PROJECT REPORT

Illness Prediction from Oral Symptoms Using Machine Learning with Small Dataset

GABRIEL ASAEL TARIGAN
18.K1.0071

HALAMAN PENGESAHAN

Judul Tugas Akhir:	:	Illness Prediction from Oral Symptoms Using Machine Learning with Small Dataset
Diajukan oleh	:	Gabriel Asael Tarigan
NIM	:	18.K1.0071
Tanggal disetujui	:	25 Oktober 2022
Telah setujui oleh		
Pembimbing	:	R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D
Penguji 1	:	Yonathan Purbo Santosa S.Kom., M.Sc
Penguji 2	:	Hironimus Leong S.Kom., M.Kom.
Penguji 3	:	R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D
Penguji 4	:	Rosita Herawati S.T., M.I.T.
Penguji 5	:	Y.b. Dwi Setianto S.T., M.Cs.
Penguji 6	:	Yulianto Tejo Putranto S.T., M.T.
Ketua Program Studi	:	Rosita Herawati S.T., M.I.T.
Dekan	:	Dr. Bernardinus Harnadi S.T., M.T.

Halaman ini merupakan halaman yang sah dan dapat diverifikasi melalui alamat di bawah ini.

sintak.unika.ac.id/skripsi/verifikasi/?id=18.K1.0071

DECLARATION OF AUTHORSHIP

I, the undersigned:

Name : Gabriel Asael Tarigan

ID : 18K10071

declare that this work, titled " Illness Prediction from Oral Symptoms Using Machine Learning with Small Dataset", and the work presented in it is my own. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at Soegijapranata Catholic University
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given.
5. Except for such quotations, this work is entirely my own work.
6. I have acknowledged all main sources of help.
7. Where the work is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Semarang, October, 27, 2022

Gabriel Asael Tarigan

18.K1.0071

DECLARATION OF RESEARCH PUBLICATION
FOR ACADEMIC PURPOSE

I, the undersigned:

Name : Gabriel Asael Tarigan

Major : Informatics Engineering

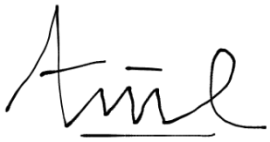
Faculty : Computer Science

Type of Work : Final Project

Agreed to give Soegijapranata Catholic University Non-exclusive Free Royalty Rights over scientific work titled "Illness Prediction from Oral Symptoms Using Machine Learning with Small Dataset". With Non-exclusive Free Royalty Rights, Soegijapranata Catholic University is allowed to save, move/make change, manage in the form of database, maintain, and publish this final project as long as my name is included as the writer/owner of the copyright.

Semarang, October, 27, 2022

Yang menyatakan

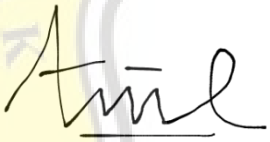


Gabriel Asael Tarigan

ACKNOWLEDGMENT

First and foremost, praises and thanks to God, the Almighty, for His blessings throughout this project titled “Illness Prediction from Oral Symptoms Using Machine Learning with Small Dataset” work to complete the project successfully. I would also like to thank my advisor R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D. for all the guidance I gave, which helped me significantly in the project completion which I highly value. I would also like to thank my parents for their support throughout my university journey. Also, Tiara Putri Dewanti not only gave me emotional support when absolutely needed but also helped by giving me valuable insights and helping in finishing this project. This project is done as a requirement to complete my bachelor's degree.

Semarang, Oktober, 27 2022



Gabriel Asael Tarigan

ABSTRACT

Dental check-ups can be quite a time-consuming process and the cost of a simple check-up can also be expensive for people in the lower economic spectrum. Currently, the most common way of a dental check-up is coming straight to the dentist and asking about what illness the patient is possibly dealing with, and that without prior knowledge of the symptoms, patients must deal with expenses and time without even needing one when the illness is mild and can be mended directly. Currently, there is no dataset of oral symptoms available publicly, and there are only a handful of tools to see the sort of illness arising from symptoms. This research aims to give a probable explanation of what the user might be dealing with and the severity of the illness by giving questions that revolve around the area of the mouth which asks about what particular symptoms the user currently has, after the list of symptoms is created, it will be processed through machine learning algorithms, in this particular project the main algorithm used is Extreme Gradient Boost, also known as XGBoost, although other algorithms are also as comparisons. Other algorithms used are random forest and TensorFlow Keras, specifically multilayer perceptrons which are used for multi-class classification. XGBoost shows that it has better performance when dealing with this issue.

***Keywords**: dental, symptoms, XGBoost, Random forest, TensorFlow MLP*

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 14

1.1. Background 14

1.2. Problem Formulation 15

1.3. Scope 15

1.4. Objective 15

CHAPTER 2 LITERATURE STUDY 16

2.1. Introduction 16

2.2. Approach, limitation, and application of AI in dentistry 16

2.3. Oral symptoms and illnesses 18

CHAPTER 3 RESEARCH METHODOLOGY 20

3.1. Dataset creation 20

3.1.1. Data gathering 20

3.1.2. Data validation and annotation 24

3.2. Preparation and introduction for each algorithm 25

CHAPTER 4 ANALYSIS AND DESIGN 28

4.1. Analysis 28

4.2. Design 29

4.3. Function 32

CHAPTER 5 IMPLEMENTATION AND RESULTS 38

5.1. Implementation 38

5.2. Results 43

CHAPTER 6 CONCLUSION 96

LIST OF FIGURES

Figure 3.1 Array of Encoded Labels 26

Figure 4.1 Example of symptoms choice 28

Figure 4.1 User input for the program 29

Figure 4.2 Flowchart of XGBoost 30

Figure 4.3 An example of a decision tree for classification 31

Figure 4.4 Multilayer perceptrons with one output layer 32

Figure 4.1 XGBoost classification example 32

Figure 4.2 Random forest classification example 33

Figure 4.3 Random forest classification example 34

Figure 4.4 How function (7) looks like 35

Figure 4.5 Example of LTU 35

Figure 4.6 Weight, input, and bias for each node 36

Figure 4.7 A more detailed example of multilayer perceptrons 36

Figure 4.8 Matrix of weight 37

Figure 5.1 Example of a multi-class confusion matrix 44

Figure 5.2 Precision on confusion matrix 44

Figure 5.3 Recall on confusion matrix 45

Figure 5.4 F1-score on confusion matrix 45

Figure 5.5 Confusion matrix of XGBoost, run 1 47

Figure 5.6 Confusion matrix of XGBoost, run 2 47

Figure 5.7 Confusion matrix of XGBoost, run 3 48

Figure 5.8 Confusion matrix of XGBoost, run 4 48

Figure 5.9 Confusion matrix of XGBoost, run 5 49

Figure 5.10 Confusion matrix of XGBoost, run 6 49

Figure 5.11 Confusion matrix of XGBoost, run 7 50
Figure 5.12 Confusion matrix of XGBoost, run 8 50
Figure 5.13 Confusion matrix of XGBoost, run 9 51
Figure 5.14 Confusion matrix of XGBoost, run 10 51
Figure 5.15 Confusion matrix of random forest, run 1 53
Figure 5.16 Confusion matrix of random forest, run 2 53
Figure 5.17 Confusion matrix of random forest, run 3 54
Figure 5.18 Confusion matrix of random forest, run 4 54
Figure 5.19 Confusion matrix of random forest, run 5 55
Figure 5.20 Confusion matrix of random forest, run 6 55
Figure 5.21 Confusion matrix of random forest, run 7 56
Figure 5.22 Confusion matrix of random forest, run 8 56
Figure 5.23 Confusion matrix of random forest, run 9 57
Figure 5.24 Confusion matrix of random forest, run 10 57
Figure 5.25 Confusion matrix of TensorFlow's multilayer perceptrons, run 1 59
Figure 5.26 Confusion matrix of TensorFlow's multilayer perceptrons, run 2 59
Figure 5.27 Confusion matrix of TensorFlow's multilayer perceptrons, run 3 60
Figure 5.28 Confusion matrix of TensorFlow's multilayer perceptrons, run 4 60
Figure 5.29 Confusion matrix of TensorFlow's multilayer perceptrons, run 5 61
Figure 5.30 Confusion matrix of TensorFlow's multilayer perceptrons, run 6 61
Figure 5.31 Confusion matrix of TensorFlow's multilayer perceptrons, run 7 62
Figure 5.32 Confusion matrix of TensorFlow's multilayer perceptrons, run 8 62
Figure 5.33 Confusion matrix of TensorFlow's multilayer perceptrons, run 9 63
Figure 5.34 Confusion matrix of TensorFlow's multilayer perceptrons, run 10 63
Figure 5.35 Bayes optimal error, the limit of the highest acceptable machine's accuracy 65

Figure 5.36 An example from the XGBoost's confusion matrix 66
Figure 5.37 An example from the random forest's confusion matrix 66
Figure 5.38 An example from TensorFlow's multilayer perceptrons' confusion matrix 67
Figure 5.1 Prediction of XGBoost for diabetes symptoms 68
Figure 5.2 Prediction of random forest for diabetes symptoms 68
Figure 5.3 Prediction of TensorFlow's multilayer perceptrons for diabetes symptoms 68
Figure 5.4 Prediction of XGBoost for oral cancer symptoms 69
Figure 5.5 Prediction of random forest for oral cancer symptoms 69
Figure 5.6 Prediction of TensorFlow's multilayer perceptrons for oral cancer symptoms 69
Figure 5.7 Prediction of XGBoost for kidney failure symptoms 69
Figure 5.8 Prediction of random forest for kidney failure symptoms 69
Figure 5.9 Prediction of TensorFlow's multilayer perceptrons for kidney failure symptoms 69
Figure 5.10 Prediction of XGBoost for bulimia symptoms 70
Figure 5.11 Prediction of random forest for bulimia symptoms 70
Figure 5.12 Prediction of random forest for kidney failure symptoms 70
Figure 5.13 Prediction of XGBoost with input 1, 1, 1, 1, 1, 1 71
Figure 5.14 Prediction of random forest with input 1, 1, 1, 1, 1, 1 71
Figure 5.15 Prediction of TensorFlow's multilayer perceptrons with input 1, 1, 1, 1, 1, 1 71
Figure 5.16 Prediction of XGBoost with input 10, 10, 10, 10, 10, 10 71
Figure 5.17 Prediction of random forest with input 10, 10, 10, 10, 10, 10 71
Figure 5.18 Prediction of TensorFlow's multilayer perceptrons with input 10, 10, 10, 10, 10, 10 71
Figure 5.19 Prediction confidence of XGBoost with data consisting of only more agreeable entries 72
Figure 5.20 Confusion matrix of XGBoost with data consisting of only more agreeable entries 72
Figure 5.21 Prediction confidence of random forest with data consisting of only more agreeable entries 73

Figure 5.22 Confusion matrix of random forest with data consisting of only more agreeable entries 74

Figure 5.23 Prediction confidence of TensorFlow's multilayer perceptrons with data consisting of only more agreeable entries 75

Figure 5.24 Confusion matrix of TensorFlow's multilayer perceptrons with data consisting of only more agreeable entries75

Figure 5.25 Prediction of XGBoost for diabetes symptoms 76

Figure 5.26 Confusion matrix of XGBoost for diabetes symptoms with 50:50 training and test split . 77

Figure 5.27 Prediction of random forest for diabetes symptoms 50:50 training and test split 78

Figure 5.28 Confusion matrix of random forest for diabetes symptoms 50:50 training and test split .. 78

Figure 5.29 Prediction of TensorFlow's multilayer perceptrons for diabetes symptoms 50:50 training and test split 79

Figure 5.30 Confusion matrix of TensorFlow's multilayer perceptrons for diabetes symptoms 50:50 training and test split 80

Figure 5.31 Prediction of XGBoost for diabetes symptoms 70:30 training and test split 81

Figure 5.32 Confusion matrix of XGBoost for diabetes symptoms 70:30 training and test split 82

Figure 5.33 Prediction of random forest for diabetes symptoms 70:30 training and test split 82

Figure 5.34 Prediction of random forest for diabetes symptoms 70:30 training and test split 83

Figure 5.35 Prediction of TensorFlow's multilayer perceptrons for diabetes symptoms 70:30 training and test split 84

Figure 5.36 Confusion matrix of TensorFlow's multilayer perceptrons symptoms 70:30 training and test split 84

Figure 5.37 Prediction of XGBoost for diabetes symptoms 20:80 training and test split 86

Figure 5.38 Confusion matrix of XGBoost for diabetes symptoms 20:80 training and test split 86

Figure 5.39 Prediction of random forest for diabetes symptoms 20:80 training and test split 87

Figure 5.40 Prediction of random forest for diabetes symptoms 20:80 training and test split 88

Figure 5.41 Prediction of TensorFlow multilayer perceptrons for diabetes symptoms 20:80 training and test split 89

Figure 5.42 Prediction of TensorFlow multilayer perceptrons for diabetes symptoms 20:80 training and test split 89

Figure 5.43 Prediction of random forest for diabetes symptoms 20:80 training and test split, 100 trees 90

Figure 5.44 Prediction of random forest for diabetes symptoms 20:80 training and test split, 10 trees 90

Figure 5.45 Prediction of random forest for diabetes symptoms 20:80 training and test split, 25 trees 91

Figure 5.46 Prediction of random forest for diabetes symptoms 20:80 training and test split, 50 trees 91

Figure 5.47 Prediction of TensorFlow multilayer perceptron with three layers 91

Figure 5.48 Loss curve of TensorFlow multilayer perceptron with three layers 92

Figure 5.49 Prediction of TensorFlow multilayer perceptron with one additional dense layer consisting of 16 nodes with ReLU activation function 92

Figure 5.50 Loss curve of TensorFlow multilayer perceptron with one additional dense layer consisting of 16 nodes with ReLU activation function 92

Figure 5.51 Prediction of TensorFlow multilayer perceptron with two additional dense layers consisting of 16 nodes with ReLU activation function 92

Figure 5.52 Loss curve of TensorFlow multilayer perceptron with two additional dense layers consisting of 16 nodes with ReLU activation function 93

Figure 5.53 Prediction of TensorFlow multilayer perceptron with three additional dense layers consisting of 16 nodes with ReLU activation function 93

Figure 5.54 Loss curve of TensorFlow multilayer perceptron with three additional dense layers consisting of 16 nodes with ReLU activation function 93

Figure 5.55 Loss curve of XGBoost 95

Figure 5.56 Loss curve of TensorFlow multilayer perceptrons 95

LIST OF TABLES

Table 3.1. Gums (*gusi*) list encoding (left: encoded as numbers, right: symptoms) 22

Table 3.2. Teeth (*gigi*) list encoding (left: encoded as numbers, right: symptoms) 22

Table 3.3. Lips (*bibir*) list encoding (left: encoded as numbers, right: symptoms)..... 22

Table 3.4. Tongue (*lidah*) list encoding (left: encoded as numbers, right: symptoms) 22

Table 3.5. Throat (*tenggorokan*) list encoding (left: encoded as numbers, right: symptoms)..... 23

Table 3.6. Area of mouth (*area mulut*) list encoding (left: encoded as numbers, right: symptoms 23

Table 3.7. Example of encoded symptoms for each feature, with labels..... 23

Table 3.1. Original entry..... 24

Table 3.2. Annotator's entry 24

Table 3.3. Example of the original dataset with diff and score column 24

Table 5.1. Diabetes and oral cancer symtomps..... 67

Table 5.2. Kidney failure symptoms..... 68