



PROJECT REPORT
Diabetes Prediction Using Decision Tree and
XGBoost Algorithm

MICHAEL WIRYASEPUTRA
19.k1.0018

Faculty of Computer Science
Soegijapranata Catholic University
2021



HALAMAN PENGESAHAN

Judul Tugas Akhir: : Diabetes Prediction Using Decision Tree and XGBoost Algorithm

Diajukan oleh : Michael Wiryaseputra

NIM : 19.K1.0018

Tanggal disetujui : 21 Desember 2022

Telah setuju oleh

Pembimbing : Y.b. Dwi Setianto S.T., M.Cs.

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=19.K1.0018

DECLARATION OF AUTHORSHIP

I, the undersigned:

Name : Michael Wiryaseputra

ID : 19.K1.0018

declare that this work, titled "Diabetes Prediction Using Decision Tree and XGBoost Algorithm", 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, January, 02, 2023



Michael Wiryaseputra

19.K1.001

HALAMAN PERNYATAAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan dibawah ini:

Nama : Michael Wiryaseputra

Program Studi : Teknik Informatika

Fakultas : Ilmu Komputer

Jenis Karya : Skripsi

Menyetujui untuk memberikan kepada Universitas Katolik Soegijapranata Semarang Hak Bebas Royalti Noneksklusif atas karya ilmiah yang berjudul "Diabetes Prediction Using Decision Tree and XGBoost Algorithm". Dengan Hak Bebas Royalti Noneksklusif ini Universitas Katolik Soegijapranata berhak menyimpan, mengalihkan media/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan tugas akhir ini selama tetap mencantumkan nama saya sebagai penulis / pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Semarang, 03 Januari 2023

Yang menyatakan

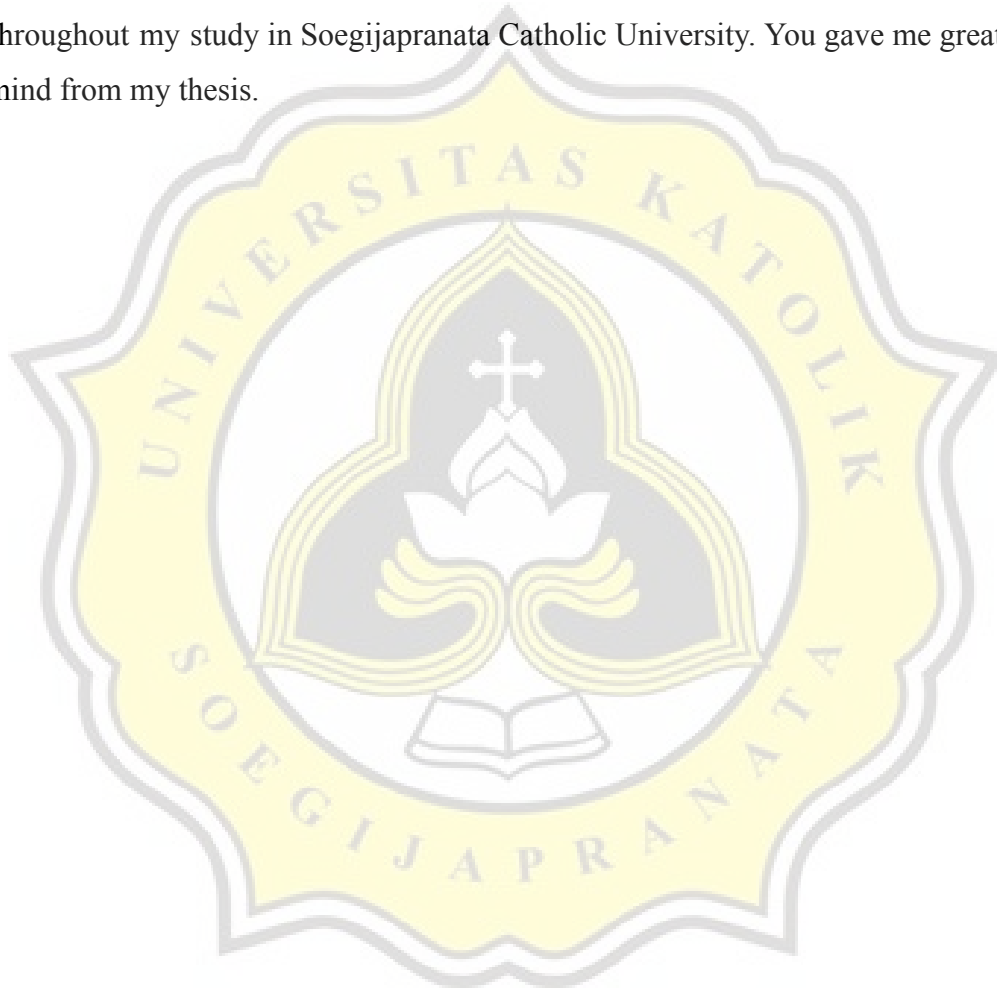


Michael Wiryaseputra
19.K1.0018

ACKNOWLEDGMENT

I have received a myriad of support, advice, and assistance throughout this document writing. I would like to thank my supervisors Y.B.Dwi Setianto for formulating this topic. I would also like to thank my friend and my family for guiding with advice to finish this document.

I would like to thank my family and friends for giving me ceaseless love, support, and advices throughout my study in Soegijapranata Catholic University. You gave me great escape to rest my mind from my thesis.



ABSTRACT (ABSTRACT TITLE)

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke.

First thing we have to do to predict the diabetes, we need to chose the good dataset for machine learning modelling. In tis case we will use pima indian diabetes as dataset. After that, test the dataset using Decision Tree and XGBoost algorithm.

The result of this project is to determine how accurate the system is in processing datasets, as well as comparing algorithms between Decision Tree with XGBoost.

Keyword: prediction, Decision Tree, XGBoost

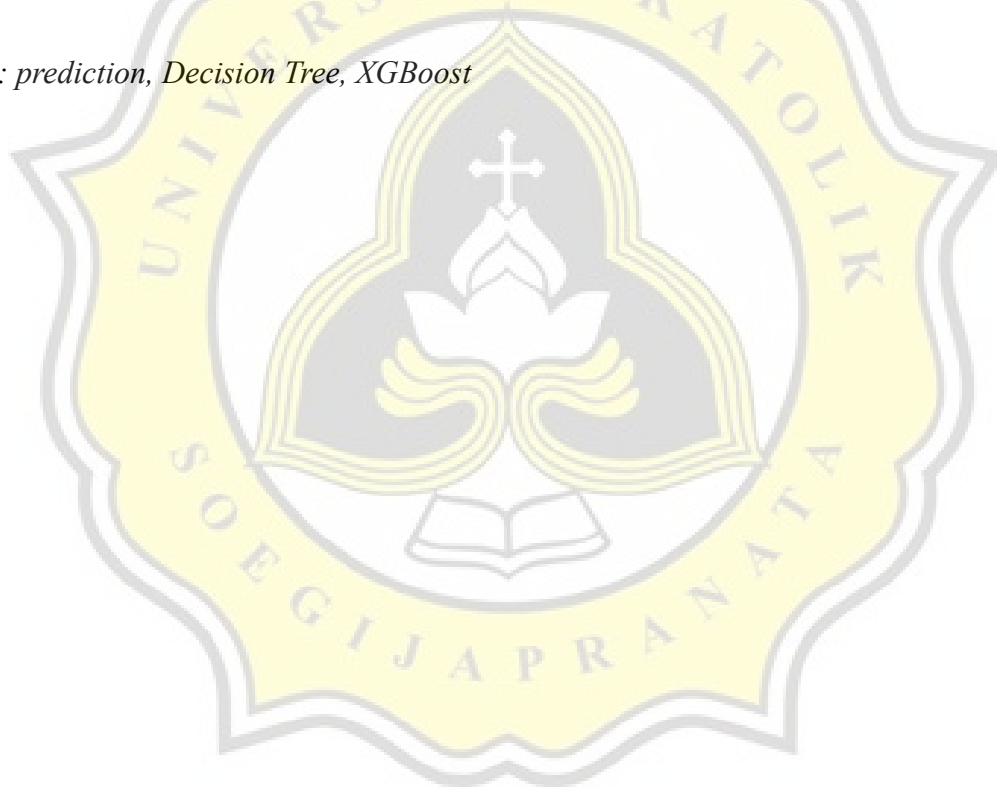


TABLE OF CONTENTS

COVER	i
APPROVAL AND RATIFICATION PAGE (Heading plain)	ii
DECLARATION OF AUTHORSHIP	iii
ACKNOWLEDGMENT	iv
ABSTRACT (Abstract Title)	v
TABLE OF CONTENTS	vi
LIST OF FIGURE	viii
LIST OF TABLE	ix
CHAPTER 1 INTRODUCTION	11
1.1. Background	11
1.2. Problem Formulation	11
1.3. Scope	11
1.4. Objective	12
CHAPTER 2 LITERATURE STUDY	13
CHAPTER 3 RESEARCH METHODOLOGY	19
CHAPTER 4 ANALYSIS AND DESIGN	24
CHAPTER 5 IMPLEMENTATION AND TESTING	26
5.1. Implementation	26
5.2. Testing	33
CHAPTER 6 CONCLUSION	46
REFERENCES	47
APPENDIX	a

LIST OF FIGURE

Figure 3.1 Decision Tree Architecture	21
Figure 4.1 Flowchart of Diabetes Prediction using Decision Tree and XGBoost	24
Figure 5.1 Import some library	26
Figure 5.2 Import the dataset	26
Figure 5.3 Show all column in Dataset	27
Figure 5.4 Show dataset info	27
Figure 5.5 Show dataset first 10 column	27
Figure 5.6 Show the sum of null data before changing the zero value	28
Figure 5.7 Changing zero value into null value	29
Figure 5.8 Null value from 5 attributes mentioned above	29
Figure 5.9 Fill null value with median	30
Figure 5.10 Checking on unbalanced data	30
Figure 5.11 Oversampling method	30
Figure 5.12 Using Random state = 42	31
Figure 5.13 Showing Outlier in each attributes using boxplot	31
Figure 5.14 Outlier Detection using boxplot	32
Figure 5.15 Applying Z-Score to remove the rows with outlier	32
Figure 5.16 Showing the new dataframe after cleaned with Z-Score	33
Figure 5.17 Showing all of the data with outlier	33
Figure 5.18 Showing correlation matrix	33
Figure 5.19 Show column relevances using Correlation Matrix Heatmap	34
Figure 5.20 Splitting X and y. X is the attributes, y is the result output	34

Figure 5.21 Decision Tree Model Evaluation with Random State = 0, Balanced Dataset	35
Figure 5.22 XGBoost Model Evaluation with Random State = 0, Balanced Dataset	36
Figure 5.23 Decision Tree Model Evaluation with Random State = 42, Balanced Dataset	37
Figure 5.24 XGBoost Model Evaluation with Random State = 42, Balanced Dataset	38
Figure 5.25 Decision Tree Model Evaluation with Random State = 0, Imbalanced Dataset	39
Figure 5.26 XGBoost Model Evaluation with Random State = 0, Imbalanced Dataset	40
Figure 5.27 Decision Tree Model Evaluation with Random State = 42, Imbalanced Dataset	41
Figure 5.28 XGBoost Model Evaluation with Random State = 42, Imbalanced Dataset	42
Figure 5.29 Decision Tree Model Evaluation with Random State = 0, Imbalanced Dataset without Age Attributes	43
Figure 5.30 XGBoost Model Evaluation with Random State = 0, Imbalanced Dataset without Age Attributes	44
Figure 5.31 Decision Tree Model Evaluation with Random State = 42, Imbalanced Dataset without Age Attributes	45
Figure 5.32 XGBoost Model Evaluation with Random State = 42, Imbalanced Dataset without Age Attributes	46

LIST OF TABLE

Table 5.1. Model Evaluation with Random State = 0, Balanced Dataset	9
Table 5.2. Model Evaluation with Random State = 42, Balanced Dataset	9
Table 5.3. Model Evaluation with Random State = 0, Imbalanced Dataset	9

Table 5.4. Model Evaluation with Random State = 42, Imbalanced Dataset	9
Table 5.5. Model Evaluation with Random State = 0, Imbalanced Dataset without Age Attributes	9
Table 5.6. Model Evaluation with Random State = 42, Imbalanced Dataset without Age Attributes	9

