

**DEVELOPMENT OF MACHINE LEARNING MODEL
FOR A DOG DISEASES EXPERT SYSTEM**

By

Rixon Hansel Patricius

12002013

BACHELOR'S DEGREE

in

Information Technology

Faculty of Engineering and Information Technology



SWISS GERMAN UNIVERSITY

The Prominence Tower

Jalan Jalur Sutera Barat No. 15, Alam Sutera

Tangerang, Banten 15143 - Indonesia

Revision after the Thesis Defense on 2 July 2024

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.



Rixon Hansel Patricius

Student

10 July 2024
Date

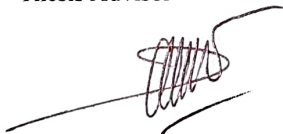
Approved by:



Dr. Maulahikmah Galinium, S.Kom., M.Sc.

Thesis Advisor

12 July 2024
Date

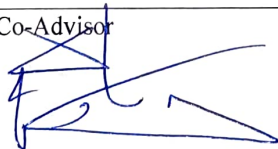


Assoc. Prof. James Purnama, S.Kom., M.Kom., M.Sc.

Thesis Co-Advisor

10/7 24

Date



Dr. Maulahikmah Galinium, S.Kom., M.Sc.

Dean

12 July 2024
Date

Rixon Hansel Patricius

ABSTRACT

DEVELOPMENT OF MACHINE LEARNING MODEL FOR A DOG DISEASES EXPERT SYSTEM

By

Rixon Hansel Patricius

Dr. Maulahikmah Galinium, S.Kom., M.Sc., Advisor

Assoc. Prof. James Purnama, S.Kom, M.Kom, M.Sc., Co-Advisor

SWISS GERMAN UNIVERSITY

Junior veterinarians often make the mistake of misdiagnosing animals during their training years. Senior veterinarians are usually too busy to assist them every time. To solve that problem, an expert system which implements machine learning model that can be used to identify dog diseases based on given symptoms is proposed to help junior veterinarians. This expert system will be in the form of web-application which is built using ReactJS. The machine learning model is trained using five algorithms (Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM)) using multi-label classification task and ROC-AUC score is used as an evaluation metric. Hyperparameter tuning was also performed to improve the performance for each model. The dataset used for the training has five different labels. After conducting the training, the best algorithm for each label are: NB for label A (0.78), SVM for label B (0.68), KNN for label C (0.70), LR for label D (0.96), and SVM for label E (0.82), with Random Forest having the longest training time and Naïve Bayes as the shortest. These models are then exported and hosted using a flask server so it can be accessed from the website by calling the API.

Keywords: Machine Learning, Expert System, Multi-label Classification, Dog Disease.

© Copyright 2024
by Rixon Hansel Patricius
All rights reserved

DEDICATION

This work is dedicated to God Almighty for His everlasting kindness
and my dearest family for their never-ending love and support.

ACKNOWLEDGEMENTS

First of all, I would like to express my biggest gratitude towards God Almighty for his blessings that I have experienced throughout my life.

I would also like to thank my parents, grandparents, and brothers that have always been on my side throughout my life. Their endless support helped me in finishing this thesis work.

Next, I would like to thank my advisor and co-advisor, Mr. Maulahikmah Galinium and Mr. James Purnama, for their guidance since the beginning of this thesis work. Without their guidance, knowledge, and experience, this thesis work would not have been complete as it is now. I would also like to give appreciation to all the lecturers in SGU for their passion in teaching and caring for the students. On top of that, I would also like to thank Ms. Nurita Kresnawati for her support in helping IT students with many administrative related things since the beginning.

I would also to express my sincere gratitude to Mr. Wilbert Adiputra, Mr. Renaldy Setiawan, and Mr. Timotius Eka Valentino as the owners of Komune Studio who gave me the opportunity to work there for a total of almost 2 years. I would also like to give a special appreciation for Mr. Wilbert Adiputra as my supervisor during my time in Komune Studio and as a teaching assistant that have been helping me and my friends with lab classes for almost the entirety of our time at SGU. I learned a lot of things from him.

Lastly, I would like to thank my friends from school and IT-2020. With IT-2020, we managed to create memorable moments throughout our time studying at SGU. despite the lockdown and online learning for the first two years of our study. My friends from school have always been there with me since the beginning, we shared many laughters and sorrows together. Shoutout to Jeremiah Aaron Rahardja as a discussion partner and for giving me additional insight in machine learning.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES	9
LIST OF TABLES.....	13
CHAPTER 1 - INTRODUCTION.....	14
1.1 Background.....	14
1.2 Research Problems	14
1.3 Research Objectives	15
1.4 Significance of Study	15
1.5 Research Questions	15
1.6 Hypothesis	15
1.7 Scope of Study.....	16
1.8 Limitations.....	16
CHAPTER 2 - LITERATURE REVIEW	17
2.1 Machine Learning.....	17
2.1.1 Multi-class Classification	18
2.1.2 Multi-label Classification.....	18
2.2 Algorithms.....	18
2.2.1 Logistic Regression	18
2.2.2 Naïve Bayes.....	19
2.2.3 K- Nearest Neighbor	20
2.2.4 Random Forest Algorithm.....	20
2.2.5 Support Vector Machine	20
2.3 Hyperparameter Tuning.....	21
2.4 Expert System.....	21
2.5 Related Works	22
2.5.1 Optimization of Naïve Bayes Method using Genetic Algorithm to Diagnose Cattle Disease	22
2.5.2 Diagnosis of Fever Symptoms Using Naive Bayes Algorithm.....	22
2.5.3 Heart Disease Prediction using Machine Learning Techniques.....	22

2.5.4	Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques	23
2.5.5	Comparing Different Supervised Machine Learning Algorithms for Disease Prediction	23
2.5.6	A Neural Network Based Expert System for Diagnosis of Diabetes ...	24
2.6	Comparison of Related Works	24
CHAPTER 3 – RESEARCH METHODS		27
3.1	Research Overview	27
3.2	Machine Learning	27
3.2.1	Experimental Environment	27
3.2.2	Dataset Gathering	27
3.2.3	Pre-processing Dataset	28
3.2.4	Training the Model	29
3.2.5	Accuracy Check	30
3.2.6	Fine-tuning the Model	31
3.3	Web-application Development	33
3.4	Integrating the Model into the Web-application	36
CHAPTER 4 – RESULTS AND DISCUSSIONS		37
4.1	Training Result (Multi-Class Classification)	37
4.1.1	First Training	37
4.1.2	Second Training	39
4.1.3	Third Training	42
4.2	Training Result (Multi-Label Classification)	43
4.3	Hyperparameter Tuning	45
4.3.1	Tuned Models for Amoeba (RandomizedSearchCV)	45
4.3.2	Tuned Models for Amoeba (GridSearchCV)	48
4.3.3	Tuned Models for Bacteria (RandomizedSearchCV)	50
4.3.4	Tuned Models for Bacteria (GridSearchCV)	52
4.3.5	Tuned Models for Giardia (RandomizedSearchCV)	55
4.3.6	Tuned Models for Giardia (GridSearchCV)	57
4.3.7	Tuned Models for Parvo (RandomizedSearchCV)	59
4.3.8	Tuned Models for Parvo (GridSearchCV)	61
4.3.9	Tuned Models for Clean (RandomizedSearchCV)	64
4.3.10	Tuned Models for Clean (GridSearchCV)	66
4.4	Web Application	68
4.5	Feedback and Evaluation	75
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS		78
5.1	Conclusions	78
5.2	Recommendations and Future Works	78
GLOSSARY		80
REFERENCES		81
CURRICULUM VITAE		83

LIST OF FIGURES

Figure	Page
Figure 2.1:Support Vector Machine (Cortes and Vapnik, 1995).....	21
Figure 3.1:Research Overview.....	27
Figure 3.2:Sequential Flow of the Machine Learning Program.....	29
Figure 3.3:Parameters of Each Machine Learning Model	29
Figure 3.4:Parameters for Logistic Regression.....	32
Figure 3.5:Parameters for Random Forest	32
Figure 3.6:Parameter for Naïve Bayes.....	33
Figure 3.7:Parameters for K-Nearest Neighbor	33
Figure 3.8:Parameters for Support Vector Machine	33
Figure 3.9:Architecture Diagram of the System	34
Figure 3.10:Activity Diagram of the Web-application	35
Figure 3.11:Use Case Diagram for the Web-application.....	36
Figure 4.1:RandomizedSearchCV Duration and Best Parameters for Logistic Regression on Amoeba	45
Figure 4.2:RandomizedSearchCV Duration and Best Parameters for Random Forest on Amoeba	46
Figure 4.3:RandomizedSearchCV Duration and Best Parameters for Naïve Bayes on Amoeba	46
Figure 4.4:RandomizedSearchCV Duration and Best Parameters for K-Nearest Neighbor on Amoeba	46
Figure 4.5:RandomizedSearchCV Duration and Best Parameters for Support Vector Machine on Amoeba	46
Figure 4.6:ROC Curve for RandomizedSearchCV Tuned Naive Bayes on Amoeba..	47
Figure 4.7:GridSearchCV Duration and Best Parameters for Logistic Regression on Amoeba	48
Figure 4.8:GridSearchCV Duration and Best Parameters for Random Forest on Amoeba	48
Figure 4.9:GridSearchCV Duration and Best Parameters for Naïve Bayes on Amoeba	48
Figure 4.10:GridSearchCV Duration and Best Parameters for K-Nearest Neighbor on Amoeba	48
Figure 4.11:GridSearchCV Duration and Best Parameters for Support Vector Machine on Amoeba	49
Figure 4.12:ROC Curve for GridSearchCV Tuned Naive Bayes on Amoeba.....	50
Figure 4.13:RandomizedSearchCV Duration and Best Parameters for Logistic Regression on Bacteria	50
Figure 4.14:RandomizedSearchCV Duration and Best Parameters for Random Forest on Bacteria	50
Figure 4.15:RandomizedSearchCV Duration and Best Parameters for Naïve Bayes on Bacteria	51

Figure 4.16:RandomizedSearchCV Duration and Best Parameters for K-Nearest Neighbor on Bacteria	51
Figure 4.17:RandomizedSearchCV Duration and Best Parameters for Support Vector Machine on Bacteria	51
Figure 4.18:ROC Curve for RandomizedSearchCV Tuned Support Vector Machine on Bacteria	52
Figure 4.19:GridSearchCV Duration and Best Parameters for Logistic Regression on Bacteria	53
Figure 4.20:GridSearchCV Duration and Best Parameters for Random Forest on Bacteria	53
Figure 4.21:GridSearchCV Duration and Best Parameters for Naïve Bayes on Bacteria	53
Figure 4.22:GridSearchCV Duration and Best Parameters for K-Nearest Neighbor on Bacteria	53
Figure 4.23:GridSearchCV Duration and Best Parameters for Support Vector Machine on Bacteria	53
Figure 4.24:ROC Curve for GridSearchCV Tuned Support Vector Machine on Bacteria	54
Figure 4.25:RandomizedSearchCV Duration and Best Parameters for Logistic Regression on Giardia.....	55
Figure 4.26:RandomizedSearchCV Duration and Best Parameters for Random Forest on Giardia.....	55
Figure 4.27:RandomizedSearchCV Duration and Best Parameters for Naïve Bayes on Giardia.....	55
Figure 4.28:RandomizedSearchCV Duration and Best Parameters for K-Nearest Neighbor on Giardia	55
Figure 4.29:RandomizedSearchCV Duration and Best Parameters for Support Vector Machine on Giardia.....	56
Figure 4.30:ROC Curve for RandomizedSearchCV Tuned Random Forest on Giardia	57
Figure 4.31:GridSearchCV Duration and Best Parameters for Logistic Regression on Giardia.....	57
Figure 4.32:GridSearchCV Duration and Best Parameters for Random Forest on Giardia.....	57
Figure 4.33:GridSearchCV Duration and Best Parameters for Naïve Bayes on Giardia	58
Figure 4.34:GridSearchCV Duration and Best Parameters for K-Nearest Neighbor on Giardia.....	58
Figure 4.35:GridSearchCV Duration and Best Parameters for Support Vector Machine on Giardia.....	58
Figure 4.36:ROC Curve for GridSearchCV Tuned K-Nearest Neighbor on Giardia..	59
Figure 4.37:RandomizedSearchCV Duration and Best Parameters for Logistic Regression on Parvo	59
Figure 4.38:RandomizedSearchCV Duration and Best Parameters for Random Forest on Parvo	60
Figure 4.39:RandomizedSearchCV Duration and Best Parameters for Naïve Bayes on Parvo	60
Figure 4.40:RandomizedSearchCV Duration and Best Parameters for K-Nearest Neighbor on Parvo	60

Figure 4.41:RandomizedSearchCV Duration and Best Parameters for Support Vector Machine on Parvo	60
Figure 4.42:ROC Curve for RandomizedSearchCV Tuned Logistic Regression on Parvo	61
Figure 4.43:GridSearchCV Duration and Best Parameters for Logistic Regression on Parvo	62
Figure 4.44:GridSearchCV Duration and Best Parameters for Random Forest on Parvo	62
Figure 4.45:GridSearchCV Duration and Best Parameters for Naïve Bayes on Parvo	62
Figure 4.46:GridSearchCV Duration and Best Parameters for K-Nearest Neighbor on Parvo	62
Figure 4.47:GridSearchCV Duration and Best Parameters for Support Vector Machine on Parvo	62
Figure 4.48:ROC Curve for GridSearchCV Tuned Logistic Regression on Parvo	63
Figure 4.49:RandomizedSearchCV Duration and Best Parameters for Logistic Regression on Clean	64
Figure 4.50:RandomizedSearchCV Duration and Best Parameters for Random Forest on Clean	64
Figure 4.51:RandomizedSearchCV Duration and Best Parameters for Naïve Bayes on Clean	64
Figure 4.52:RandomizedSearchCV Duration and Best Parameters for K-Nearest Neighbor on Clean	64
Figure 4.53:RandomizedSearchCV Duration and Best Parameters for Support Vector Machine on Clean	65
Figure 4.54:ROC Curve for RandomizedSearchCV Tuned Support Vector Machine on Clean	66
Figure 4.55:GridSearchCV Duration and Best Parameters for Logistic Regression on Clean	66
Figure 4.56:GridSearchCV Duration and Best Parameters for Random Forest on Clean	66
Figure 4.57:GridSearchCV Duration and Best Parameters for Naïve Bayes on Clean	67
Figure 4.58:GridSearchCV Duration and Best Parameters for K-Nearest Neighbor on Clean	67
Figure 4.59:GridSearchCV Duration and Best Parameters for Support Vector Machine on Clean	67
Figure 4.60:ROC Curve for GridSearchCV Tuned Logistic Regression on Clean	68
Figure 4.61:General Information Form of the Diagnose Flow	69
Figure 4.62:Pre-diagnose Form of the Diagnose Flow	69
Figure 4.63:Symptom Form Page of the Diagnose Flow	70
Figure 4.64:Result Page of the Diagnose Flow	70
Figure 4.65:Disease List Page.....	71
Figure 4.66:Create New Disease Page.....	72
Figure 4.67:Update Disease Page	73
Figure 4.68:Symptom List Page	74
Figure 4.69:Create New Symptom Page.....	74
Figure 4.70:Update Symptom Page	75
Figure 4.71:New Pre-diagnose Form (1)	76

Figure 4.72:New Pre-diagnose Form (2)	76
Figure 4.73:New Pre-diagnose Form (3)	77

LIST OF TABLES

Table	Page
Table 2.1:Advantages and Limitations of Each Algorithm the Authors Tested.....	24
Table 2.2:Comparison of Related Works.....	25
Table 4.1:Attributes of the First Dataset from Medical Record	37
Table 4.2:Result of First Training with 70/30 Train/Test Split	38
Table 4.3:Result of First Training with 75/25 Train/Test Split	38
Table 4.4:Result of First Training with 80/20 Train/Test Split	38
Table 4.5:Attributes of the Second Dataset Made with the Guidance from Senior Veterinarians	39
Table 4.6:Result of Second Training with 70/30 Train/Test Split.....	41
Table 4.7:Result of Second Training with 75/25 Train/Test Split.....	41
Table 4.8:Result of Second Training with 80/20 Train/Test Split.....	41
Table 4.9:Attributes of the Third Dataset from Medical Record	42
Table 4.10:Result of Third Training with 70/30 Train/Test Split.....	42
Table 4.11:Result of Third Training with 75/25 Train/Test Split.....	43
Table 4.12:Result of Third Training with 80/20 Train/Test Split.....	43
Table 4.13:Training Result for Amoeba	44
Table 4.14:Training Result for Bacteria	44
Table 4.15:Training Result for Giardia.....	44
Table 4.16:Training Result for Parvo	44
Table 4.17:Training result for Clean.....	45
Table 4.18:ROC-AUC Scores for Amoeba after Hyperparameter Tuning using RandomizedSearchCV	47
Table 4.19:ROC-AUC Scores for Amoeba after Hyperparameter Tuning using GridSearchCV	49
Table 4.20:ROC-AUC Scores for Bacteria after Hyperparameter Tuning.....	51
Table 4.21:ROC-AUC Scores for Bacteria after Hyperparameter Tuning using GridSearchCV	54
Table 4.22:ROC-AUC Scores for Giardia after Hyperparameter Tuning using RandomizedSearchCV	56
Table 4.23:ROC-AUC Scores for Giardia after Hyperparameter Tuning using GridSearchCV	58
Table 4.24:ROC-AUC Scores for Parvo after Hyperparameter Tuning using RandomizedSearchCV	60
Table 4.25:ROC-AUC Scores for Parvo after Hyperparameter Tuning using GridSearchCV	63
Table 4.26:ROC-AUC Scores for Clean after Hyperparameter Tuning using RandomizedSearchCV	65
Table 4.27:ROC-AUC Scores for Clean after Hyperparameter Tuning using GridSearchCV	67