



PROJECT REPORT
AN EVALUATION OF HEART DISEASE PREDICTION
USING AN EXTREME GRADIENT BOOSTING
ALGORITHM

DAVIN CHANG
19.K1.0005

Faculty of Computer Science
Soegijapranata Catholic University
2022

HALAMAN PENGESAHAN



Judul Tugas Akhir: : An Evaluation Of Heart Disease Prediction Using An Extreme Gradient Boosting Algorithm

Diajukan oleh : Davin Chang

NIM : 19.K1.0005

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.K1.0005

DECLARATION OF AUTHORSHIP

I, the undersigned:

Name : Davin Chang

ID : 19.K1.0005

declare that this work, titled “An Evaluation Of Heart Disease Prediction Using An Extreme Gradient Boosting 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, December, 29, 2022



DAVIN CHANG

19.K1.0005

**HALAMAN PERNYATAAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan dibawah ini:

Nama : Davin Chang
Program Studi : Teknik Informatika
Fakultas : Ilmu Komputer
Jenis Karya : Skripsi

Menyetujui untuk memberikan kepada Universitas Katolik Soegijapranata Semarang Hak Bebas Royalti Noneklusif atas karya ilmiah yang berjudul “An Evaluation Of Heart Disease Prediction Using An Extreme Gradient Boosting Algorithm”. Dengan Hak Bebas Royalti Noneklusif 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, 29 Desember 2022

Yang menyatakan



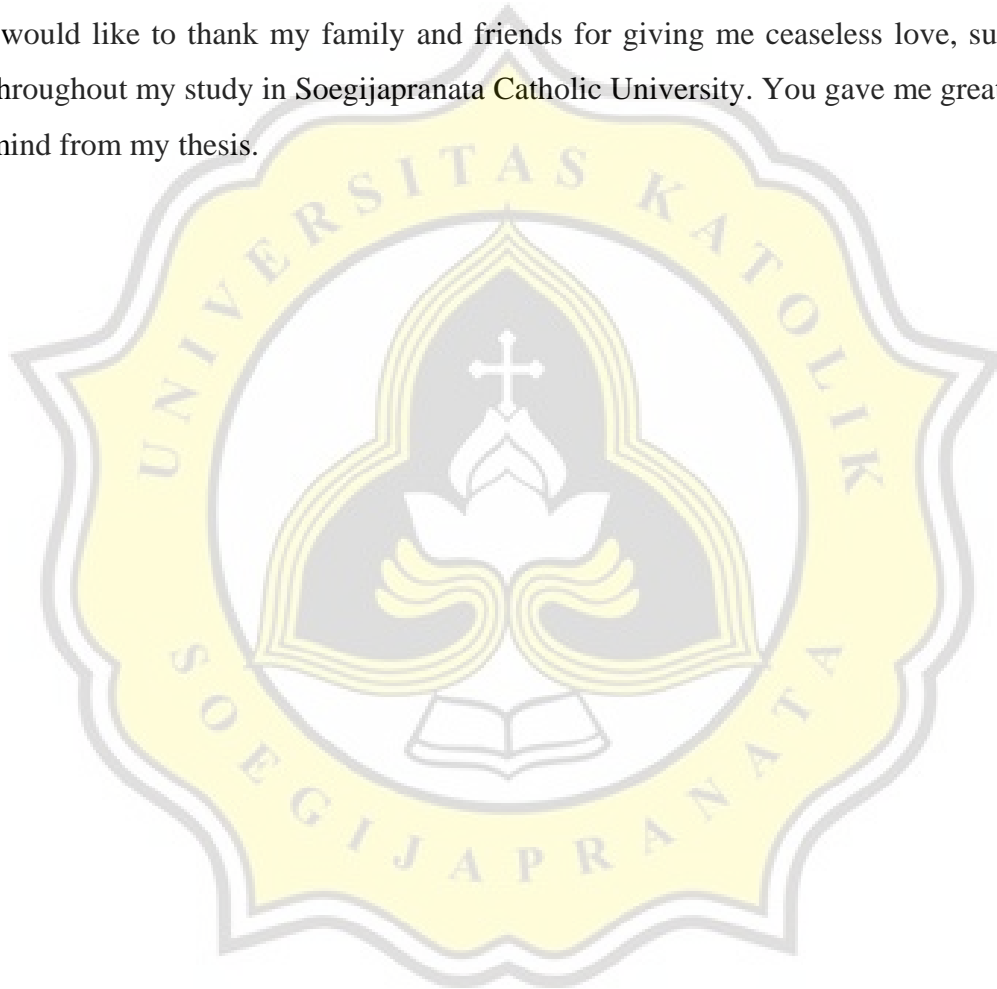
DAVIN CHANG

19.K1.0005

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 S.T., M.Cs. for formulating this topic. I would also like to thank my friend Samuel Kurniawan Santoso, Jevon Carla, Andre Nugroho Pranoto, and Tan, Yudistira Suryakencana Adisatria 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

Heart disease has recorded the most death cause in the world. A lot of researchers are trying to find better and more reliable machine learning to diagnose heart disease. Accuracy and the speed of computation become the main concern when classifying heart disease at its early stages related to human life.

This paper researched about Extreme Gradient Boosting (XGBoost) as an ensemble learning with boosting method to predict heart disease. The data will be preprocessed using handling missing value and removing outliers. The algorithm will be compared with 2 different datasets (with feature selection and without feature selection).

The outcome of this research hopefully can present the performance result of the Extreme Gradient Boosting algorithm using tenfold cross-validation and performance measures (Precision, Recall, F1-score, ROC Area, and Accuracy) when using feature selection and without using feature selection.

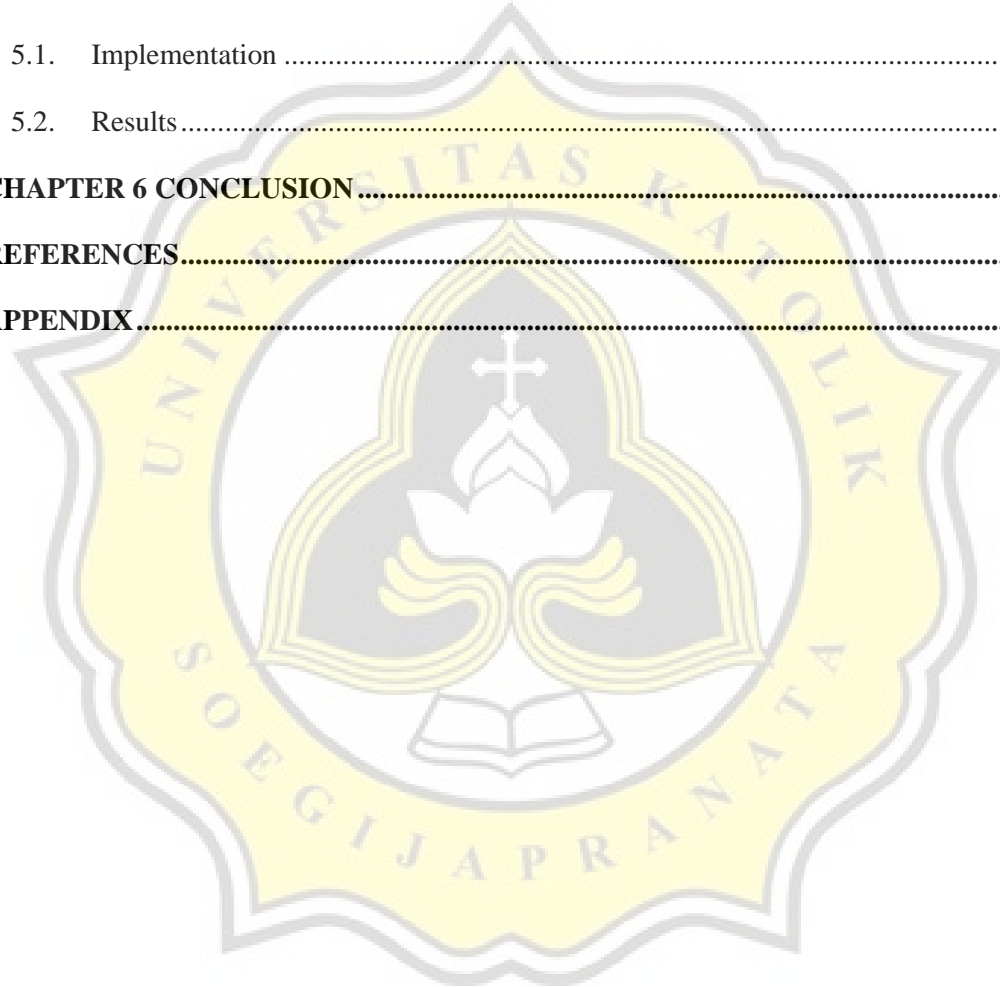
Keywords: heart disease, machine learning, ensemble learning, xgboost, feature selection



TABLE OF CONTENTS

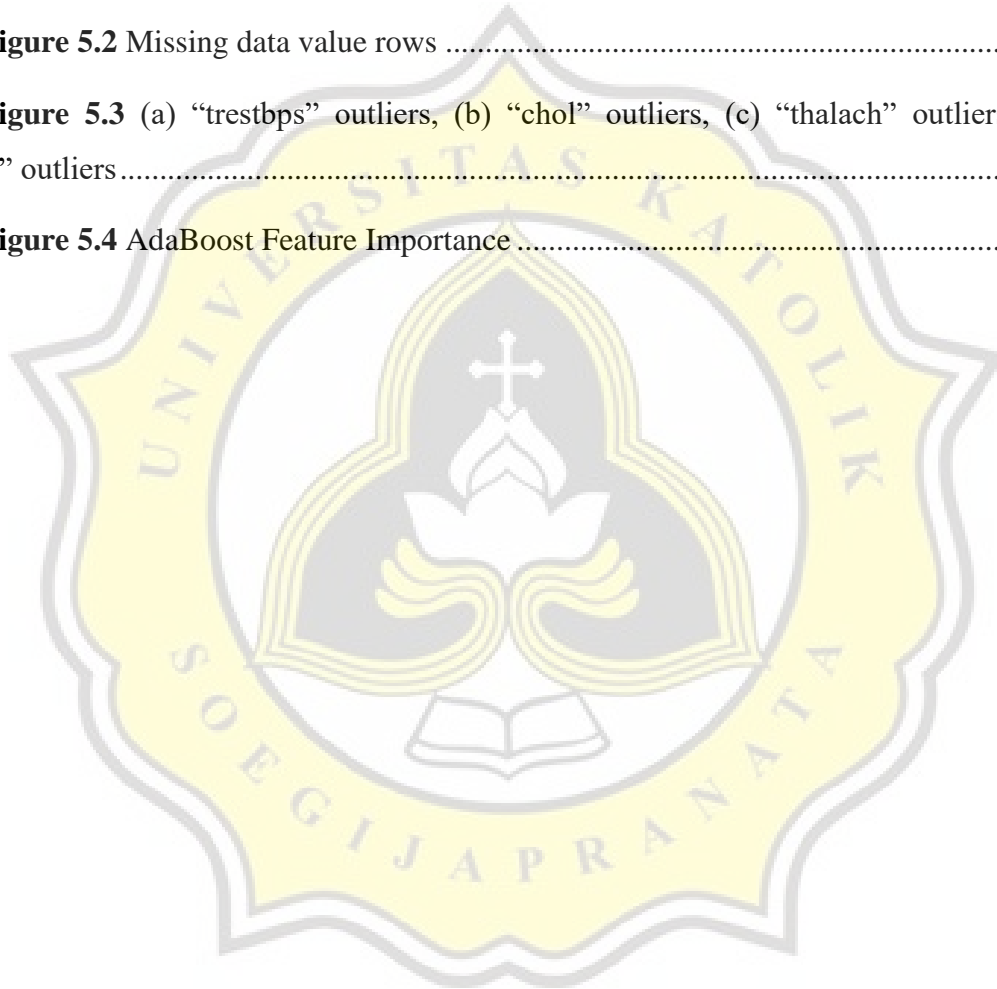
COVER	i
APPROVAL AND RATIFICATION PAGE	Error! Bookmark not defined.
DECLARATION OF AUTHORSHIP	iii
HALAMAN PERNYATAAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS	Error! Bookmark not defined.
ACKNOWLEDGMENT	v
TABLE OF CONTENTS	vii
LIST OF FIGURE	ix
LIST OF TABLE	x
CHAPTER 1 INTRODUCTION	12
1.1. Background	12
1.2. Problem Formulation	13
1.3. Scope	13
1.4. Objective	13
CHAPTER 2 LITERATURE STUDY	14
CHAPTER 3 RESEARCH METHODOLOGY	18
3.1. Overview	18
3.2. Data Collection.....	19
3.3. Data Portioning	20
3.4. Data Preprocessing.....	20
3.5. Bayesian Optimization	20
3.6. Extreme Gradient Boosting Algorithm	21
3.7. Feature Selection.....	22
3.8. Model Evaluation	22
CHAPTER 4 ANALYSIS AND DESIGN	23

4.1. Analysis.....	23
4.1.1. Handling Missing Values	23
4.1.2. Feature Selection	24
4.1.3. Model Evaluation	24
4.2. Design	24
CHAPTER 5 IMPLEMENTATION AND RESULTS	26
5.1. Implementation	26
5.2. Results	31
CHAPTER 6 CONCLUSION	49
REFERENCES.....	50
APPENDIX.....	a



LIST OF FIGURE

Figure 3.1 Research Methodology	18
Figure 3.2 Preprocessed Target Attribute	20
Figure 4.1 Handling Missing Values Flowchart	23
Figure 5.1 Heart Disease target column	26
Figure 5.2 Missing data value rows	27
Figure 5.3 (a) “trestbps” outliers, (b) “chol” outliers, (c) “thalach” outliers, and (d) “oldpeak” outliers.....	33
Figure 5.4 AdaBoost Feature Importance	42



LIST OF TABLE

Table 3.1. Data attributes and description	19
Table 3.2. Hyperparameter Search Space.....	22
Table 4.1. XGBoost Hyperparameter Search Space	25
Table 4.2. AdaBoost Optimal Parameter.....	25
Table 5.1. Handling Missing Values Evaluation.....	31
Table 5.2. Median and Mode of “ca” and “thal”	32
Table 5.3. Outliers Evaluation.....	33
Table 5.4. Feature Selection Rankings with XGBoost.....	34
Table 5.5. Chi Square Feature Selection with XGBoost.....	35
Table 5.6. Mutual Information Feature Selection with XGBoost.....	35
Table 5.7. ANOVA Feature Selection with XGBoost	36
Table 5.8. Forward Feature Selection with XGBoost	37
Table 5.9. Backward Feature Selection with XGBoost	37
Table 5.10. Recursive Feature Elimination with XGBoost.....	38
Table 5.11. Feature Importance Feature Selection with XGBoost	39
Table 5.12. Feature Selection Evaluation Comparison with XGBoost.....	40
Table 5.13. Feature Selection XGBoost Comparison	40
Table 5.14. Feature Selection Rankings with AdaBoost.....	41
Table 5.15. Feature Importance Rankings with AdaBoost	41
Table 5.16. Chi Square Feature Selection with AdaBoost.....	42
Table 5.17. Mutual Information Feature Selection with AdaBoost	43
Table 5.18. ANOVA Feature Selection with AdaBoost	44
Table 5.19. Forward Feature Selection with AdaBoost	44

Table 5.20. Backward Feature Selection with AdaBoost.....	45
Table 5.21. Recursive Feature Elimination with AdaBoost.....	46
Table 5.22. Feature Importance Feature Selection with AdaBoost	46
Table 5.23. Feature Selection Evaluation Comparison with AdaBoost.....	47
Table 5.24. Model with Feature Selection Comparison.....	48
Table 5.25. Model without Feature Selection Comparison.....	48

