Diabetes Risk Level Prediction Using Data Mining Techniques

By:
Shajan Mohammed Mahdi

Supervisor
Dr. Aysh Alhroob

The Thesis was Submitted in Partial Fulfillment for the Requirements of the Master’s Degree in Software Engineering

Faculty of Graduate Studies
Isra University

June, 2019
The undersigned have examined the thesis entitled Diabetes Risk Level Prediction Using Data Mining Techniques presented by Student Name, a candidate for the degree of Master of Science in Software Engineering and hereby certify that it is worthy of acceptance.

_____________________ _____________________
Date

Dr.Aysh Alhroob

_____________________ _____________________
Date

Dr. Adnan Hadi

_____________________ _____________________
Date

Dr. Mohammed Ali Alabbadi
جامعة الإسراء

إقرار تفويض

أنا شجن محمد مهدي ، أفوض جامعة الإسراء بتزويج نسخ من رسالتي ورقية وإلكترونيا للمكتبات أو المنظمات أو الهيئات والمؤسسات المعنية بالأبحاث والدراسات العليا عند طلبها.

التوقيع:
التاريخ:
AUTHORIZATION STATEMENT

I, Shajan Mohammed Mahdi, authorize Isra University to provide hard copies or soft copies of my thesis to libraries, institutions or individuals upon their request.

Signature:

Date:
الإهداء

إلى... النخلة التي ألَّفْتَ روحي
إلى... الجذع الذي حملنا معاً مكاناً وشامخاً
إلى كلٍّهما وهما يحملان على أكاسياتهم وبحقان بي بعيداً نحو سماء النجاح
لك وانت تجد في غرسي داخل رحم الحياة
لك وانت تتمضين بي مراراً وتناوهي فرصة البدايات الجديدة على اعتاب النهاية
لكما وانتما تتجاربان الكأس فارغًا تسدقياً ينابيع الحب
أمي... أبي

ولتلك الروح التي علمني بأن الحضور قد يعني الفيائل المكتمل أو الرحيل الأبدى الذي للاقطاعاً غوادة ولايشود نبض
لروحك جدتي

إلى كل من محتشي الدفء في صناع الغربة ويُزغ قمرأناً تحالكت في عيني النّا

أهدي هذا الاجتهد
شكر وإمتنان

(رب أورغَني أن أشكر نعمتك اللي أعلم علي وعلي وليذويو أن عمل صالحًا فرضه وأدخلني
برحمتك في عداد الصالحين) (91) النمل

أشكر الله الذي أستغنى على بفوف نعه وواسع توفيقه وكريمه

ولمشرفي واستاذي الجليل الدكتور "عباش الحروب" الذي انتهت من مناهج علمه ومعرفته دون مثل منه أو كلونه الذي
غمرني بفوف ثقه وتشجيعه الذي كان عكازي في اشد لحظات التردد وأعى مراحل الضغط

كل من اشعر قيسا للعلم في دربي .. استاذتي الأفضل

لعائلتي التي مبارحت ان تكون لي سند شدة وداعما

احباني واصدقاني وكل من وقف معي لتحقيق حلمي والتشديث به

واخيرا للروح التي تسكن اعماقي ولم تضجه عواصف اليأس ولم تثثها غيوم التفاوض والعجز .
DEDICATION

To the beloved lady who suffered, cared, and prayed for my success, my mother.

To the great man who always supported me in every step of my life, my father.

To the candles of my life; my brothers & sisters, and all the faithful friends for their unlimited love and support, for all them, I dedicate this humble work.
ACKNOWLEDGMENTS

First of all, I give thanks and praise to Allah for his mercy, and his blessing for giving me the opportunity for this achievement.

I would like to express my thanks and gratitude to my supervisor Dr. Aysh Alhroob for his continuous academic support, suggestions, guidance, and patience throughout the project progress, besides his brotherly care.

Thanks are also Dr. Wael Alzyadat, and the Faculty members who were generous in their teaching and academic services.

Finally, my deepest thanks and unlimited love for my family and all my friends who were so helpful and supportive for me to complete my thesis.
# Table of Contents

Dedication ........................................................................................................... V
Acknowledgments ............................................................................................. VI
Table of Contents ............................................................................................... VII
List of Figures ...................................................................................................... X
List of Tables ....................................................................................................... IX
List of Equations ................................................................................................. XI
List of Abbreviations ......................................................................................... XII

ABSTRACT ........................................................................................................ XIII

1 CHAPTER ONE (Introduction) ......................................................................... 1
1.1 Overview .................................................................................................... 1
1.2 Problem Statement .................................................................................... 2
1.3 Research Questions .................................................................................... 2
1.4 Aim and Objectives ................................................................................... 2
1.5 Motivations ............................................................................................... 3
1.6 Significance ............................................................................................... 3
1.7 Scope ......................................................................................................... 3
1.8 General Methodology ............................................................................... 3
1.8.1 ............................................................................................................... 4
1.9 Using Data Mining Techniques and AI appellations for Diabetic Diagnostic .......6
1.9.1 Level One (preprocess data) .................................................................. 7
1.9.2 Level two (identify risk levels) ............................................................... 7
1.9.3 Level three (Learning) ............................................................................ 8
1.9.4 Level four (evaluations) ........................................................................ 8

2 CHAPTER TWO (Background and Previous Work) ........................................ 9
2.1 Overview .................................................................................................... 9
2.2 Terminologies ............................................................................................. 9
2.2.1 Data mining .......................................................................................... 9
2.2.2 Big data ............................................................................................... 10
2.2.3 Machine learning (ML) ....................................................................... 10
2.2.4 Fuzzy C-means ................................................................................... 10
2.2.5 Classification ....................................................................................... 11
2.2.6 Deep Learning .................................................................................... 12
2.3 Related works .......................................................................................... 13
2.4 Summary ................................................................................................... 20

3 Chapter Three (Proposed Approach) ............................................................ 22
LEST OF TABLES

Table 2-1 : Summary of related work ................................................................................................................. 20
Table 3-1: Description of Data Set ........................................................................................................................ 24
Table 3-2: Risk matrix – Low level ......................................................................................................................... 33
Table 3-3: Confusion Matrix ................................................................................................................................. 35
Table 4-1: Deep Learning Result .......................................................................................................................... 50
## List of figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>General Research Methodology to handle the Thesis.</td>
<td>4</td>
</tr>
<tr>
<td>1-2</td>
<td>Layers of Approach.</td>
<td>6</td>
</tr>
<tr>
<td>3-1</td>
<td>Model of Approach.</td>
<td>23</td>
</tr>
<tr>
<td>3-2</td>
<td>Data Pre-pressing</td>
<td>28</td>
</tr>
<tr>
<td>4-1</td>
<td>Data Collection</td>
<td>39</td>
</tr>
<tr>
<td>4-2</td>
<td>Input Data</td>
<td>40</td>
</tr>
<tr>
<td>4-3</td>
<td>Data View</td>
<td>40</td>
</tr>
<tr>
<td>4-4</td>
<td>Explore Data</td>
<td>41</td>
</tr>
<tr>
<td>4-5</td>
<td>Remove Attribute (Weight)</td>
<td>42</td>
</tr>
<tr>
<td>4-6</td>
<td>Remove Attribute (Payer_code)</td>
<td>42</td>
</tr>
<tr>
<td>4-7</td>
<td>Remove Attribute (medical specialty)</td>
<td>42</td>
</tr>
<tr>
<td>4-8</td>
<td>Remove Duplicate Instances</td>
<td>43</td>
</tr>
<tr>
<td>4-9</td>
<td>Weight of Attributes</td>
<td>44</td>
</tr>
<tr>
<td>4-10</td>
<td>Clustering</td>
<td>45</td>
</tr>
<tr>
<td>4-11</td>
<td>Clustering</td>
<td>46</td>
</tr>
<tr>
<td>4-12</td>
<td>SVM Outcomes</td>
<td>47</td>
</tr>
<tr>
<td>4-13</td>
<td>Low Level Rule</td>
<td>48</td>
</tr>
<tr>
<td>4-14</td>
<td>Deep Learning Report</td>
<td>49</td>
</tr>
<tr>
<td>4-15</td>
<td>Deep Learning Report</td>
<td>50</td>
</tr>
<tr>
<td>4-16</td>
<td>Evaluation the Model</td>
<td>51</td>
</tr>
<tr>
<td>Equation</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>3-1 Correlation Equation</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>3-2 Accuracy Equation</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>3-3 Accuracy Equation</td>
<td>36</td>
<td></td>
</tr>
</tbody>
</table>
**List of Abbreviations**

<table>
<thead>
<tr>
<th>#</th>
<th>Abbreviation</th>
<th>Full Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>2.</td>
<td>CSV</td>
<td>Comma-Separated Values</td>
</tr>
<tr>
<td>3.</td>
<td>FCM</td>
<td>Fuzzy C-means</td>
</tr>
<tr>
<td>4.</td>
<td>GDA</td>
<td>Generalized DiscriminantAnalysis</td>
</tr>
<tr>
<td>5.</td>
<td>LR</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>6.</td>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>7.</td>
<td>NNs</td>
<td>Neural Networks</td>
</tr>
<tr>
<td>9.</td>
<td>T2DM</td>
<td>Type Two of Diabetes Mellitus</td>
</tr>
<tr>
<td>10.</td>
<td>UCI</td>
<td>University of California, Irvine</td>
</tr>
<tr>
<td>11.</td>
<td>WEKA</td>
<td>Waikato Environment for Knowledge Analyze</td>
</tr>
</tbody>
</table>
Abstract

Big data faces many challenges in various aspects that appear through characteristics such as: volume, velocity, and variety; big data processes and analysis challenges acquiring quality information to support accurate decision-making values. Health care produces large amount of data by follow up the patients. This data can be used for diagnosing, detecting abnormal behavior and decision-making. Nevertheless, in critical fields that are directly related to human health care, the data must be treated in manner to overcome unwanted medical actions related to Big Data. Diabetics Big Data is rich in medical details, due to the frequency of updating case, and rich in gaps and unwanted data as well. Therefore, precise work on big data makes the diagnoses prediction of diabetics in terms of risk level possible. This prediction helps the doctor to overcome the ambiguous problem of the case in future and predict the optimal treatment at early stage of the case. In this work, an approach is proposed to pre-process the benchmark dataset UCI and select the correlated features based on target attribute. Fuzzy C-Means is used to values clustering and Support Vector Machine (SVM) is used for classification as well. Clustering and classification techniques are used to increase the clarity of data to enrich the rules that will be generated from dataset. Risk Matrix was proposed to represent rules of three levels of diabetes (low, high, medium), and use Risk Matrix to train deep learning and build an expert system that can predict the risk level automatically. The approach is tested in the fourth layer using the evaluation Metrics of machine learning algorithms. The approach experiments use Diabetes patient data and symptom in rapidminer tool. This approach Achieved 97.8% accuracy to automatically predict the level of risk and can be applied at the field of health care to target diabetic patients at variant levels of risks and provide customized care to reduce the re-admission rate.

Keywords: Big Data, Fuzzy C-Means, Diabetic, Healthcare, Support Vector Machine (SVM), Risk Matrix.