



UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA

**ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE
SISTEMAS**

“Aplicación de minería de datos para mejorar el diagnóstico de la
tuberculosis pulmonar en el Hospital de EsSalud “Aurelio Díaz Ufano y
Peral”

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

AUTOR:

Huerta Carrión, Jimmy Roger

ASESOR:

Ing. Fernando Mendoza Apaza

LÍNEA DE INVESTIGACIÓN:

Sistemas de información estratégica y de toma de decisiones

LIMA-PERÚ

2014

El Jurado encargado de evaluar la tesis presentada por don (a) Huerta Carrión Jimmy Huerta cuyo título es: Aplicación de minería de datos para mejorar el diagnóstico de la tuberculosis pulmonar en el hospital de EsSalud Aurelio Díaz Ufano y Peral.

Reunido en la fecha, escuchó la sustentación y la resolución de preguntas por el estudiante, otorgándole el calificativo de: 17 (Diecisiete).


Lima, San Juan de Lurigancho 05 de diciembre del 2014



PRESIDENTE



SECRETARIO



VOCAL



Elaboró

Dirección de
Investigación

Revisó



Responsable del SGC



Aprobó

Vicerrectorado
de Investigación

Dedicatoria

Dedico esta tesis a Dios, que me guardo en este tiempo me dio sabiduría y no dejo que caiga en desgano ni conformidad, a mi madre Adelina Carrión Flores que, con su fe, su amor y su fuerza siempre está para incentivar me a seguir adelante.

Agradecimiento

A mis hermanas Yuliza y Morehima a Juan Morales que ha sido como un padre para mí, a mis docentes y compañeros por lo que compartimos una vida de experiencias, aprendizaje, a la familia Varillas Carreño, y a todos q aportaron en esta etapa de mi vida.


Declaratoria de autenticidad

Yo, Jimmy Roger Huerta Carrión, estudiante de la facultad de Ingeniería de Sistemas de la Escuela de Pregrado de la Universidad César Vallejo, identificada con DNI 41499515, con la tesis titulada "Aplicación de la técnica de árboles de decisión para predecir el diagnóstico de la enfermedad de tuberculosis pulmonar en el Hospital de EsSalud Aurelio Díaz Ufano y Peral "declaro bajo juramento que:

1. La tesis es de mi autoría.
2. He respetado las normas internacionales de citas y referencias para las fuentes consultadas. Por tanto, la presente tesis no ha sido plagiada ni total ni parcialmente.
3. La tesis no ha sido auto plagiada; es decir, no ha sido publicada ni presentada anteriormente para obtener algún grado académico previo o título profesional.
4. Los datos presentados en los resultados son reales, no han sido falseados, ni duplicados, ni copiados y por tanto los resultados que se presenten en la tesis se constituirán en aportes a la realidad investigada.

De identificarse la falta de fraude (datos falsos), plagio (información sin citar a autores), autoplagio (presentar como nuevo algún trabajo de investigación propio que ya ha sido publicado), piratería (uso ilegal de información ajena) o falsificación (representar falsamente las ideas de otros), asumo las consecuencias y sanciones que de mi acción se deriven, sometiéndome a la normatividad vigente de la Universidad César Vallejo.

San Juan de Lurigancho, 12 de diciembre del 2014



.....
Jimmy Roger Huerta Carrión

Presentación


Miembros del Jurado:

Dando cumplimiento a las normas establecidas en el Reglamento de Grados y Títulos de la Universidad César Vallejo para obtener el título de Ingeniería de Sistemas, presento el trabajo de investigación que tiene como título: "Aplicación de la técnica de árboles de decisión para predecir el diagnóstico de la enfermedad de tuberculosis pulmonar en el Hospital de EsSalud Aurelio Díaz Ufano y Peral".

La investigación tiene como propósito fundamental: Determinar el algoritmo de aprendizaje más adecuado de minería de datos que permitirá realizar el diagnóstico de la tuberculosis pulmonar en el hospital de EsSalud Aurelio Díaz Ufano y Peral.

La presentación investigación está estructurada bajo el esquema de siete capítulos: en el capítulo I se expone la introducción, en el capítulo II se expone el marco metodológico y el método de investigación, en el capítulo III se muestra los resultados de la investigación, en el capítulo IV se expone las discusiones, en el capítulo V las conclusiones, en el capítulo VI se expones las recomendaciones, en el capítulo VII se especifica las referencias bibliográficas y los anexos.

Espero que la presente investigación se ajuste a los requerimientos establecidos y que este trabajo sirva como base para posteriores estudios.



.....
Jimmy Roger Huerta Carrión

ÍNDICE

Página del Jurado	ii
Dedicatoria	iii
Agradecimiento	iv
Declaratoria de autenticidad.....	v
Presentación	vi
Resumen.....	10
Abstract	11
I. INTRODUCCIÓN.....	12
1.1 Realidad problemática.....	12
1.2 Trabajos Previos.....	16
1.3 Teoría relacionada sobre el tema	22
1.4 Formulación del Problema.....	46
1.5 Justificación del estudio.....	47
1.6 Hipótesis.....	48
1.7 Objetivos	48
II. Método	50
2.1 Diseño de investigación.....	50
2.2 Operacionalización de variables:.....	53
2.3 Población, muestra.....	55
2.4 Técnicas e instrumentos de recolección de datos, validez y confiabilidad ..	56
2.5 Métodos de análisis de datos	57
2.6 Aspectos Éticos	57
III. RESULTADO.....	59
IV. DISCUSIÓN.....	89
V. CONCLUSIONES.....	92
VI. RECOMENDACIONES	94
VII. REFERENCIAS BIBLIOGRAFICAS	95
ANEXOS	104
Anexo 01. Matriz de Consistencia.....	105
Anexo 02: Instrumento	107
Anexo 03: Validación de instrumento	108
Anexo 04: Acta de aprobación de originalidad de tesis.....	118
Anexo 05: Autorización de publicación de tesis	120

Anexo 06: Autorización de la versión final del trabajo de investigación	121
---	-----

ÍNDICE DE TABLAS

Tabla 1: Resultado de las pruebas de los modelos del trabajo de Asia Nездredin..	21
Tabla 2: Resultado de las pruebas de los modelos del trabajo de Nagabhushanam D, Naresh N, Raghunath A, PraveenKumar.	22
Tabla 3: Operacionalización de Variables	53
Tabla 4: Resumen del procesamiento de los casos	59
Tabla 5: Estadísticos de fiabilidad	59
Tabla 6: Abreviación y descripción de las variables	64
Tabla 7: Posibles valores para datos recolectados	65
Tabla 8: Comparación de resultados de los algoritmos de selección de atributos	67

ÍNDICE DE FIGURAS

Figura 1: Proporción de casos nuevos de tuberculosis por regiones de salud, Peru-2012	14
Figura 2: Tasa de mortalidad por 100000 habitantes por tuberculosis por distrito DISA IV LE, 2006- 2012.....	16
Figura 3: Etapas de la minería de datos.....	28
Figura 4: Esquema de una red neuronal totalmente conectadas	34
Figura 5: Esquema de un perceptrón multicapa, con una capa oculta	36
Figura 6: estructura de una red bayesiana que representa el funcionamiento de un automóvil.....	38
Figura 7: Ejemplo de 3 nodos conectados en serie.....	39
Figura 8: Ejemplo de una conexión divergente	39
Figura 9: Ejemplo de una conexión convergente	40
Figura 10: Topología de un clasificador Naive Bayes	40
Figura 11: Estructura algoritmo TAN	41
Figura 12: Fases del modelo CRISP-DM	46
Figura 13: Relación de variable independiente y dependiente	50
Figura 14: Base de datos Tuberculosis Data-weka	66
Figura 15: Análisis gráfico de relación de los atributos Disnea y Class.....	68
Figura 16: Resultados generados por el algoritmo J48	69
Figura 17: Matriz de confusión con los parámetros por defectos del algoritmo J48.....	70

Figura 18: Visualización grafica del resultado del modelo de árbol de decisiones J48.....	73
Figura 19: Resultados del modelo clasificador MLP.....	74
Figura 20: Resumen de evaluación de los datos entrenados con MLP.....	77
Figura 21: Resumen detallado de precisión por clase y matriz de confusión del MLP.....	77
Figura 22: Red neuronal perceptrón multicapa	79
Figura 23: Resultado gráfico del modelo clasificador de Redes bayesianas TAN...80	
Figura 24: Resumen de evaluación de los datos entrenados con red bayesiana TAN.....	81
Figura 25: Resumen detallado de precisión por clase y matriz de confusión con red bayesiana TAN.....	82
Figura 26: Reglas generadas por el modelo de clasificación de árboles de decisión J48.....	87

ÍNDICE DE GRAFICOS

Gráfico 1: Pacientes atendidos en el PCT del HADUYP	60
Gráfico 2: Número total pacientes de Tb pulmonar por rango de edad.....	61
Gráfico 3: Precisión de los algoritmos de clasificación.....	82
Gráfico 4: Tiempo tomado para el entrenamiento del modelo de prueba con datos.....	83

Resumen

Con la minería de datos es posible extraer patrones ocultos a partir de grandes depósitos de datos, en otras palabras, es transformar esos grandes depósitos de datos en información. Esos patrones extraídos se utilizan para interpretar los datos nuevos o existentes en una información útil para realizar predicciones, optimizar procesos, diagnosticar enfermedades, y en otras áreas.

Con colaboración de la tecnología informática se han implementado técnicas de minería de datos como las redes neuronales, clustering, algoritmos genéticos, arboles de decisión y las máquinas de vectores soporte, la minería de datos puede aplicarse en todas las organizaciones donde se almacene datos, como en negocios, hospitales, laboratorios u otros.

La presente tesis propone determinar el mejor algoritmo de las técnicas de clasificación de minería de datos para predecir el diagnóstico correcto de la enfermedad de tuberculosis a partir de la población de pacientes con diagnóstico de tuberculosis del programa de control y prevención de tuberculosis del hospital de EsSalud Aurelio Díaz Ufano del distrito de San Juan de Lurigancho, se propone un muestreo aleatorio simple de 502 pacientes, el diseño de la investigación fue pre-experimental, la técnica de recolección de datos fue el fichaje y el instrumento fue la ficha de registro los cuales fueron validados por el juicio de expertos; se detalla aspectos teóricos del proceso de minería de datos, herramientas y la metodología para determinar la mejor técnica de diagnóstico de la tuberculosis, siendo la mejor herramienta de minería de datos Weka y la mejor metodología de desarrollo Crisp-DM.

Finalmente, se demuestra que las técnicas de clasificación de árboles de decisión con el algoritmo J48, redes neuronales artificiales con el algoritmo perceptrón multicapa y las redes bayesianas con el algoritmo TAN cumplieron con los objetivos de la investigación, permitieron una clasificación correcta de datos con una precisión muy alta de 0.988 %, 0.998 % y 0.988 % respectivamente, y en un periodo de tiempo muy corto, llegando a la conclusión que las herramientas de minería de datos pueden mejorar con el diagnóstico de la enfermedad.

Palabras claves: minería de datos, algoritmos, técnicas de clasificación, diagnostico, tuberculosis.

Abstract

With data mining it is possible to extract hidden patterns from large data repositories, in other words it is transforming those large data repositories into information. These extracted patterns are used to interpret new data or are used in useful information to make predictions, optimize processes, diagnose diseases, and in other areas.

With the collaboration of computer technology, data mining techniques such as neural networks, grouping, genetic algorithms, decision trees and supporting support machines have been implemented; data mining can be carried out in all the organizations where they are stored data such as in business, hospitals, laboratories or others.

This thesis proposes to determine the best algorithm of data classification techniques to predict the correct diagnosis of tuberculosis disease from the population of patients diagnosed with TB tuberculosis control and prevention program of the hospital of EsSalud Aurelio Diaz Ufano from the district of San Juan de Lurigancho, a simple random sample of 502 patients is proposed, the design of the research was pre-experimental, the data collection technique was the signing and the instrument was the registration form of which they were validated by expert judgment; the theoretical details of the data mining process, tools and methodology to determine the best diagnostic technique for tuberculosis, being the best Weka data mining tool and the best Crisp-DM development methodology.

Finally, it is demonstrated that the techniques of classification of decision trees with the J48 algorithm, artificial neural networks with the multilayer perceptron algorithm and Bayesian networks with the TAN algorithm fulfilled the research objectives, allowed a correct classification of data with a very high precision of 0.988%, 0.998% and 0.988% respectively, and in a very short period of time, reaching the conclusion that the tools of data mining can improve with the diagnosis of the disease.

Key words: data mining, algorithms, classification techniques, diagnosis, tuberculosis.

Anexo 04:

 UCV UNIVERSIDAD CÉSAR VALLEJO	ACTA DE APROBACIÓN DE ORIGINALIDAD DE TESIS	Código : F06-PP-PR-02.02 Versión : 09 Fecha : 23-03-2018 Página : 1 de 1
--	---	---

Yo, María Eudelia Acuña Meléndez, docente de la Facultad de Ingeniería y Escuela Profesional de Ingeniería de Sistemas de la Universidad César Vallejo Lima Este, revisor (a) de la tesis titulada

“Aplicación de minería de datos para mejorar el diagnóstico de la tuberculosis pulmonar en el hospital de EsSalud Aurelio Díaz Ufano y Peral”, del (de la) estudiante Jimmy Roger Huerta Carrión, constato que la investigación tiene un índice de similitud de 30% verificable en el reporte de originalidad del programa Turnitin.

El/la suscrito (a) analizó dicho reporte y concluyó que cada una de las coincidencias detectadas no constituyen plagio. A mi leal saber y entender la tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad César Vallejo.

Lima, San Juan de Lurigancho 11 de abril de 2019



Firma

María Eudelia Acuña Meléndez

DNI: 19083126

					
Elabora	Dirección de Investigación	Revisó	Responsable del SGC	Aprobó	Vicerectorado de Investigación