Faculty of Information and Communication Technology

AN IMPROVED DIABETES RISK PREDICTION FRAMEWORK: AN INDONESIAN CASE STUDY

Daniel Hartono Sutanto

Doctor of Philosophy

2018

Faculty of Information and Communication Technology

AN IMPROVED DIABETES RISK PREDICTION FRAMEWORK: AN INDONESIAN CASE STUDY

Daniel Hartono Sutanto

Doctor of Philosophy

2018

AN IMPROVED DIABETES RISK PREDICTION FRAMEWORK: AN INDONESIAN CASE STUDY

DANIEL HARTONO SUTANTO

A thesis submitted
in fulfillment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2018

DECLARATION

I declare that this thesis entitled "An Improved Diabetes Risk Prediction Framework: An Indonesian Case Study" is the result of my own research except as cited in the references. The thesis has not been accepted or not concurrently submitted in the candidature of any other degree.

Signature : ……..

Name : Daniel Hartono Sutanto

Date : ……..

APPROVAL

I declare that I have read this thesis and in my opinion, this thesis is sufficient in terms of scope and quality for the award of Doctor of Philosophy.

Signature : ……..

Supervisor Name : Professor Dr. Mohd. Khanapi Abd. Ghani

Date : ……..

DEDICATION

A special feeling gratitude to God Jesus Christ for the completion of my Ph.D. with a great blessing and experiences. I dedicate my work to my wife, daughter, beloved parents, and my sister. I am highly thankful and always appreciate the love, prayer, support, and consideration.

ABSTRACT

Lack of diagnosis for diabetes often transpire in some ASEAN countries with relatively diminutive doctor to patient ratio. Essentially, it is believed that a systematic framework to predict diabetes risk factors is crucial for refining diagnostics and improving accuracy. However, there is the issue of noisy dataset detected as incomplete data and the outlier class problem that affects sampling bias. Existing frameworks were deemed difficult in identifying the critical risk factors of diabetes; some of which were considerably inaccurate and consume substantial computation time. The purpose of this study is to develop a suitable framework for predicting diabetes risks. From a complete blood test, the framework can predict and classify the output of either having diabetes risk or no diabetes risk. A Diabetes Risk Prediction Framework (DRPF) was developed from the literature review and case studies were afterwards conducted in three private hospitals in Semarang. Analyses were conducted to find a suitable component of the framework—due to lack of comparison and analysis on the combination of feature selection and classification algorithm. DRPF comprises four main sections: pre-processing, outlier detection, risk weighting, and learning. Pre-processing resolves the issue of missing data and hence normalizes the data. Outlier treatment employs k-mean clustering to validate the class. Suitable components were selected through comparison of classifier algorithms and feature selection. Attribute weighting based feature selection was selected for assigning weightage. Weighted risk factor was used on training dataset in order to improve accuracy and computation time of the prediction. In the learning section, Support Vector Machine and Artificial Neural Network were selected as suitable classification algorithms, while Gradient Boosted Tree was employed to interpret the rule based on the black box classifiers. Testing the framework involved Pima Indian Dataset as public dataset and Semarang Hospital Dataset as private dataset (800 patients' data). In validating the DRPF, four case studies investigated Subject Matter Expert (SME) groups based on the agreement level. The questionnaire consists of a DRPF component, implementation of DRPF, and viability of DRPF. DRPF components were validated by the SMEs, whereby the group ascertained five highest risk factors: HbA1c, systole/diastole, blood glucose, and creatinine and blood urea nitrogen that were assigned by attribute weighting. Results from the questionnaire revealed an average agreement level of 80%. In conclusion, DRPF is implementable as prototype and has been highly accepted by Indonesian practitioners as aid for the diagnostics of diabetes.

ABSTRAK

Kekurangan diagnosis untuk penyakit kencing manis sering berlaku di sesetengah negara ASEAN dengan nisbah relatif dokter lebih kecil daripada pesakit. Pada asasnya, dipercayai bahawa rangka kerja sistematik untuk meramalkan faktor risiko penyakit kencing manis adalah penting untuk memperbaiki diagnostik dan meningkatkan ketepatan. Walau bagaimanapun, terdapat terbitan data bising yang dikesan sebagai data yang tidak lengkap dan masalah kelas terluar yang memberi kesan kepada kepekatan sampel. Rangka kerja sedia ada dianggap sukar untuk mengenalpasti faktor risiko kritikal penyakit kencing manis; sebahagiannya tidak tepat dan mengambil masa pengiraan yang besar. Tujuan kajian ini adalah untuk membangunkan rangka kerja yang sesuai untuk meramalkan risiko penyakit kencing manis. Keputusan ujian darah yang lengkap, rangka kerja dapat meramalkan dan mengklasifikasikan hasil sama ada dengan risiko penyakit kencing manis atau bebas risiko penyakit kencing manis. Rangka kerja Ramalan Risiko Diabetis (DRPF) telah dibangunkan dari kajian kepustakaan dan kajian kes dijalankan di tiga buah hospital swasta di Semarang. Analisis telah dijalankan untuk mencari komponen kerangka yang sesuai - kerana kekurangan perbandingan dan analisis mengenai gabungan algoritma pemilihan ciri dan klasifikasi. DRPF terdiri daripada empat bahagian utama: pra pemprosesan, pengesanan luar, pengurangan risiko, dan pembelajaran. Pra pemprosesan menyelesaikan masalah data yang hilang dan dengan itu menormalkan data. Analisis lanjut menggunakan k-mean clustering untuk mengesahkan kelas telah dijalankan. Komponen yang sesuai dipilih melalui perbandingan algoritma pengkelas dan pemilihan ciri. Pemilihan ciri berasaskan weighting attribute telah dipilih untuk menentukan berat badan. Faktor risiko bersesuaian digunakan pada dataset latihan untuk meningkatkan ketepatan dan masa pengiraan ramalan. Dalam bahagian pembelajaran, Mesin Sokongan Vektor dan Jaringan Neural Buatan telah dipilih sebagai algoritma klasifikasi yang sesuai, manakala Pokok Penyokong Kecerunan digunakan untuk menafsirkan peraturan berdasarkan klasifikasi kotak hitam. Menguji rangka kerja yang melibatkan Pima Indian Dataset sebagai dataset awam dan Hospital Dataset Semarang sebagai dataset swasta (800 data pesakit). Dalam mengesahkan DRPF, empat kajian kes menyiasat kumpulan Pakar Meta Subjek (SME) berdasarkan tahap perjanjian. Soal selidik terdiri daripada komponen DRPF, pelaksanaan DRPF, dan daya maju DRPF. Komponen DRPF telah disahkan oleh SME, di mana kumpulan tersebut menentukan lima faktor risiko tertinggi: HbA1c, systole/diastole, gula darah, dan creatin dan nitrogen urea darah yang diberikan oleh weighting atribut. Keputusan dari soal selidik menunjukkan tahap persetujuan purata sebanyak 80%. Kesimpulannya, DRPF boleh dilaksanakan sebagai prototaip dan telah diterima oleh pengamal perubatan Indonesia sebagai bantuan untuk mendiagnosis penyakit kencing manis.

ACKNOWLEDGEMENTS

First and foremost, I would wish to convey my deepest gratitude to my supervisor Professor Dr. Mohd. Khanapi Abd. Ghani for his immense support and patience in sharing knowledge as well as in providing the motivation and enthusiasm in completing the thesis.

My deepest appreciation is also extended to the evaluation committee for their encouragement, insightful commentaries, and rigorous review. My special thank goes to all my fellows who have offered help and support, especially to Dr. Fahmi Arif, Dr. Romi Satrio Wahono, Dr. Affandy, Dr. Ridzuan, Mohamed Doheir, Sri Winarno, and others.

Special thank is dedicated to The Minister of Finance of Indonesia (No. Kep56/LPDP/2014) for funding the research and publications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND GLOSSARY

ANN	*Artificial Neural Network*. A learning algorithm based on a loose analogy of how the human brain functions. Learning is achieved by adjusting the weights on the connections between nodes, which are analogous to synapses and neurons (Sammut & Webb 2010).
AUC	*Area under Curve*. An empirical measure of classification performance based on the area under an ROC curve. It evaluates the performance of a scoring classifier on a test set, but ignores the magnitude of the scores and only takes their rank order into account. AUC is expressed on a scale of 0 to 1, where 0 means that all negatives are ranked before all positives, and 1 means that all positives are ranked before all negatives (Sammut & Webb 2010).
BCP-NCD	*Benchmark Classification Algorithm Procedure in Non-Communicable Disease.*
BFP-NCD	*Benchmark Feature selection Procedure in Non-Communicable Disease.*
CDSS	*Clinical Decision Support System* is a system to make decisions with the help of available data (or information) and domain knowledge for unstructured and semi-structured problems (Nelson Ford 1985).
DM	*Diabetes Mellitus.* One of chronic disease or Non-Communicable Disease that cannot be cured directly. The blood glucose must to be maintained due to hyperglycaemia cannot process the certain of blood glucose level (American Diabetes Association 2011).
DRPF	*Diabetes Risk Prediction Framework.* This is a proposed framework for predicting diabetes risk, the input framework collects from complete blood test and the output diagnosed people with diabetes or no diabetes risk.
DT	*Decision Tree*. A tree-structured classification model, which is easy to understand, even by non-expert users, and can be efficiently induced from

data. The induction of decision trees is one of the oldest and most popular algorithms for learning discriminatory models, which has been developed independently in the statistical and machine learning communities (Sammut & Webb 2010).

ISI — *Institute for Scientific Information*. It was founded by Eugene Garfield in 1960 and then acquired by Thomson Scientific & Healthcare in 1992, which then became known as Thomson ISI. ISI offers bibliographic database services with its specialty on citation indexing and analysis. It maintains citation databases covering thousands of academic journals. This database allows a researcher to identify which articles have been cited most frequently, and who has cited them. The database not only provides an objective measure of the academic impact of the papers indexed in it, but also increases their impact by making them more visible and providing them with a quality label. The ISI also publishes the annual Journal Citation Reports which list an impact factor for each of the journals that it tracks.

k-nn — *k-nearest neighbour*. This algorithm represents on classification method, in which a new object is labelled based on its closest (k) neighbouring objects. In principle, given a training dataset (left) and a new object to be classified (right), the distance (referring to some kind of similarity) between the new object and the training objects is first computed, and the nearest (most similar) k objects are then chosen (Gorunescu 2011).

LDA — *Linear Discriminant Analysis*. A method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used as a linear classifier, or for dimensionality reduction before later classification. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempts to express one dependent variable as a linear combination of other features or measurements. However, ANOVA uses categorical independent variables and a continuous

dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable.

N/A *Not Available.*

NB *Naïve Bayes*. A simple learning algorithm that utilizes Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy. Coupled with its computational efficiency and many other desirable features, this leads to naïve Bayes being widely applied in practice (Sammut & Webb 2010).

R *Average Rank*. Average rank is used in the Friedman test, as a performance comparison measurement of the classification algorithms on each dataset. Friedman test has been suggested as preferable when comparing algorithms over several datasets and when normal distribution cannot be assumed (Berndtsson et al. 2008).

RC *Research Contribution*. Research contribution refers to a sharing of new ideas, inventions, theories and results with the rest of the world and expanding what is already known. These are shown as expansions to the world's body of knowledge. The definition of research itself is a considered activity, which aims to make an original contribution to knowledge (Dawson 2009).

RO *Research Objective*. Research objective identifies specific, measurable achievement that build towards the ultimate aim of a research. Research aims identify the highest level of the result to achieve a research. It is a broad statement of intent that identifies a research's purpose. Research objectives are more precise than research aims as they are quantitative and qualitative measures by which completion of the research will be judged (Dawson 2009).

ROC *Receiver Operating Characteristic*. ROC analysis investigates and employs the relationship between sensitivity and specificity of a binary classifier. Sensitivity or true positive rate measures the proportion of

positives correctly classified; specificity or true negative rate measures the proportion of negatives correctly classified (Sammut & Webb 2010).

RP *Research Problem*. A statement or description of the selected problem in a research. It has the same meaning of problem statement or problem description.

RQ *Research Question*. An open-ended opportunity to satisfy one's curiosity, they are often linked closely with one or more hypotheses. Our notion of research simply denotes a structured process for solving complex problems, formulated as research questions.

RS *Research Stage.* A stage of explaining the research, it starts from the literature review stage until the conclusion stage.

SLR *Systematic Literature Review*. A process of identifying, assessing, and interpreting all available research evidences with the purpose of providing answers for specific research questions. It is a tool that aims to produce a scientific summary of the evidences in a particular area, in contrast to traditional narrative review (Kitchenham et al. 2009)

SME *Subject Matter Expert*. Domain expert is a person who has a deep understanding of a particular process, function, technology, machine, material or type of equipment (Kawamoto et al. 2013).

SVM *Support Vector Machine*. A class of linear algorithms that can be used for classification, regression, density estimation, novelty detection, and other applications. In the simplest case of two-class classification, SVMs find a hyperplane that separate the two classes of data with as wide a margin as possible (Sammut & Webb 2010).

LIST OF PUBLICATIONS

Some of the research results represent the contributions of this thesis which have been published in the form of the following publications.

A. International Journal Articles

1. Sutanto, D. H, Herman, N. S, Ghani, M. K. A., 2014. The trend of Case-Based Reasoning for Chronic Disease Diagnosis: A Review. Advanced Science Letters, 20, pp. 1740 - 1744.

2. Sutanto, D. H, Ghani, M. K. A., 2015. A Benchmark of Classification Framework for Non-Communicable Disease Prediction. ARPN Journal of Engineering and Applied Sciences, 10 (16), pp. 9941 - 9955.

3. Sutanto, D. H, Ghani, M. K. A., 2015. A Benchmark Feature selection Framework for Non-Communicable Disease Prediction Model. Advanced Science Letters, 21, pp. 3409-3416.

4. Ghani, M. K. A., Sutanto, D. H., 2015. Improving Classification Accuracy For Non-Communicable Disease Prediction Model Based On Support Vector Machine. Jurnal Teknologi, 77 (18), pp. 29 - 36.

5. Sutanto, D. H, Ghani, M. K. A., 2015. Improving Classification Performance of K-Nearest Neighbour by Hybrid Clustering and Feature selection for Non-Communicable Disease Prediction. ARPN Journal of Engineering and Applied Sciences, 10 (16), pp. 6817 - 6825.