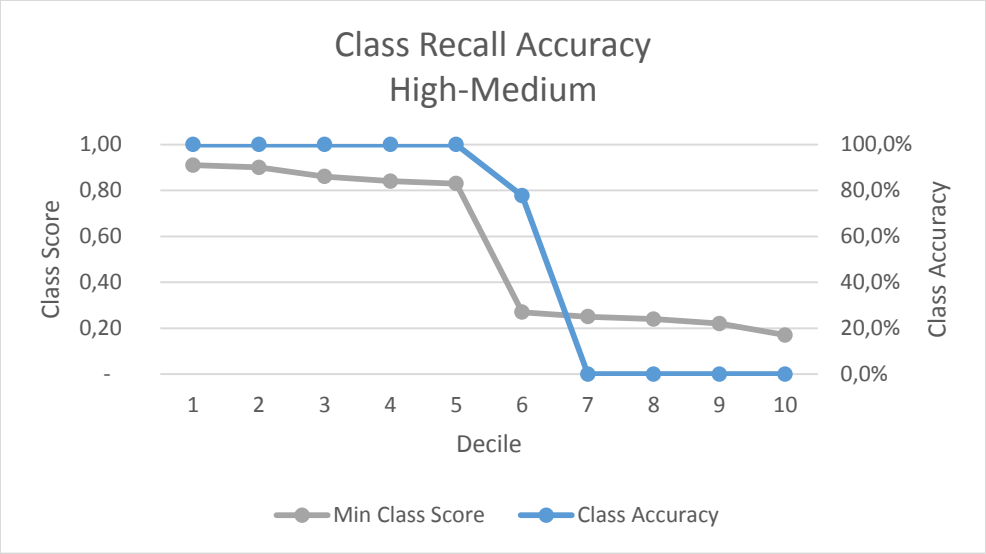


Figura30

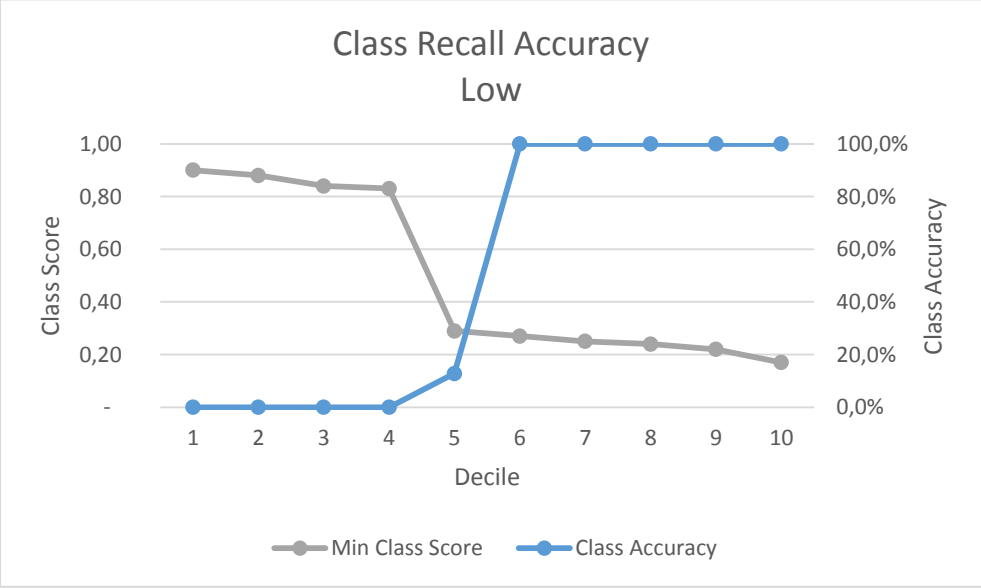


Figura 314

Del total de 6045 planes que tenían la clase “High-Medium”, el modelo fue capaz de predecir 3051 con una precisión perfecta (los primeros cinco deciles), y de los 775 planes que tenían las clase “Low”, el modelo predijo 400 con esta condición. Esta información es bastante importante y permite al analista obtener un mayor entendimiento de las predicciones y su relación con los

mismos planes. Además, entrega al analista distintas herramientas para evaluar futuros modelos y comprenderlos.

Trabajando sólo con la muestra en donde el modelo tuvo un 100% de concordancia, un test de medias fue desarrollado para observar si las medias de cada clase eran estadísticamente diferentes. Como se muestra en la Figura 32, la media del indicador "markjobcreation" de los documentos con la clase "High-Medium" fue 13,69 y la de los documentos con la clase "Low" fue de 9,42. La diferencia entre estos dos valores es estadísticamente significativa (t-test 46.951, p-value 0.000%). Además, la desviación estándar entre grupos es estadísticamente significativa(Levene's Test F=88.184 p-value 0.00%), además se puede decir que el grupo de documentos con la clase "High-Medium" tiene mayor dispersión que el grupo con la clase "Low".

Group Statistics

	Flag_H-M_prediction	N	Mean	Std. Deviation	Std. Error Mean
mark_job_creation	High-Medium	3051	13,69	2,314	,042
	Low	399	9,42	1,608	,081
$-N-potential workers Score	High-Medium	3051	,87167871	,031501611	,000570311
	Low	399	,22777921	,013826041	,000692168

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
mark_job_creation	Equal variances assumed	88,184	,000	35,667	3448	,000	4,261	,119	4,027	4,496
	Equal variances not assumed			46,951	636,699	,000	4,261	,091	4,083	4,440
$-N-potential workers Score	Equal variances assumed	1105,510	,000	403,205	3448	,000	,643899499	,001596953	,640768430	,647030567
	Equal variances not assumed			717,952	1058,191	,000	,643899499	,000896856	,642139681	,645659317

Figura 32

Finalmente, dos gráficos de vela fueron generados para visualizar de mejor manera la relación entre los valores predichos y los reales. Como se puede observar en la Figura 33, la diferencia entre la clase "High-Medium" predicha y la clase "Low" predicha es bastante clara.

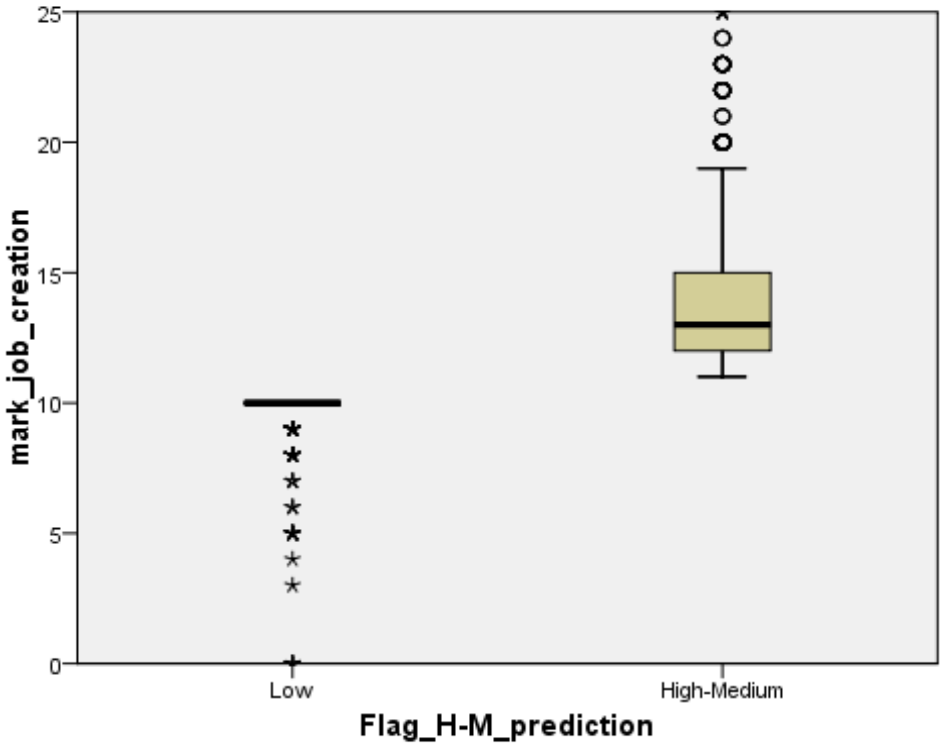


Figura 33

En la Figura 34 se observa que cuando el score se encuentra sobre 80%, la clase más predicha es “High-Medium” mientras que cuando es menor a 25% la clase “Low” aparece con mayor frecuencia. Esta diferencia es significativa (véase Figura 32).

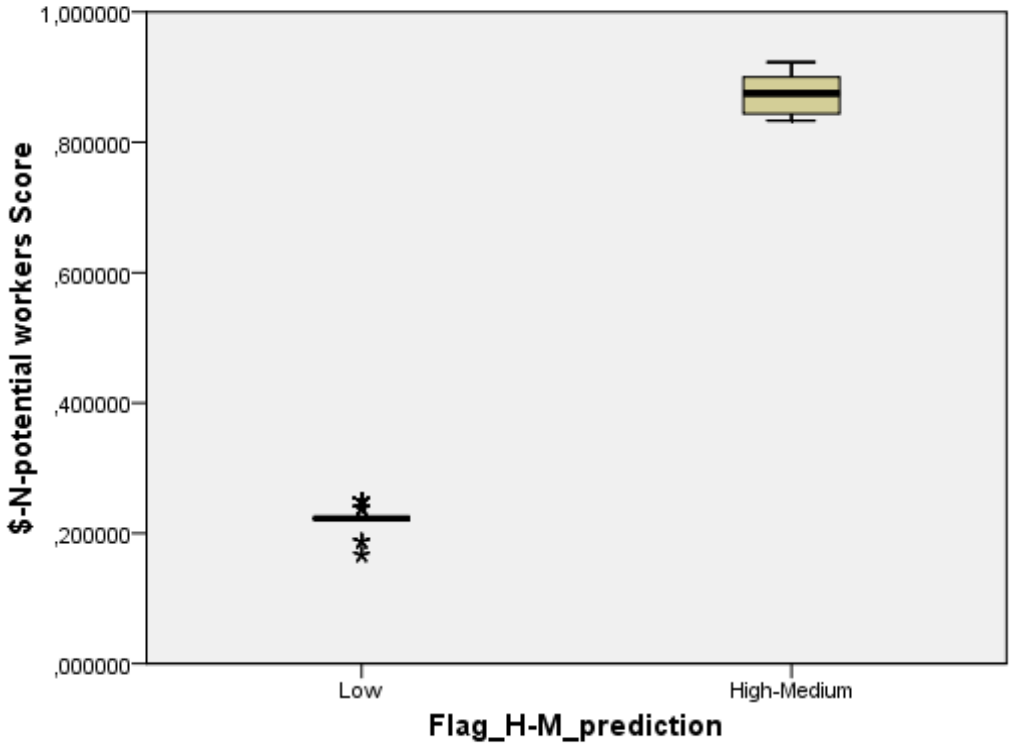


Figura 34

5

4.2 Respuestas a preguntas de investigación

Dadas las preguntas presentadas en el capítulo 2 de esta tesis, se procede a contestarlas basado en los resultados obtenidos.

1) Es posible generar un modelo que en base al análisis de un plan de negocio prediga el potencial de empleo con una precisión superior a un modelo aleatorio.

H1: No es posible generar un modelo que realice la tarea

Los resultados mostrados anteriormente nos entregan evidencia suficiente para rechazar esta hipótesis, debido a que si es posible crear un modelo que en base al análisis de un plan de

negocio prediga el potencial de empleo con un nivel de predicción bastante superior a un modelo aleatorio.

H2: Es posible generar un modelo que realice la tarea con un 100% de precisión

Dada la evidencia observada es posible rechazar esta tesis debido a que no se obtuvo un 100% de precisión en la muestra, sólo en algunas zonas.

H3: Es posible generar un modelo que realice la tarea con una precisión superior a la obtenida por un modelo aleatorio pero menor al 100%.

Es posible rechazar esta tesis por lo que se asume como verdadera. Los resultados muestran que el rendimiento del modelo es notoriamente superior a uno aleatorio.

H4: Es posible generar un modelo que realice la tarea con un mejor rendimiento que un modelo no lingüístico.

Es posible rechazar esta hipótesis por lo que se asume como verdadera. Se observa que un modelo no lingüístico si obtuvo resultados similares en la clase High_Medium, pero notablemente menores en la clase Low, por lo que eso nos entrega la evidencia necesaria.

La pregunta e hipótesis planteadas muestran que crear un modelo que prediga el potencial de empleo que pueden generar distintos planes de negocio es posible con una precisión razonable. Además destaca la hipótesis número 4 en donde se le compara con un modelo no lingüístico. La importancia de esto último radica en que en ese caso la máquina elige los términos que más discriminan entre los planes por lo que con esto se demuestra que aplicar sentido humano al análisis entrega mejores resultados. Esto genera un precedente para la discusión existente en la realidad acerca de si en un futuro cercano las máquinas serán capaces de funcionar perfectamente por su propia cuenta. En este caso vemos que la respuesta es negativa.

2) Es posible encontrar ciertos factores claves que explican el nivel de empleo generado

H1: el factor más importante es la industria del plan de negocios

Se encontró evidencia suficiente para rechazar esta hipótesis. Además, los resultados mostraron que el concepto industria no sólo, no es el más importante, sino que tampoco aparece dentro de

los primeros. Esto entrega información relacionada con que existen otros factores que tienen un mayor peso a la hora de discriminar entre los distintos planes de negocio.

H2: el factor más importante es aportar a la reducción del desempleo.

No se encontró evidencia suficiente para rechazar esta hipótesis, por lo que se asume verdadera. La evidencia muestra que aportar a la reducción del desempleo es uno de los principales factores, lo que plantea la discusión acerca de si los emprendedores sólo se basan en obtener beneficios o también en ayudar gente por medio del empleo. Otros conceptos que no son los más importantes, pero que aparecen dentro de los que más discriminan son el de comunidad y reducción de la pobreza, lo que también se relaciona a la idea expresada anteriormente.

H3: el factor más importante es entregar ayudar a la comunidad.

Se encontró evidencia suficiente para rechazar esta hipótesis, sin embargo este concepto aparece dentro de los primeros por lo que es posible decir que cuenta con un grado de importancia a la hora de discriminar.

Esta pregunta se centra en las razones por las cuales los emprendedores pudieran generar empleo por medio de sus planes de negocios. Pese a la creencia popular de que sólo se centran en obtener beneficios y ven el empleo como un medio para ello, se observa que el factor que más relevancia ostenta es el de reducción del desempleo y que el factor de comunidad también cuenta con un grado de importancia. Esto no significa que se dejen de lado los beneficios, pero sin duda amplia el marco de investigación relacionado con el tema. Una gran cantidad de aplicaciones podrían desarrollarse fruto de esto, debido a que plantea un cambio en la mentalidad de los emprendedores y por consiguiente la literatura debería adaptarse a este cambio. Un análisis que podría desarrollarse a partir de esto podría ser el de analizar según zonas geográficas el grado de “apoyo a la comunidad” de los planes pertenecientes a esa zona, lo que permitiría a una entidad global, como el banco mundial por ejemplo, premiar o castigar en función de eso.

3) Es posible predecir el potencial de empleo con una seguridad del 100%

H1: Es posible predecir el potencial de empleo de la totalidad de los planes de negocio con una seguridad del 100%.

Los resultados nos obligan a rechazar esta hipótesis debido a que no es posible para el modelo predecir con un 100% de seguridad, pese a que en cierto número de planes si es posible.

H2: Es posible predecir el potencial de empleo de al menos un 20% de planes de negocio, con un seguridad del 100%.

Dada la evidencia, no es posible rechazar esta hipótesis por lo que se asume verdadera.

Esta pregunta es de vital importancia debido a que, como se mencionó anteriormente, cada predicción normalmente tiene asociado un costo. Es por esto que el poder desarrollar predicciones con una seguridad total es extremadamente valorado. Además, contar con un 20% de la base en la que se puedan hacer predicciones con 100% de seguridad es muy importante ya que permite tomar decisiones de manera rápida y sin perder exactitud, lo que hace a cualquier tipo de aplicación más eficiente.

4) Es posible aplicar el modelo a un concurso distinto del que se usó para entrenarlo.

H1: Es posible aplicar el modelo a otro concurso y obtener una precisión mayor a la obtenida con un modelo aleatorio.

No existe evidencia para rechazar esta hipótesis por lo que se asume verdadera. Como se mostró en los resultados, el modelo se aplicó a otro concurso y si bien decayó en sus predicciones, obtuvo un rendimiento notoriamente mejor que el de un modelo aleatorio.

H2: Es posible aplicar el modelo a otro concurso y obtener al menos un 20% de predicciones con un 100% de seguridad.

No existe evidencia para rechazar esta hipótesis por lo que se asume verdadera. Los resultados expuestos sugieren que al aplicar el modelo a otro concurso se obtuvo un número de predicciones con 100% de seguridad mayor al 20%.

Esta pregunta apunta a medir el real grado de aplicabilidad que tiene el modelo debido a que uno de los objetivos principales es que sea usado en una gran cantidad de concursos y no sólo en el concurso para el cual fue creado. Es por esto que medir la precisión y la seguridad de las predicciones se vuelve imprescindible a la hora de evaluar su real grado de potencial uso.

5. Conclusiones e investigación futura

El presente estudio se enfoca en la difícil y poco abordada tarea de predecir el potencial de empleo que un plan de negocio puede entregar. En el pasado no se ha intentado hacer esto de manera completa ni formal, solamente se han testeado relaciones entre diversos factores y la creación de empleo, pero todos en retrospectiva. En otras palabras, lo que se analiza es el empleo creado o el que se está creando en este momento, nunca el que se creará. Esto genera que muchas empresas nuevas que podrían haber sido capaces de generar grandes cantidades de empleo, quiebren o se desintegren debido a que no recibieron apoyo a tiempo.

Para esta tarea, decidimos apoyarnos en el uso del KDD y de algoritmos de DM. Para ser más específicos usamos tanto algoritmos de Text Mining como árboles de decisión. Esto último nos permitió generar nuevos insights debido a que esta tecnología es bastante innovadora y permite realizar investigaciones que de otra manera habrían sido bastante más complejas.

Dada la revisión bibliográfica realizada, se descubrió que la literatura correspondiente a la predicción del potencial de la creación de empleo es bastante escasa. La literatura existente se basa más en análisis del empleo que ya se generó o que está siendo generado en este momento, no en el que se generará. Esto muestra diversas desventajas, debido a que existen empresas que pudieron haber generado grandes cantidades de empleos, pero que dado que no recibieron apoyo a tiempo, terminaron cerrando y por consiguiente sin entregar esos empleos. Si se observa ahora la literatura correspondiente a técnicas de Data y Text Mining junto con sus aplicaciones, se observa que existen varias pero no en temas relacionados con la creación de empleo.

Los resultados presentados en esta tesis muestran que el uso de un modelo de Text Mining para analizar y predecir el potencial de creación de empleo de un plan de negocios de manera automática, no es sólo posible, sino que también escalable a otros concursos con niveles razonables de predicción. Futuros análisis pueden incluso mejorar los modelos *linguisticos*de

Text Mining presentados con el objetivo de considerar un mayor espectro de tipos de planes de negocios, lo que implica tener planes de distintos lenguajes, entornos culturales, ciclos de vida, entre otros. Con lo cual se obtendría no sólo un modelo más completo, si no que con una capacidad de extrapolación notablemente mayor.

Un hecho importante expuesto aquí, es que los modelos no sólo deben ser juzgados por su precisión promedio, esto debido a que si estudiamos en detalle la confianza que se le asignó a cada predicción se pueden obtener reglas que funcionan de manera bastante acertada. Usarsólo predicciones que tengan un alto grado de confianza puede ayudar de sobremanera al momento de discriminar entre diferentes planes de negocio que se encuentran en las colas de una distribución de creación de empleo. Este análisis puede incluso mostrar donde el modelo funciona mejor y donde puede ser mejorado, otorgando al analista una medida efectiva de la precisión de esta técnica. Esto último se puede aprovechar de sobremanera, debido a que si bien hoy en día el modelo no es capaz de predecir de manera perfecta, si tiene un gran rendimiento en el primer tercio de la muestra. Esto permitiría premiar a los planes que entregarán empleo o desechar a los que no de manera automática, rápida y eficiente.

Como limitación de nuestro estudio se encuentra la gran brecha aún existente entre los niveles de precisión obtenidos y el 100% que sería el ideal. Esto puede deberse a varias razones dentro de las cuales sobresalen la dificultad para trabajar con palabras que implican un sentido humano, el tiempo destinado a la investigación, el cual era limitado, y finalmente el hecho de que la clasificación inicial fuera realizada por analistas sin un conocimiento acabado en el tema, lo que podría plantear el caso de que el modelo funcionara mejor que el propio analista. Una segunda limitación se relaciona con la baja capacidad de extrapolación del modelo, la razón para este punto es bastante simple y se basa principalmente en que el modelo se entrenó con los términos de un sólo concurso por lo que a medida que se vayan agregando concursos a su base, su capacidad de extrapolación ira mejorando paulatinamente.

Desde este último punto se desprende la investigación futura o los pasos a seguir para poder mejorar los niveles de predicción expuestos por el modelo desarrollado. Esta se basa en desarrollar la anotación manual y las reglas de extracción para otros concursos, logrando de esta manera aumentar la base de conceptos del modelo permitiendo que este funcione de mejor manera. Mientras mayor sea la variedad de conceptos que el modelo reconozca, mayor será su

capacidad de extrapolación a concursos nuevos y mayor será su precisión en los concursos existentes.

6. Referencias

- [1] Díaz, D. (2013). Material de Apoyo Curso Business Intelligence and Analytics.
- [2] Cantillion, R. (1730). Essai sur la Nature du Commerce en Général.
- [3] Schumpeter, J. (1942). Capitalism, Socialism, and Democracy. New York: Harper Brothers.
- [4] Ayyagari, M., Demirguc-Kunt, A., &Maksimovic, V. (2011). Small vs Young Firms across the World, Contribution to Employment, Job Creation and Growth.
- [5] Haltiwanger, J., Jarmin, R. & Miranda, J. (2009). Jobs Created from Business Startups in the United States. Kauffman Foundation WorkingPaper. http://www.kauffman.org/.
- [6] Klapper, L. & Love, I. (2010). New Firm Creation.
- [7] Ghani, E., Kerr, W. & O´Connell, S. (2011). Who creates Jobs.
- [8] Osterwalder, A. (2004). The Business Model Ontology, a Preposition in a Design Science Approach.
- [9] Chu, Y. & Hsu, W. (2006). Organization Ontology for Innovation and Entrepreneurship for Cross-Border Knowledge Services in the Globalizing IC Design Industries.
- [10] Sousa, P., Manso, V., Costa, J. &Almeyda, F. (2012). Ontology for Entrepreneurs-Risk Analysis for Start-Up Tech.
- [11] Lytras, M., &Gracia, R. (2008). Semantic Web Applications: A Framework for Industry and Business Exploitation – What is it needed for a successful semantic web based application. International Journal of Knowledge and Learning 4(1): 93-108.
- [12] Timmers, P. (1998). Business Models for Electronic Markets. JournalonElectronicMarkets 8(2): 3-8.
- [13] Weill, P. & Vitale, M. (2001). Place to space: Migrating to eBusiness Models. Boston, Harvard Business School Press.
- [14] Osterwalder, A., Pigneur, Y. &Tucci, C. (2005). Clarifying Business Models: Origins, Present, and Future of the Concept.
- [15] Linder, J. & Cantrell, S. (2000). Changing Business Models: Surveying the Landscape, accenture Institute for Strategic Change.

- [16] Markides, C., &Oyon, D. (2010). What to Do Against Disruptive Business Models (When and How to Play Two Games at Once). MIT Sloan Management Review 51(4): 26-32.
- [17] Saab, D.J., Fonseca, F., (2008). Ontological complexity and human culture. Conference paper presented at the Philosophy's Relevance in Information Science, Paderborn.
- [18] Morecroft, J.D. (1994) Executive Knowledge, Models, and Learning. In Morecroft, J.D. ; and Sternman, J.D. (editors) Modeling for Learning Organizations, pp. 3-28, Portland : Productivity Press.
- [19] Osterwalder, A. &Pigneur, Y. (2002). An e-Business Model Ontology for Modeling e-Business.
- [20] Kaplan, R. S. & D. P. Norton (1992). The balanced scorecard--measures that drive performance. Harvard Business Review 70.
- [21] Markides, C. (1999). All the Right Moves. Boston, Harvard Business School Press.
- [22] Jones, M., Coviello, N. &Tang, Y. (2011). International Entrepreneurship research (1989-2009): A domain ontology and thematic analysis. Journal of Business Venturing 26.
- [23] Dunham, M. (2002). Data Mining: Introductory and Advanced Topics.
- [24] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases.
- [25] Venkatadri, M., & Reddy, L. (2011). A Review on Data Mining from Past to the Future.
- [26] Rud, O. (2009). Business Intelligence Success Factors, Tools for Aligning Your Business in the Global Economy.
- [27] Tan, A. (1999). Text Mining: The State of the Art and the Challenges.
- [28] Feldman, R. & Sanger, J. (2004). The Text Mining Handbook. Cambridge University Press.
- [29] Hearst, M. (2003). What is text mining? Retrieved from http://www.sims.berkeley.edu/~hearst/text-mining.html
- [30] Hendry, S. &Madeley, A. (2010). Text Mining and the Information Content of Bank of Canada Communications.

- [31] Li, X., Xie, H., Chen, L., Wang, J. & Deng, X. (2014). News Impact on Stock Price Return via Sentiment Analysis.
- [32] IBM. (2010). CRISP-DM 1.0 Step by Step Data Mining Guide.
- [33] Ayre, L. (2006). Data Mining: What It Is and How It Is Related to Information Retrieval and Text Mining.
- [34] Fayyad, U., Piatetsky-Shapiro, G. &Smyth,P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework.
- [35] Stevenson, H., Roberts, M. &Grousbeck, H. (1999). New Business Ventures and the Entrepreneur. 5th ed., Boston: Irwin, McGraw-Hill
- [36] Berlanga, V., Rubio, M.J., Vilá, R. (2013). Como aplicar árboles de decisión en SPSS. Universitat de Barcelona. Institut de Ciéncies de l´educació.

7. Anexos

Anexo 1

Rules for High-Medium, 19 rules	
Rule 1 for High-Medium (8; 1.0)	if Category_Reduce unemployment = T
	then High-Medium
Rule 2 for High-Medium (31; 0.935)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = T
	then High-Medium
Rule 3 for High-Medium (14; 0.929)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = T
	then High-Medium
Rule 4 for High-Medium (14; 0.929)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = T
	then High-Medium
Rule 5 for High-Medium (35; 0.886)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = T
	then High-Medium
Rule 6 for High-Medium (4; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = T
	then High-Medium
Rule 7 for High-Medium (4; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = T
	then High-Medium
Rule 8 for High-Medium (30; 0.867)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = T
	then High-Medium

Rule 9 for High-Medium (8; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = T
	and Category_human resources/developers = T
	and Category_Need of people = F
	then High-Medium
Rule 10 for High-Medium (10; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = T
	and Category_human resources/developers = F
	and Category_social/expansion = T
	then High-Medium
Rule 11 for High-Medium (18; 0.833)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = T
	and Category_human resources/developers = F
	and Category_social/expansion = F
	and Category_Social/Young = F
	and Category_human resources/professionals/members = F
	then High-Medium

Rule 12 for High-Medium (13; 0.923)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = F
	and Category_Need of people/Create job = F
	then High-Medium
Rule 13 for High-Medium (8; 0.875)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = F
	then High-Medium
Rule 14 for High-Medium (5; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = F
	and Category_human resources/professionals/members = T
	then High-Medium

Rule 15 for High-Medium (10; 0.7)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = F
	and Category_human resources/professionals/members = F
	and Category_Employ people/People = F
	and Category_social/community = F
	then High-Medium
Rule 16 for High-Medium (11; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = T
	then High-Medium
Rule 17 for High-Medium (9; 0.889)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = T
	and Category_social/community = T
	then High-Medium
Rule 18 for High-Medium (4; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = F
	and Category_Employ people/employ = T
	and Category_human resources/developers = F
	and Category_Expansion = T
	then High-Medium
Rule 19 for High-Medium (4; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = F
	and Category_Employ people/employ = T
	and Category_human resources/developers = F
	and Category_Expansion = F
	and Category_Job opportunity = F
	then High-Medium

Rules for Low - contains 17 rule(s)	
Rule 1 for Low (2; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = T
	then Low
Rule 2 for Low (2; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = T
	then Low
Rule 3 for Low (4; 0.75)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = T
	then Low
Rule 4 for Low (2; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = T
	then Low
Rule 5 for Low (2; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = T
	then Low
Rule 6 for Low (7; 0.857)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = T
	and Category_human resources/developers = T
	and Category_Need of people = T
	then Low

Rule 7 for Low (5; 0.8)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = T
	and Category_human resources/developers = F
	and Category_social/expansion = F
	and Category_Social/Young = T
	then Low
Rule 8 for Low (5; 0.8)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = T
	and Category_human resources/developers = F
	and Category_social/expansion = F
	and Category_Social/Young = F
	and Category_human resources/professionals/members = T
	then Low
Rule 9 for Low (14; 0.857)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = T
	and Category_social/development = F
	then Low

Rule 10 for Low (2; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = T
	and Category_Industry = T
	and _Category_Employ people = F
	and Category_Need of people/Create job = T
	then Low
Rule 11 for Low (19; 0.789)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = F
	and Category_human resources/professionals/members = F
	and Category_Employ people/People = T
	then Low
Rule 12 for Low (9; 0.778)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = T
	and Category_Social/Agriculture = F
	and Category_social/poverty = F
	and Category_Employ people/employment opportunity = F
	and Category_Expansion/Double team = F
	and Category_social/poverty reduction = F
	and Category_Social/Woman = F
	and Category_social/women = F
	and Category_Employ people/employ = F
	and Category_Create job/Provide job = F
	and Category_human resources/professionals/members = F
	and Category_Employ people/People = F
	and Category_social/community = T
	then Low

Rule 13 for Low (6; 0.833)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = T
	and Category_social/farmer groups = F
	and Category_social/unemployed youths = F
	and Category_Create job = F
	then Low
Rule 14 for Low (2; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = T
	and Category_social/community = F
	then Low
Rule 15 for Low (4; 1.0)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = F
	and Category_Employ people/employ = T
	and Category_human resources/developers = T
	then Low
Rule 16 for Low (10; 0.8)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = F
	and Category_Employ people/employ = T
	and Category_human resources/developers = F
	and Category_Expansion = F
	and Category_Job opportunity = T
	then Low
Rule 17 for Low (236; 0.78)	if Category_Reduce unemployment = F
	and Category_Create jobs/Many job opportunity = F
	and Category_Social/Local = F
	and Category_Create jobs = F
	and Category_employ opportunity = F
	and Category_Social/Young = F
	and Category_Employ people/employ = F
	then Low

Anexo 2

Model 1		
Categories	Subcategories	Rules and concepts
Employ people		
	_employ & more & people	
	_employ & people	
	employ	
		_employ
	employment opportunity	
		_employ opportunity
	work	
		_work
Industry		
	manufacturing industry	
		_manufacturing industry
	skilled technology	
		_skilled technology
	pharmacies	
		_pharmacies
	pharmacy network	
		_pharmacy network
	technology	
		_technology
human resources		
	_human resources	
	fulltime employees	
		_fulltime employees
	professionals	
		_professionals
		members
	technology researchers	
		_technology researchers
	ict professionals	
		_ict professionals
	indirect staff	
		_indirect staff
	developers	
		_developers
	researchers	
		_researchers
	staff strength	
		_staff strength

Model 1		
Categories	Subcategories	Rules and concepts
internationalization process		
	_internationalization process	
social		
	_social	
	cultural development1	
		_cultural development
	development	
		_development
	disabilities	
		_disabilities
	disabilities turn social	
		_disabilities turn social
	displaced families	
		_displaced families
	employ of refugee	
		_employ of refugee
	expansion	
		_expansion
	farmer members	
		_farmer members
	generate local incomes	
		_generate local incomes
	generate regional income	
		_generate regional income
	poverty	
		_poverty
	poverty reduction	
		_poverty reduction
	refugees	
		_refugees
	unemployed	
		_unemployed
	women	
		_women
	farmer groups	
		_farmer groups
	handicapped employ	
		_handicapped employ
	engagement	
		_engagement
	cultural development	
		_cultural development
	regular rural communities	
		_regular rural communities
	rural employ opportunity	
		_rural employ opportunity
	self-employment	
		_self-employment
	community	
		_community
	unemployed youths	
		_unemployed youths

Model 1		
Categories	Subcategories	Rules and concepts
Create jobs		
	_create & job	
	_creating & job	
	_generate & employ	
	_producing & job	
	_job create	
	job	
		_job
	job create	
		_job create
	Many job opportunity	
		_many & job opportunity
		_job opportunity
	Many jobs	
		_many & job
	Produce jobs	
		_produce & job
	job creator	
		_job creator
	new job	
		_new job
	job opportunity	
		_job opportunity
employ opportunity		
	_employ opportunity	
	adequate employ opportunity	
		_adequate employ opportunity
	extensive employement	
		_extensive employement
	generate employ	
		_generate employ
	higher employment opportunity	
		_higher & employment opportunity
	Extensive employment	
		_extensive employement
		higher quality employ
	Lots of employment	
		_lots of employment
		lots of employ
	serious employ	
		_serious employ

Model 2		
Categories	Subcategories	Rules and concepts
Employ people		
	Researcher	
		_hire researchers
	Professional	
		_job & professionals
		_hire & key professionals
		_hire & professionals
	People	
		_get & people
		_hire people
		_job & individuals
		_hire & people
		_job & persons
		_job & people
Social		
	_increase & workers	
	Local	
		_create & local job
		_job & local people
		_provide & local job
		_hire local software developers
	Rural	
		_job create & rural communities
		_create & rural job place
	Woman	
		_create & job & woman
	Young	
		_job & youths
		_hire & innovative young people
	Graduate	
		_policy of hire & fresh graduates
		_policy of hire & unemployed graduates
	Agriculture	
		_create & job opportunities & agricultural workers
		_increase & farmers
	Refugee	
		_job of refugee
	Source of rent	
		_provide & source of rent

Model 2		
Categories	Subcategories	Rules and concepts
Create job		
	_create & job	
	_create & rural job place	
	_job production	
	_job create	
	_get & job	
	Create new job	
		_create & new job
		_provide & new job
		_open new job
	Increase job	
		_share of job
		_increase & job
		_increase job
	Provide job	
		_provide & job
	Mandays	
		_create & mandays of job
	Additional job	
		_additional & job
Expansion		
	Expand	
		_will be employed & expansion
		_planning & expand
	Increase job	
		_preview & increase & job
	Double team	
		_double & manpower
		_double & team
Development		
	_job & development	
	Development benefit	
		_developmental benefit & job job
		_development benefit & job benefit
Reduce unemployment		
	Unemployment	
		_decline & unemployment
		_decline & egyptian unemployment
	Unemployment rate	
		_decline & unemployment rate

Model 2		
Categories	Subcategories	Rules and concepts
New business		
	Opportunity	
		_create & opportunities & new business
Industry		
	Job opportunity	
		_graha agri industry & provide & job opportunities
	Job creator	
		_manufacturing industry & job creator
Job opportunity		
	_job opportunities	
	New job	
		_new job opportunities
	Variety	
		_provide & variety job opportunities
	Rural job	
		_rural job opportunities
	Exciting	
		_create & exciting job opportunities
	Offer	
		_offers job opportunities
		_offer job
	Create	
		_create & job opportunities
		_provide & job opportunities
		_create & opportunities
Need of people		
	Skilled labor	
		_require & skilled labor
	Hire	
		_need & hire
		_need & employees
	Create job	
		_require & create & job