



PROJECT REPORT

Machine Learning to Predict the onset of Diabetes Disease based on Diagnostic Measures using Two Different Libraries and Algorithms.

**EWEN LIAM
17.K1.0052**

**Faculty of Computer Science
Soegijapranata Catholic University
2021**

APPROVAL AND RATIFICATION PAGE



Judul Tugas Akhir : Machine Learning to Predict the onset of
Diabetes Disease based on
Diagnostic Measures using Two Different
Libraries and Algorithms.

Diajukan oleh : Ewen Liam

NIM : 17.K1.0052

Tanggal disetujui : 07 Juli 2021

Telah setuju oleh

Pembimbing : R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D

Penguji 1 : R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D

Penguji 2 : Rosita Herawati S.T., M.I.T.

Penguji 3 : Hironimus Leong S.Kom., M.Kom.

Penguji 4 : Y.b. Dwi Setianto

Penguji 5 : Yulianto Tejo Putranto S.T., M.T.

Ketua Program Studi : Rosita Herawati S.T., M.I.T.

Dekan : R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D

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

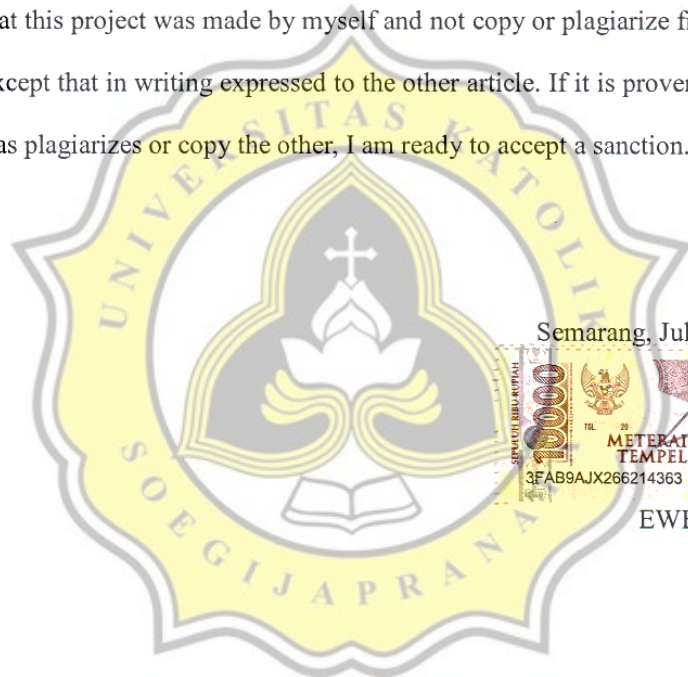
STATEMENT OF ORIGINALITY

I, the undersigned:

Name : EWEN LIAM

ID : 17.K1.0052

Certify that this project was made by myself and not copy or plagiarize from other people, except that in writing expressed to the other article. If it is proven that this project was plagiarizes or copy the other, I am ready to accept a sanction.



Semarang, July 7, 2021



EWEN LIAM

**APPROVAL PAGE FOR PUBLICATION OF SCIENTIFIC PAPERS
FOR ACADEMIC INTEREST**

The undersigned below:

Name : Ewen Liam
Undergraduate Program : INFORMATICS ENGINEERING
Faculty : COMPUTER SCIENCE
Type of work : THESIS

Approved to give Non-Exclusive Royalty Free Right to Soegijapranata Catholic University Semarang for scientific work entitled "Machine Learning to Predict the onset of Diabetes Disease based on Diagnostic Measures using Two Different Libraries and Algorithms." along with the existing tools (if needed). With this Non-Exclusive Royalty Free Right Soegijapranata Catholic University has the right to store, transfer data / format, manage in the form of database, maintain and publish this final project as long as I keep my name as a writer / creator and as a Copyright owner.

This statement I made in truth

Semarang, July 7, 2021

Sincerely



EWEN LIAM

ACKNOWLEDGEMENTS

First of all, all glory belongs to Jesus that allows me to make this thesis and always strengthens me in the making of this thesis. This thesis is dedicated to faith in humanity as the motto of Soegijapranata Catholic University that goes "Talenta Pro Patria at Humanitate". I hope this thesis could be useful for the development of machine learning implementation in medicine and could help people that struggle with diabetes diseases. This thesis also for the completion of the requirement to finish a Bachelor Degree in Informatics Engineering, Soegijapranata Catholic University.

This thesis can be successfully done also with the support of my surroundings, special thanks to:

1. All of the lectures, that always acknowledge me and make a time to discuss this problem with me, especially my supervising lecturer R. Setiawan Aji Nugroho S.T., M CompIT., Ph.D.
2. My dearest family, Law Mie Tjoe, Oei Sien Ling, Gagarin, Bethea, and Hansen. For always supporting me with materials and love.
3. All of my dearest friends (Abigail Anita, Stevanus, Aditya Morning Star, David, Lucas, Dinar) for always be there and encourage me when things get hard. And all of my friends that I cannot mention.
4. Last but not least, I want to thank me, I want to thank me for believing in me, I want to thank me for doing all this hard work, I want to thank me for having no days off, I want to thank me for never quitting, I want to thank me for always being a giver, and trying to give more than I receive, I want to thank me for trying to do more right than wrong, I want to thank me for just being me at all times.

Semarang, July 7, 2021

Sincerely



EWEN LIAM

ABSTRACT

Diabetes is a very serious problem that needs to be solved, the cure has not been found yet, but we still could prevent it by diagnosing people who have diabetes disease. If people positively have a diabetes disease, they can quickly prevent it from a young age by watching their diet. So, it is an important thing to predict the patient who has diabetes disease. This study will show about how technology could help the medic as a milestone in medical study

The technology that used in this project is machine learning based, using logistic regression to predict the onset of diabetes. This project will also be using two methods, the first method is using Naïve Bayes from Sci-Kit learn library and the second method is Logistic Regression without using library at all.

The final result is the Naïve Bayes with Sci-Kit library performs better compare to the logistic regression without library. It could happen because, the more complex the algorithm the more the algorithm has higher accuracy. Although the model of the logistic regression without library does not have good accuracy as the Naïve Bayes with library, the logistic regression also high value which means it can be trusted code.

Keyword: Logistic Regression, Naïve Bayes, Sci-kit Learn Library, Diabetes, Prediction

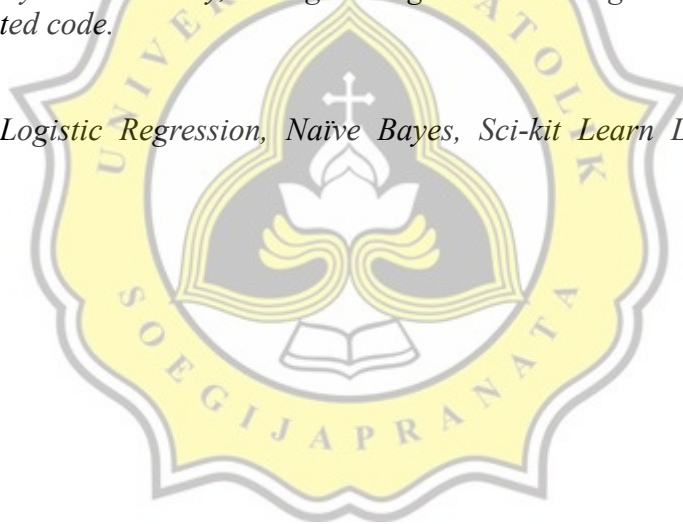


TABLE OF CONTENTS

Cover	i
APPROVAL AND RATIFICATION PAGE	i
STATEMENT OF ORIGINALITY	Error! Bookmark not defined.
ABSTRACT	vi
TABLE OF CONTENTS	vii
ILLUSTRATION INDEX	viii
INDEX OF TABLES	ix
CHAPTER 1 Introduction	1
1.1 Background	1
1.2 Problem Formulation	2
1.3 Scope	2
1.4 Objective	2
CHAPTER 2 Literature Study	3
CHAPTER 3 Research Methodology	9
3.1 Literature Study	9
3.2 Collecting Data	9
3.3 Analysis	10
3.4 Implementation and Testing	12
3.5 Conclusion	12
CHAPTER 4 Analysis and Design	13
4.1 Analysis	13
4.2 Desain	19
CHAPTER 5 Implementation and Testing	21
5.1 Implementation	21
5.2 Testing	41
CHAPTER 6 Conclusion	47
References	1
Appendix	A

ILLUSTRATION INDEX

Illustration 4.1: Information about the dataset	11
Illustration 4.2: Logistic Regression Graph (Abin, 2019).....	15
Illustration 4.3: Naïve Bayes classifier explained in image (Jake, 2016)	15
Illustration 4.4: Sigmoid function graph. (Qef, 2008)	17
Illustration 4.5: Gradient Descent Graph illustrated (2020)	17
Illustration 4.6: Flow chart of this project.....	19
Illustration 5.1: The head of the dataset	21
Illustration 5.2: Heatmap from the dataset.....	22
Illustration 5.3: Pregnancy and it's density graph.....	24
Illustration 5.4: Glucose and it's density graph.....	24
Illustration 5.5: Blood pressure and it's density graph.....	25
Illustration 5.6: Skin thickness and it's density graph.....	25
Illustration 5.7: Insulin and it's density graph.....	26
Illustration 5.8: BMI and it's density graph	26
Illustration 5.9: Diabetes pedigree function and it's density graph.....	27
Illustration 5.10: Age and it's density graph	27
Illustration 5.11: The new histogram of Pregnancy	29
Illustration 5.12: The new histogram of Glucose.....	29
Illustration 5.13: The new histogram of Blood pressure.....	30
Illustration 5.14: The new histogram of Skin thickness.....	30
Illustration 5.15: The new histogram of Insulin.....	31
Illustration 5.16: The new histogram of BMI	31
Illustration 5.17: The new histogram of Diabetes pedigree function.....	32
Illustration 5.18: The new histogram of Age	32
Illustration 5.19: Cleaning and scaling the data.	33
Illustration 5.20: The count plot of 'outcome' data.	35
Illustration 5.21: ROC Curve from the dataset	42
Illustration 5.22: Confusion matrix from the dataset	42
Illustration 5.23: The Accuracy Result of SK-Learn LR	43
Illustration 5.24: Mean Squared Errors of the model.....	44
Illustration 5.25: MSE for the best learning rate.....	45
Illustration 5.26: The Accuracy Result of LR Without Library	45

INDEX OF TABLES

Table 3.1: Attributes of the dataset	8
Table 3.2: Output variables of the result	9

