

Tabla de Contenido

1. Introducción	1
1.1. Antecedentes generales	1
1.2. Justificación y descripción del proyecto	4
1.2.1. Análisis de las alternativas de solución actual	4
1.2.2. Uso de antibióticos	5
1.2.3. Características de la solución que se necesita	6
1.2.4. Beneficios de la implementación de un nuevo modelo	7
1.2.5. Hipótesis de Investigación	8
1.3. Objetivos	9
1.3.1. Objetivos Generales	9
1.3.2. Objetivos Específicos	9
1.4. Apoyos institucionales	9
1.5. Metodología	10
1.5.1. Investigación de puntuaciones de riesgo en casos de neutropenia febril	10
1.5.2. Implementación y evaluación de los modelos hoy existentes	10
1.5.3. Procesamiento y creación de la base final de datos	11
1.5.4. Investigación y selección de los modelos a aplicar	11
1.5.5. Creación del modelo final de machine learning	12
1.5.6. Evaluación de la interpretabilidad y rendimiento de los los modelos .	12
1.5.7. Prototipo del módulo de visualización de riesgo	12
1.5.8. Evaluación del impacto económico y social de la solución	13
1.6. Alcances	13
1.7. Resultados esperados	14
1.8. Estructura del informe	15
2. Marco Teórico	16
2.1. Herramientas tecnológicas y datos utilizados	16
2.1.1. Datos utilizados	16
2.1.2. Python	16
2.1.3. Django	16
2.1.4. Jupyter Notebook	17
2.1.5. Git	17
2.2. Conceptos médicos relevantes	17

2.2.1.	Cáncer	17
2.2.2.	Tipos de cáncer	18
2.2.3.	Tratamientos para el cáncer	18
2.2.4.	Neutropenia Febril	19
2.2.5.	Tratamiento para la NF	20
2.2.6.	Complicaciones	21
2.2.7.	Cáncer y neutropenia febril en niños	22
2.3.	Puntajes de predicción de riesgo	22
2.4.	Descubrimiento de Conocimiento en Base de Datos	23
2.5.	Modelos ML	25
2.5.1.	Logistic Regression	25
2.5.2.	Support Vector Machine (SVM)	25
2.5.3.	Clasificador Bayesiano Ingenuo	25
2.5.4.	Arboles de decisión	26
2.5.5.	Random Forest	26
2.5.6.	Gradient Boost Machine	27
2.5.7.	Redes neuronales	28
2.6.	Conceptos relevantes de machine learning y diseño de aplicaciones Web	28
2.6.1.	Matriz de confusión y métricas de desempeño asociadas	28
2.6.2.	Validación cruzada de K conjuntos	30
2.6.3.	Modelo Vista Controlador	31
2.6.4.	Métodos de selección de variables	31
2.6.5.	Balanceo de datos	33
3.	Desarrollo del modelo final de clasificación	34
3.1.	Aplicación de los modelos de estratificación existentes	34
3.1.1.	Modelo propuesto por Rondinelli et al.	34
3.1.2.	Modelo de Das et al.	36
3.1.3.	Modelo propuesto por Agyeman et al.	38
3.1.4.	Modelo propuesto por Hakim et al.	40
3.1.5.	Modelo propuesto por Santolaya et al.	41
3.1.6.	Modelos no aplicados	42
3.1.7.	Resumen modelos aplicados	43
3.2.	Selección y creación de conjuntos de variables para los modelos	45
3.2.1.	Modelo utilizando variables encontradas en la literatura	46
3.2.2.	Métodos de filtro	48
3.2.3.	Balanceo y transformación de datos	52
3.2.4.	Métodos envolventes	52
3.3.	Desarrollo del modelo de clasificación	59
4.	Implementación del modelo de visualización	63
4.1.	Descripción del prototipo desde la vista del usuario	63
4.2.	Creación del modelo funcional	65
5.	Estimación del impacto económico del modelo	68
5.1.	Impacto para la red PINDA metropolitana	68
5.1.1.	Ahorro directo por menores costos de tratamiento	69

5.2. Impacto para Chile	71
5.2.1. Ahorro directo por menores costos de tratamiento	71
5.3. Impacto en el tratamiento general de neutropenia febril	73
6. Conclusiones	74
6.1. Conclusiones generales	74
6.2. Trabajo Futuro	76
7. Anexos	78
7.1. Marco teorico	78
7.1.1. Variables por modelos de la literatura	78
7.2. CRISP-DM	79
7.3. Mejores variables por modelos	80
8. Bibliografia	80

Índice de Tablas

1.1. Comparación estratificación de riesgo calculada versus declarada, basada en las tablas entregadas por el PINDA	7
2.1. Tabla de sensibilidad y especificidad para los modelos de Das et al [1]	22
2.2. Ejemplo matriz confusión.	29
3.1. Tabla de puntajes de riesgo, elaboración propia en base a [2]	35
3.2. Rendimiento modelo Rondanelli et al., elaboración propia	36
3.3. Tabla de puntajes de riesgo Das et al., elaboración propia en base a [1] . . .	37
3.4. Rendimiento modelo Das et al., elaboración propia	38
3.5. Tabla de puntajes de riesgo, elaboración propia en base a [3]	39
3.6. Rendimiento modelo de Agyeman et al., elaboración propia	39
3.7. Tabla de puntajes de riesgo Hakim, elaboración propia en base a [4]	40
3.8. Rendimiento modelo de Hakim et al., elaboración propia	41
3.9. Rendimiento modelo de Santolaya et al., elaboración propia	42
3.10. Variables utilizadas para cada modelo	43
3.11. Mejor rendimiento para cada modelo	44
3.12. Mejor rendimiento para cada modelo	47
3.13. Variables eliminadas a través del método de variabilidad, elaboración propia	48
3.14. 10 variables con menor significancia en chi-cuadrado, basada en las bases entregadas por el PINDA	50
3.15. 15 variables con menor significancia en t-test, elaboración propia	51
3.16. Coeficientes de la regresión logística, modelo Rondinelli, elaboración propia .	53
3.17. Coeficientes de la regresión logística, modelo Agyeman, elaboración propia .	54
3.18. Coeficientes de la regresión logística, modelo Hakim, elaboración propia . . .	54
3.19. Coeficientes de la regresión logística, modelo Das, elaboración propia	55
3.20. Coeficientes de la regresión logística, modelo Santolaya, elaboración propia .	56
3.21. Variables utilizadas para cada modelo y su correspondiente dirección	56
3.22. Variables utilizadas para cada modelo y su correspondiente dirección	57
3.23. Variables utilizadas para cada modelo y su correspondiente dirección	61
5.1. Episodios de neutropenia febril por mes durante 2009 al 2015, elaboración propia en base a datos del PINDA	69

5.2. Porcentaje de mejores clasificados, de acuerdo al umbral definido para la clasificación	70
5.3. Evolución población Chilena hasta los 18 años de edad, incluyendo la cantidad de casos de neutropenia febril estimados.	72
7.1. Modelos utilizados en el mundo para estratificar riesgo, elaboración propia en base a Das 2017	79
7.2. Resumen rendimientos modelos con métodos inductivos y deductivos	81

Índice de Ilustraciones

2.1. Tratamiento de neutropenia febril, Paganini 2011	21
2.2. Ciclo de vida de un proyecto basado en KDD, fuente IBM [5]	24
2.3. Ejemplo Curva AUC	30
3.1. Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo de Rondanelli et al.,elaboración propia	36
3.2. Elaboración propia en base a base del PINDA	37
3.3. Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Das et al,elaboración propia	38
3.4. Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Agyeman, elaboración propia	40
3.5. Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Hakim, elaboración propia	41
3.6. Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Santolaya, elaboración propia	43
3.7. Gráfico AUC de los diferentes modelos, elaboración propia	44
3.8. Rendimiento por cantidad de variables, elaboración propia	60
4.1. Vista del usuario del prototipo funcional al ingresar datos	65
4.2. Vista del usuario del prototipo funcional al mostrar el resultado de la clasificación	66
7.1. Ciclo de vida de un proyecto basado en CRISP DM, fuente IBM [6]	80