IU

جامعة الإسراء

Isra University

Diabetes Risk Level PredictionUsing 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

DateDr.Mohammed Ali Alabbadi

جامعة الإسراء

إقرار تفويض

أنا شجن محمد مهدي ، أفوض جامعة الإسراء بتزويد نسخ من رسالتي ورقياً 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 copes or soft copes of my thesis to libraries , institutions or individuals upon their request.

Signature :

Date:

ألإهداء

إلى.. النَخلةِ التي أثمَرتْ روحي

إلى.. الجِذعِ الذي حَمَلنا معاً مُكابراً وشامِخاً

إلى كِليهُما وهما يَحمِلاني على أكتافِألأملِ ويُحلِقان بي بَعيداً نَحو سماءِ النِجاح

لكَ وأنتَ تَجُد في غَرسي داخِل رحِم الحَياة

لكِ وأنتِ تَتَمخَضين بي مِراراً وتَمنَحيني فُرصة البدايات الجَديدة على أعتابِ النهاية

لَكما وأنتما تَتَجرَعان الكأسَ فارِغاً لتُسقيانا يَنابيع الحب

أمي .. أبي

ولِتلكَ الروح التي عَلمتني بأن الحُضور قد يَعني الغِيابَ المُكتَمل او الرَحيل ألأبدي الذي لاتُقاطعهُ عَودة ولايَشوبهُ وِصال

لروحكِ جدتي

إلى كُلِ مَن مَنحَني الدفءَ في صَقيعِ الغُربة وبَزغَ قمراًإن تَحالَكت في عَيني الدُنا

أهدي هذا ألاجتهاد

شكر وإمتنان

﴿رَبِّ أَوْزِعْنِيأَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَىٰ وَالِدَيَّوَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ وَأَدْخِلْنِي
بِرَحْمَتِكَ فِي عِبَادِكَ الصَّالِحِين﴾ (َ١٩) النمل

أشَكرُ الله الذي أسَبَغ عليَ بفيضِ نِعمه وواسع تَوفيقهُ وكرمه

ولمشرفي واستاذي الجليل الدكتور "عايش الحروب" الذي انتهلت من مناهل علمه ومعرفته دون ملل منه او كللوالذي
غمرني بفيض ثناءه وتشجيعه الذي كان عكازي في اشد لحظات التردد واعتى مراحل الضغط

لكل من اشعل قبساً للعلم في دربي .. اساتذتي الافاضل

لعائلتي التي مابرحت ان تكون لي سنداً وداعماً

احبائي واصدقائي وكل من وقف معي لتحقيق حلمي والتشبث به

واخيرا للروح التي تسكن اعماقي ولم تُضعِفها عواصف اليأس ولم تُثنها غيوم التقاعس 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 continuousacademic 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

LEST OF TABLES

List of figures

List of Equation

List of Abbreviations

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

Abstract

Big data faces many challenges in various aspects that appear through characteristicssuch As: volume, velocity, and variety; big data processes and analyzis 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 ambiguousproblem 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.