



**The Influence of Using Up Sampling Method in Predicting Patients' Disease
Types using a Combination of Natural Language Processing, Naive Bayes
Algorithm, XGBoost and Support Vector Machine**

Thesis Report

Mhd Avicenna W.R
41517010028

DEPARTMENT OF INFORMATICS
FACULTY OF COMPUTER SCIENCE
UNIVERSITAS MERCU BUANA
JAKARTA
2021



**The Influence of Using Up Sampling Method in Predicting Patients' Disease
Types using a Combination of Natural Language Processing, Naive Bayes
Algorithm, XGBoost and Support Vector Machine**

Thesis Report

Submitted to Complete Terms
Completed a Computer Bachelor Degree

Created By:
Mhd Avicenna W.R
41517010028

DEPARTMENT OF INFORMATICS
FACULTY OF COMPUTER SCIENCE
UNIVERSITAS MERCU BUANA
JAKARTA
2021

Universitas Mercu Buana

ORIGINALITY STATEMENT SHEET

The undersigned below:

Student Number : 41517010028

Student Name : Mhd Avicenna W.R

Title of Final Project :The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

Stating that my Final Project Report is my own and not plagiarism. If it is found in my Final Project Report that there is an element of plagiarism, then I am ready to get academic sanctions related to it.

Jakarta, 19 March 2021



Mhd Avicenna W.R

UNIVERSITAS
MERCU BUANA

FINAL PROJECT PUBLICATION STATEMENT

As a Universitas Mercu Buana student, I, the undersigned below :

Student Name : Mhd Avicenna W.R
Student Number : 41517010028
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

By giving permission and approval of **Non-exclusive Royalty Free Right** to Universitas Mercu Buana for my scientific work entitled above along with the available devices (if necessary).

With this Non-exclusive Royalty Free Right, Universitas Mercu Buana has the right to store, transfer / format, manage in form of database, administer and publish my final project.

Furthermore, in sake of science development in Universitas Mercu Buana environment, I give the permission to Researcher in Research Lab of Computer Science Faculty, Universitas Mercu Buana to use and develop existing result of the research of my final project for the research and publication purpose as long as my name is stated as author / creator and Copyright owner.

Hereby I made this statement in truthfulness.

UNIVERSITAS Jakarta, 19 March 2021
MERCU BUANA



Mhd Avicenna W.R

FINAL PROJECT OUTPUT STATEMENT LETTER

As a Universitas Mercu Buana student, I, the undersigned below:

Student Name : Mhd Avicenna W.R
 Student Number : 41517010028
 Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

Declare that :

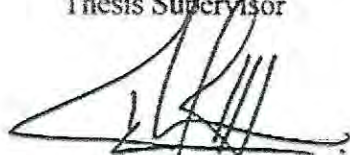
1. My Final Project Output as follows :

No	Output	Type	Status
1	Scientific Publication	Not Accredited national Journal	Submitted ✓
		Accredited National Journal	
		Not Reputable International Journal	Approved
		Reputable International Journal ✓	
Submitted / Published :	Journal Name :	The Science and Information Organization	
	ISSN :		
	Journal Link :		
	Published Journal Link File :		

2. Willing to complete the entire article publication process starting from submitting, revising the article until it is declared that it can be published in the intended journal.
3. Asked to attach a scanned ID card and a statement letter (see the HKI document attachment), for the purpose of registering HKI if needed.

This statement I made in truth.

Approved
 Thesis Supervisor



Dr. Mujiono Sadikin. MT. CISA. CGEIT..

Jakarta, 19 March 2021



Mhd Avicenna W.R

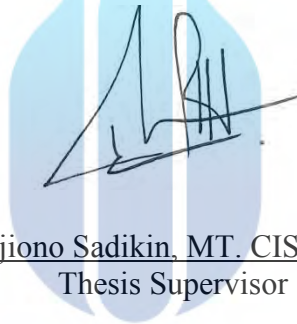
APPROVAL SHEET

Student Name : Mhd Avicenna W.R
Student Number : 41517010028
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

The Final Project has been reviewed and accepted for thesis defense.

Jakarta, 19 March 2021

Approved and Accepted,

A handwritten signature in black ink, appearing to read 'M. Sadikin', is written over a light blue circular stamp. The stamp contains a stylized graphic of a flame or a similar shape.

(Dr. Mujiono Sadikin, MT, CISA, CGEIT.)
Thesis Supervisor

UNIVERSITAS
MERCU BUANA

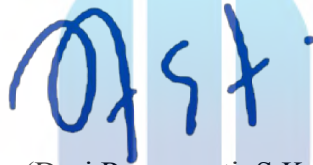
COMMITTEE APPROVAL SHEET

Student Number : 41517010028
Student Name : Mhd Avicenna W.R
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

This Thesis has been examined and tried as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 19 March 2021

Approved,



(Desi Ramayanti, S.Kom.,MT)
Head of Defense Committee

UNIVERSITAS
MERCU BUANA



(Dr. Ida Nuraida, MT)
Defense Committee 1

(Dr. Leonard Goeirmento)
Defense Committee 2

COMMITTE APPROVAL SHEET

Student Number : 41517010028
Student Name : Mhd Avicenna W.R
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

This Thesis has been examined and tried as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 19 March 2021

Approved



(Desi Ramayanti, S.Kom.,MT)

MERCU BUANA

COMMITTE APPROVAL SHEET

Student Number : Mhd Avicenna W.R
Student Name : 41517010028
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

This Thesis has been examined and tried as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 19 March 2021

Approved

(Dr. Leonard Goeirmanto)

UNIVERSITAS
MERCU BUANA

COMMITTE APPROVAL SHEET

Studen Number : 41517010028
Student Name : Mhd Avicenna W.R
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

This Thesis has been examined and tried as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 19 March 2021

Approved



UNIVERSITAS
MERCU BUANA

(Dr. Ida Nuraida, MT)

VALIDITY SHEET

Student Number : 41517010028
Student Name : Mhd Avicenna W.R
Title of Final Project : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

This Thesis has been examined and tried as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 19 March 2021

Approved,



(Dr. Mujiono Sadikin, MT. CISA, CGEIT.)
Thesis Supervisor

Acknowledged,



(Diky Firdaus, S.Kom, MM)
Informatics Thesis Coordinator

(Desi Ramayanti, S.Kom, MT)
Head of Informatics Department

ABSTRAK

Nama : Mhd Avicenna W.R
NIM : 41517010028
Pembimbing TA : Dr. Mujiono Sadikin, MT. CISA, CGEIT.
Judul : Pengaruh Penggunaan Metode Up Sampling dalam Memprediksi Jenis Penyakit Pasien Menggunakan Kombinasi Natural Language Processing, Algoritma Naive Bayes, XGBoost dan Support Vector Machine

Rumah sakit adalah tempat penyedia pelayanan kesehatan yang sangat penting. Untuk memastikan pelayanan yang diberikan dirasa maksimal, menerapkan teknologi yang dapat membantu pelayanan ini sangat dibutuhkan. Teknologi yang dapat digunakan seperti pengolahan data-data kegiatan rumah sakit secara maksimal, dikarenakan masih sedikit rumah sakit yang memanfaatkan data-data yang mereka miliki selain dijadikan histori kegiatan maka dibutuhkan pengelolaan data yang lebih lanjut. Data-data kegiatan yang dapat dimanfaatkan seperti data pemeriksaan pasien, data obat-obatan, data penyakit, hingga data yang berhubungan dengan penanganan pasien. Dalam data-data kegiatan ini bisa dimanfaatkan sebagai salah satu acuan tenaga medis dalam mengambil keputusan untuk memberikan pelayanan terbaik terlebih khusus dokter agar terhindar dari adanya Medical Error. Dalam penelitian ini ingin menunjukkan bahwa pengaplikasian Data mining dengan mengkombinasikan Natural Language Processing, Algoritma Naïve Bayes, XGBoost, dan Support Vector Machine (SVM) untuk menguji data yang akan diolah dalam klasifikasi jenis penyakit berdasarkan dataset yang telah ada dan juga ingin mengetahui seberapa besar pengaruh dari dataset yang tidak seimbang (Unbalance) dan data set yang seimbang (Balance). Penggunaan XGBoost dan SVM dipilih karena memiliki kemampuan komputasi yang baik, efisien dalam waktu pengolahan data dan Akurasi yang cukup tinggi. Dengan menggunakan ketiga algoritma didapatkan hasil klasifikasi jenis penyakit dengan akurasi rata-rata setiap algoritma sebesar 69.22% dengan algoritma Naïve Bayes, 92.13% dengan XGBoost dan 88.49% dengan SVM. Dalam penelitian ini juga diperoleh pengaruh dari dataset yang tidak seimbang dan seimbang dengan rata-rata keefektifan penggunaan metode Up sampling pada data unbalance ini sebesar 2.72%.

Kata kunci:

Data mining, Natural Language Processing, Word2Vec, Naïve Bayes, XGBoost, Support Vector Machine, Unbalanced

ABSTRACT

Student Name : Mhd Avicenna W.R
Student Number : 41517010028
Counsellor : Dr. Mujiono Sadikin, MT. CISA, CGEIT.
Title : The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine

The hospital is a very important place for health care providers. To ensure that the services provided are considered to be maximal, applying technology that can help these services is urgently needed. Technology that can be used is such as processing hospital activity data optimally, because there are still a few hospitals that use the data they have besides being used as activity history, further data management is needed. Activity data that can be utilized, such as patient examination data, drug data, disease data, and data related to patient handling. In this activity data can be used as a reference for medical personnel in making decisions to provide the best service, especially for doctors to avoid medical errors. In this study, we want to show that the application of data mining by combining Natural Language Processing, Naïve Bayes Algorithm, XGBoost, and Support Vector Machine (SVM) to test data to be processed in the classification of disease types based on existing datasets and also wants to know how much influence it has. from an unbalanced dataset (Unbalance) and a data set that is balanced (Balance). The use of XGBoost and SVM was chosen because they have good computational capabilities, are efficient in data processing time and are quite high in accuracy. By using the three algorithms, the classification results of the type of disease are obtained with an average accuracy of each algorithm of 69.22% with the Naïve Bayes algorithm, 92.13% with XGBoost and 88.49% with SVM. In this study also obtained the effect of an unbalanced and balanced dataset with an average effectiveness of using the Up sampling method on this unbalanced data of 2.72%.

Key words:

Data mining, Natural Language Processing, Word2Vec, Naïve Bayes, XGBoost, Support Vector Machine, Unbalanced

PREFACE

Praise the writers to the Almighty God who has bestowed His grace and gifts, so that the author can complete the final project entitled “*The Influence of Using Up Sampling Method in Predicting Patients' Disease Types using a Combination of Natural Language Processing, Naive Bayes Algorithm, XGBoost and Support Vector Machine*” just in time. This final project is structured to fulfill one of the requirements to obtain a bachelor's degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana. The author is fully aware that in completing this thesis report will not escape the support and guidance of the closest people, therefore the author would like to express my gratitude as possible to:

1. Dr. Mujiono Sadikin, MT. CISA, CGEIT. as the Thesis Supervisor who has taken the time to provide guidance and direction in the preparation of this final project to completion.
2. Desi Ramayanti, S.Kom, MT. Head of the Informatics Department at the Universitas Mercu Buana, as well as being the academic advisor, thank you for the knowledge you have deliberated to me as guidance completing the thesis report.
3. Diky Firdaus, S.Kom., MM. as the Thesis Coordinator for the Informatics Engineering Study Program, Faculty of Computer Science, Mercu Buana University.
4. Mrs. Prastika Indriyanti, S.Kom., M.Cs as Head of International Informatics Department of Universitas Mercu Buana.
5. Parents who always provide prayers and support to the author.
6. Friends who always encourage the author during the implementation of the final project.
7. All parties that the author cannot mention one by one who have helped a lot in the preparation of the final project.

Finally, the authors hope that this final project can be useful for readers to increase knowledge and insight.

Jakarta, March 2021
Mhd Avicenna W.R

TABLE OF CONTENTS

TITLE PAGE	i
ORIGINALITY STATEMENT SHEET	ii
FINAL PROJECT PUBLICATION STATEMENT	iii
FINAL PROJECT OUTPUT STATEMENT LETTER.....	iv
APPROVAL SHEET	v
COMMITTEE APPROVAL SHEET	vi
VALIDITY SHEET	x
ABSTRAK	xi
ABSTRACT.....	xii
PREFACE.....	xiii
TABLE OF CONTENTS.....	xiv
JOURNAL TEXT	1
WORKING PAPER.....	11
PART 1. LITERATUR REVIEW	13
PART 2. ANALYSIS AND DESIGN.....	20
PART 3. SOURCE CODE.....	22
PART 4. DATASET.....	69
PART 5. EXPERIMENT STAGE	73
PART 6. RESULTS ALL EXPERIMENTS	78
PART 7. BIBLIOGRAPHY	96
ATTACHMENT OF HAKI DOCUMENTS	98

The Influence of Using Up Sampling Method in Predicting Patients' Disease Types Using a Combination of Natural Language Processing, Naïve Bayes Algorithm, XGBoost and Support Vector Machine

Mhd Avicenna W.R¹

Faculty of Computer Science,

Universitas Mercu Buana,

Meruya Selatan No. 1, Kembangan, Jakarta, Indonesia

41517010028@student.mercubuana.ac.id

Mujiono Sadikin²

Faculty of Computer Science,

Universitas Mercu Buana,

Meruya Selatan No. 1, Kembangan, Jakarta, Indonesia

mujiono@mercubuana.ac.id

Abstract— *The hospital is a very important place for health care providers. To ensure that the services provided are considered to be maximal, applying technology that can help these services is urgently needed. Technology that can be used is such as processing hospital activity data optimally, because there are still a few hospitals that use the data they have besides being used as activity history, further data management is needed. Activity data that can be utilized, such as patient examination data, drug data, disease data, and data related to patient handling. In this activity data can be used as a reference for medical personnel in making decisions to provide the best service, especially for doctors to avoid medical errors. In this study, we want to show that the application of data mining by combining Natural Language Processing, Naïve Bayes Algorithm, XGBoost, and Support Vector Machine (SVM) to test data to be processed in the classification of disease types based on existing datasets and also wants to know how much influence it has from an unbalanced dataset (Unbalance) and a data set that is balanced (Balance). The use of XGBoost and SVM was chosen because they have good computational capabilities, are efficient in data processing time and are quite high in accuracy. By using the three algorithms, the classification results of the type of disease are obtained with an average accuracy of each algorithm of 69.22% with the Naïve Bayes algorithm, 92.13% with XGBoost and 88.49% with SVM. In this study also obtained the effect of an unbalanced and balanced dataset with an average effectiveness of using the Up sampling method on this unbalanced data of 2.72%.*

Keywords— *Data mining, Natural Language Processing, Word2Vec, Naïve Bayes, XGBoost, Support Vector Machine, Unbalanced*

I. Introduction

Hospital is an agency established to help people get health assistance in the form of care, consultation and other services related to health. This health service is vital, so it needs a sufficient number of medical personnel and experts to provide maximum health services.

However, in some hospitals there are only a few medical personnel, which causes the quality of health services to decline. In terms of existing technology, there are still few hospitals that use patient-related data for further use. Hospital

activity data has various types such as medical check up data, diagnosis data, disease type data, disease class, and patient visit history. The amount of this data can be used further to assist health workers in carrying out their duties. Medical error is a case that is caused by an incorrect decision made by the health worker, as such it must be avoided [1]. This decision making is very risky, because it can threaten the life of the patient if the treatment is wrong or late so we need a technology that can help health workers to reduce cases of medical error [2].

Data Mining is one part of the Knowledge Discovery in Dataset (KDD) stage. The steps in KDD include cleaning data (data cleaning), data integration (data integration), data selection, data transformation, mining process (data mining), evaluation of patterns that have been made (Pattern Evaluation) and the delivery of new required knowledge (Knowledge Presentation) [3], [4]. In its use, Data Mining can be used as a Decision Support System [5] for health workers, especially doctors. By using certain methods it can provide drug recommendations from medical record data besides this Decision Support System can be used as one of the quick steps that health professionals can take to help medical management of patients who have similar things, thus giving doctors alertness in making a decision [6], [7].

In the examination data, there are many elements that can be used in providing decision recommendations to health professionals such as anamnesis (Medical Abstract), diagnosis and the patient's disease class. This anamnesis data is data from the results of examinations conducted by doctors to patients in the form of Medical Abstracts which are structured text. To process Medical Abstract data, it is necessary to do text mining using Natural-Language Processing (NLP) as preprocessing [8], [9], but not only data from Medical Abstracts are processed using Natural-Language Processing but text data from diagnosis and disease classes are also processed. So that all text data can be processed in data mining [10]. By using the Naïve Bayes

algorithm, XGBoost and the Support Vector Machine, you can get a classification of the type of disease suffered by a patient based on the doctor's examination.

In this research, the Naïve Bayes algorithm, XGBoost, and Support Vector Machine combined with Natural-Language Processing were used in data processing to obtain a classification of diseases suffered by patients based on the results of a doctor's examination. Types of disease can be divided into two, namely "acute" and "chronic". A disease is called acute if it is temporary and can be cured after receiving treatment, while chronic illness is a disease that affects patients for a long time, recurs, and requires a relatively long and regular treatment time, and the ability to limit a person's lifestyle [11] so as not to affect his health. There are two main stages carried out in this research, namely: 1) Preprocessing which includes Cleaning Data and representing Medical Abstracts, Diagnoses and Disease Groups in the form of Vector which is obtained from the Natural-Language Processing training data model utilizing the Word2Vec Word Embedding method of python gensim library [12], [13]; 2) Classification uses the Naïve Bayes algorithm, XGBoost and Support Machine Vector and tests the accuracy of each algorithm. The final result of this research is a classification of disease types that can be used as a recommendation so that it can increase the alertness of doctors and other medical personnel in making decisions to treat patients and minimize medical errors.

In this research, we also wanted to know whether the effect of unstructured data from the Word2Vec training dataset affects the accuracy results, and also wanted to know whether the effect of unbalanced data could affect accuracy. So that this research consists of four main scenarios.

II. Research Methods

In this research, the Naïve Bayes, XGBoost, and Support vector machine (SVM) classification techniques are used in the python programming language to run the algorithm. In general, the research stage consists of data collection, training the Word2Vec model, data preprocessing, displaying medical abstract data into vectors of length N, training and testing the model using the Naïve Bayes algorithm, XGBoost algorithm, and the SVM algorithm. At the data collection stage, the dataset is obtained in the form of medical abstracts, diastolic blood pressure, systolic blood pressure, respiratory rate,

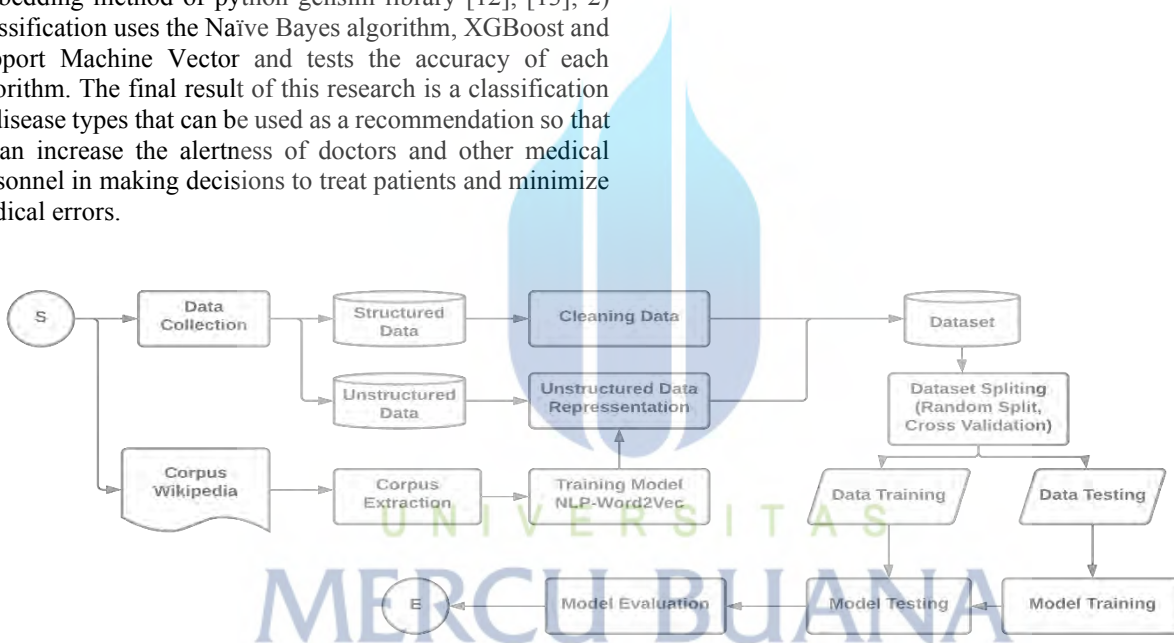


Figure 1. Research Stages - first Scenario

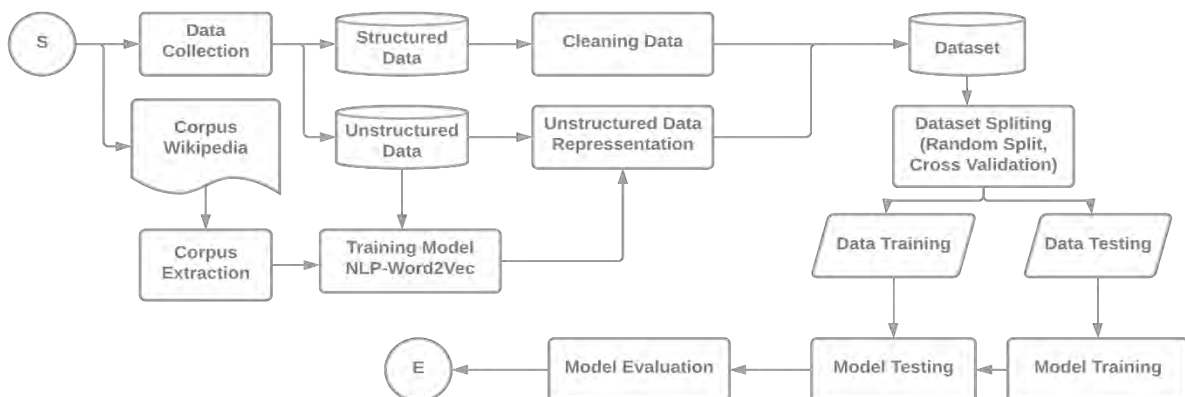


Figure 2. Research Stages - second Scenario

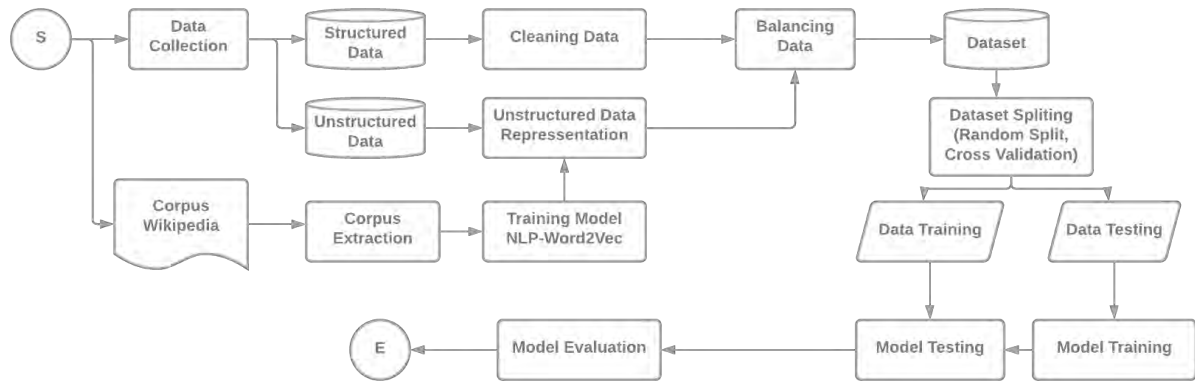


Figure 3. Research Stages - first Scenario Balancing

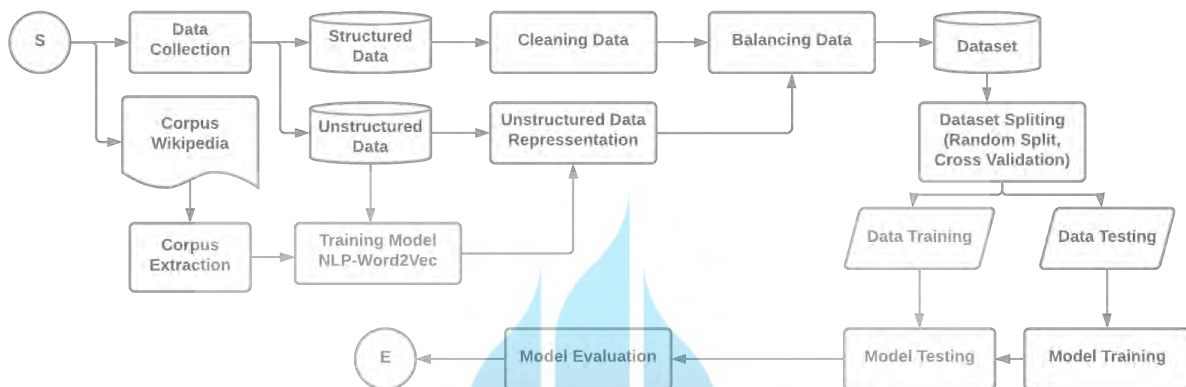


Figure 4. Research Stages - second Scenario Balancing

temperature, pulse, age, gender, dizziness, diagnosis results, decisions, diagnosis, disease class and type of disease. Word2Vec is used to represent abstract text data. This word2vec training generates numeric vectors along N dimensions. At the model testing data training stage, the Naïve Bayes, XGBoost, and SVM algorithms were used.

The process of representing medical abstracts into vector data requires the Word2Vector model to map keywords to vector data. However, some of the keywords in the medical abstracts were not found in the Word2Vec model, due to the inconsistency of doctors entering the medical abstracts input. In order to avoid missing keywords, this research was conducted using two methods to train the Word2Vec model. **Figure 1** shows a diagram of the stages of research with the Word2Vec model training process using the Wikipedia corpus to process the dataset. However, if the medical abstract keyword is not found in the Word2Vec model, an update is made to the Word2Vec model by adding keywords to the model and then retraining the model. So the model can avoid inconsistencies from medical abstract input. **Figure 2** shows a diagram of the research stages with the second scenario. The second scenario process for the Word2Vec model does not only use the wikipedia corpus as the dataset, but uses the wikipedia corpus combined with the medical abstract as the data set. This way the Word2Vec model avoids inconsistencies in medical abstracts. The process of representing medical abstracts is also carried out on the diagnosis and class data. This data is also carried out by the same process so that a vector that contains diagnosis data is generated and a vector containing data for the sick group is also produced.

In this research, the prediction target, namely the type of disease, has an unbalanced number of comparison data between "ACUTE" and "CHRONIC" with a ratio of 90.5: 9.5. Which is why it is necessary to do a method, namely resampling, so that the data used produces better accuracy. By applying the Oversampling method to the dataset, making the comparison between acute and chronic becomes 50:50. This is also done to find out the effect of the data used which is not balanced with the balance. This oversampling method is also applied to 2 scenarios in Word2Vec, **Figure 3** shows a balanced Wiki corpus scenario and **Figure 4** shows a wiki corpus scenario combined with balancing medical abstract data.

A. Data Collection

The stages of data collection were carried out by making observations and interviews. From the results of observations and interviews that have been conducted, it is found that there is no use of doctor's examination data to serve as a Decision Support System in the form of a classification of disease types. The data held is only stored and used as patient history, so there is no further use. The data collected were 4,404 examination records with details as in **Table 1**.

Table 1. Structure of the Patient Examination Dataset

Atribut	Tipe data	Range Nilai
Anamnesis	Text	Text
Tenang	Varchar	Ordinal
Cemas	Varchar	Ordinal
rr	Integer	Continue

Umur	Integer	Continue
Jnskelamin	Varchar	Ordinal
Systol	Integer	Continue
Diastol	Integer	Continue
Suhu	Numeric	Continue
Nadi	Integer	Continue
Kndslimbung	Varcar	Ordinal
Hasil	Varchar	Ordinal
Keputusan	Varchar	Ordinal
Diagnosa	Text	Text
Golongan sakit	Text	Text
Jnspenyakit	Varchar	Ordinal

Description Table 1.

- Anamnesis : Medical abstracts in the form of text descriptions from the patient during the examination.
- Continuous range of values: are values of numeric and numeric types
- Ordinal value range: is a scale that differentiates categories by level / order.
- The range of values is Calm (Tenang), Anxious (Cemas), unsteady (kndslimbung), is Y and N.
- rr (Respiratory Rate) : a record of the patient's respiratory rate.
- The range of values for jnskelamin (sex) was L and P
- Result (hasil) Value Range is "TR (Tidak Resiko / No Risk)", "RR (Resiko Rendah / Low Risk)", and "RT (Resiko Tinggi / High Risk)"
- Decision Value Range is S (Calm) and 0 (Uneasy)
- Range of jnspenyakit value is "Chronic" and "Acute".

B. Cleaning Data

Cleaning Data is a process that must be done before the data algorithm can be implemented. The process of cleaning data includes removing invalid data in the dataset, such as large numbers of empty values. Some data that has blank values can be tolerated and still usable. By using a method used to swap the blank value with the average value that is recognized from the variable value [14], [15]. The average value obtained becomes a constant that fills in the blank value and has no effect on the relationship between properties that can affect the use of the Data Mining algorithm. At this stage the data that is carried out cleaning includes calm, anxiety, rr (respiratory rate), age, sex, systol, dsyastol, temperature, pulse, kndslimbung, results, decisions and disease. At this point, 3108 data records are ready for use.

C. Convert Text Data

Text conversion is a way of converting text-type data into variable data that is understood by data mining algorithms. In converting text data into variable data that can be understood by a data mining algorithm, the Word2Vec Word Embedding method is used. Word2Vec Word Embedding is a method for converting text data into variables that can be understood by data mining algorithm by dividing the text into groups of words which are then converted into vectors based on the label generated using probability calculations by the corpus model that has been made [16], [17]. Word2Vec has several models including Skip-Gram and Continuous Bag-Of-Word (CBOW). The Skip-Gram Model has a context prediction method of a word as input, while the CBOW (Continuous Bag-Of-Word) Model uses prediction with attributes that are nearby as a reference [17]. Visualization of the concept is seen in Figure 5 below[18].

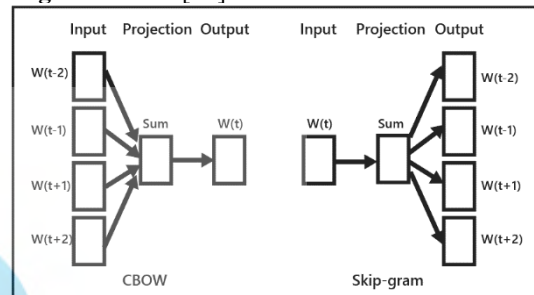


Figure 5. Architecture CBOW and Skip-Gram

Converting text data in this research, converting text data in the form of Medical Abstracts, Diagnoses and Disease Groups. The process of representing abstract medical data into vector data using the word embedding method requires a training data model. To create a training data model, a large number of Indonesian texts are required. The text obtained from the wikipedia corpus was 408,952 articles in Indonesian which were extracted using wikicorpus. Furthermore, text data that has been created using the Gensim Word2Vec library by initiating the Word2Vec function and entering all the text into this function. The result of the text is a model with a size of 25 vectors which is used to map the semantic proximity positions between words from an input text.

After the training data model has been created, the next process is cleaning text data for Medical Abstracts, Diagnoses, and Diseases. This process is carried out before the representation process becomes vector data. The process of cleaning text data uses the Indonesian Natural-Language Processing library in python. The steps taken include Lemmatization, Removing number, Stopword removal and Post Tagger.

Lemmatization stage is the stage used to convert words into basic forms in text data. Lemmatization changes words by considering the context of the word, so that the word does not just remove some characters in a word but also takes into account its meaning and accuracy. The next step is removing numbers, which aims to remove numbers from text data, because numbers in text data have no effect on getting the root word. Stopword removal is used to remove common words that are considered meaningless. An example of a stopword in Indonesian is "yang", "dan", "di", "dari" and much more. By using the Stopword process, you can focus

the text data on only important words. Post Tagger is used to get Part-Of-Speech tags from text which is useful for categorizing word classes. After the text data has been cleaned successfully, the data is ready to be represented as vector data.

Table 2. Anamnesa Vector Data Representation

Teks	V1	V2	...	V25
batuk	-0.00452	0.01768	...	0.00079
dahak	0.01737	0.00581	...	-0.01433
nafas	0.01186	0.01459	...	-0.01567
demam	0.00434	0.00833	...	-0.00749

Table 3. Diagnosis Vector Data Representation

Teks	V1	V2	...	V25
Tuberculosis	-0.08049	-0.11207	...	-0.47349
Vertigo	-2.83492	4.25373	...	4.37310
Dyspepsia	-0.02305	-0.01820	...	-0.01567
Haemoptysis	-0.00071	-0.01652	...	-0.00908

Table 4. Golongan Sakit Vector Data Representation

Teks	V1	V2	...	V25
Tuberkolus is	0.02735	0.12261	...	-0.05846
Neoplasma	-0.33051	0.11408	...	-1.14892
Bronkitis	-1.31546	-0.34373	...	-0.20741
Asma	-0.40387	1.00817	...	2.21707

D. Balancing Data

Balancing Data is a process for balancing datasets that have more dominant values than other data [19]. This unbalanced data can result in the training data learning to be more dominant [20]. That is why resampling is carried out in order to produce balanced data, one of which is using the Up Sampling method, this method was chosen because the number of minority data is below 50% of the total data so this method makes the scenario as if the minority data has the same amount of data as the majority data by doing recalculate each value in the dataset and duplicate data that are similar to minorities, so that minority data can have the same number of datasets as the majority data.

E. Dataset Splitting

The process of separating the dataset is carried out to divide the dataset into two categories, namely testing data and training data. In this research, using two types of splitting methods, namely Random Split and K-Fold Cross Validation. The Random Split method is done by taking random data for testing and training with a certain ratio. The K-Fold Cross Validation method is carried out by dividing the data into K sections on the dataset of the same size in order to eliminate data habits. The training and testing process is carried out as many as K which has been determined [21], [22]. **Figure 6** Shows the iteration progress in the K-Fold Cross Validation method.

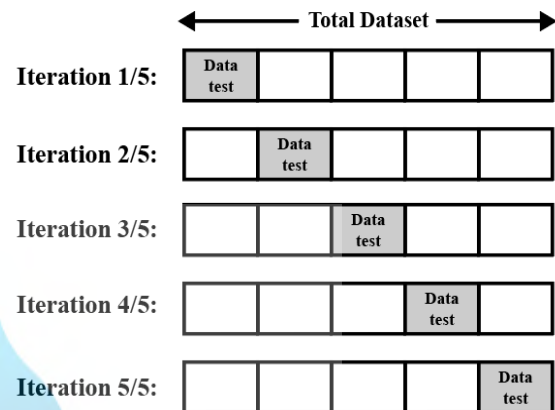


Figure 6. Cross Validation K-Fold with 5K

This research uses two methods that have the same number of comparisons for each algorithm and the Word2Vec training model scenario. From a total of 3108 rows of data, it is divided into three comparisons, namely 9: 1, 8: 2 and 7: 3. While K-fold Cross Validation uses 3 K values, namely 5K, 10K, and 15K.

F. Naïve Bayes Algorithm

The Naïve Bayes algorithm is an algorithm that uses a probability and statistical model invented by a British scientist named Thomas Bayes [23], [24]. Classification that is done to predict the future based on past experiences. This algorithm aims to predict the class based on the training data provided **Figure 7** is a form of the Naïve Bayes general formula.

The Naïve Bayes algorithm is an algorithm that uses a probability and statistical model invented by a British scientist named Thomas Bayes [23], [24]. Classification that is done to predict the future based on past experiences. This algorithm aims to predict the class based on the training data provided **Figure 7** is a form of the Naïve Bayes general formula.

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

Figure 7. Naïve Bayes Theorem

Explanation :

X = Data with an unknown class

y = The hypothesis of class X data is a specific class
 $P(y|X)$ = The probability of the hypothesis y based on condition X
 $P(y)$ = Hypothesis probability y
 $P(X|y)$ = Probability of X based on the condition
 $P(X)$ = Probability of X

Implementation of the Naïve Bayes Algorithm is carried out on the patient examination dataset. This implementation is done by creating a Naïve Bayes model using training data. The model is used to predict testing data, then perform calculations in the form of predictive accuracy from the testing data.

G. XGBoost

XGBoost (Extreme Gradient Boosting) is a method in machine learning which is a variant of the boosting tree model algorithm that can perform more optimal and faster computations [25]. This method is a development of the Gradient Boosting Machine, which has the ability to perform parallel computations and can avoid overfitting in the dataset. Therefore XGBoost is often used and very popular because of its speed and ease of use, **Figure 8** is a general formula XGBoost [26].

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F$$

Figure 8. XGBoost Theorem

Explanation:

\hat{y}_i = Probabilitas hipotesis \hat{y}_i
 K = Total tree
 F = Basic Tree Model

XGBoost implementation was carried out on the patient examination dataset. This implementation is done by creating an XGBoost model using training data. The model is used to predict the testing data, then perform calculations in the form of predictive accuracy from the testing data.

H. Support Vector Machine

Support Vector Machine (SVM) is a method in supervised learning used for classification and regression. SVM is used to find the best Hyperplane by maximizing the distance between classes [27], [28].

In SVM there are several kernels that can be used to assist in classification including Linear Kernel, Radial Basis Function (RBF), Polynomial, and Sigmoid. [29]. The following is the basic formula for each kernel.

$$k(X_p, X_q) = X_p^T X_q$$

Figure 9. Linear Kernel Theorem

$$k(X_p, X_q) = \exp(-\alpha \|X_p - X_q\|), \alpha > 0$$

Figure 10. RBF Kernel Theorem

$$k(X_p, X_q) = (\alpha X_p^T + X_q), \alpha > 0$$

Figure 11. Polynomial Kernel Theorem

$$k(X_p, X_q) = \tanh(\alpha X_p^T X_q + h)$$

Figure 12. Sigmoid Kernel Theorem

In SVM there is a parameter C, which is used to adjust the amount of margin that will be used and set the amount of penalty in the classification that can be included in the margin so that the accuracy of the SVM can be better. The parameter value C in SVM has a default of $C = 1$, and does not have special settings for the value of the C parameter itself [30].

The implementation of each SVM model is carried out on the patient examination dataset. This implementation is done by creating a Kernel model in SVM using training data. The model is used to predict the testing data, then perform calculations in the form of predictive accuracy from the testing data.

III. RESULT AND DISCUSSION

To get a prediction of the type of disease, in this research using a combination of Natural-Language Processing, Naïve Bayes Algorithm, XGBoost, and Support Vector machine. Natural-Language Processing is used in Medical Abstracts, Diagnoses and Illness Groups in order to represent data in vector form, so that unstructured data and other text data can be processed in data mining algorithms. The Naïve Bayes algorithm, XGBoost and Support Vector Machine are used to classify and test the accuracy of each algorithm.

To be able to represent Medical Abstracts, Diagnosis and Group of sickness required Natural Language Processing Model. The training data model was made from a total of 415,307 Indonesian articles from the Corpus Wikipedia extracted via Wikicorpus. Word2Vec modeling was carried out in 2 scenarios, both of these scenarios were carried out to overcome inconsistencies in abstract medical data input, diagnosis and sickness groups. The training process for the Word2Vec model in the first scenario only uses the wikipedia corpus as the trainer data, which takes 25 minutes 39 seconds and produces a txt file measuring 790 MB. The next stage is to train the txt data that has been created using the Word2vec generator library. The data training process takes 30 minutes 16 seconds to produce 3 training data model files measuring 25 vectors.

Word2Vec's second scenario model training process does not only use the wikipedia corpus as the dataset. However, combining the corpus with medical abstracts has been done by simple preprocessing, namely lemmatization and eliminating numeric characters and symbols into the dataset. The extraction process of the articles produced a txt file measuring 790 MB in 27 minutes 51 seconds. Next, train the txt data file that was created using the Word2Vec generator library. The training data process takes 23 minutes 15 seconds to produce 3 training data model files measuring 25 vectors.

The next step is to change the medical abstract data, diagnoses and sickness groups using previously made models so that the data can be represented as vector data. In this research, four key words were taken from medical abstract data, diagnosis and sickness groups. These keywords are

represented using two model scenarios that have been made. Each scenario model measures 25 proximity semantic vector data for input keywords. So that from one Medical Abstract record, 100 vector data is obtained for each scenario, as well as the Diagnosis and Disease Group.

Training text data into vectors using Natural Language Processing processes each text data into 3 different datasets with 2 different scenarios resulting in 6 datasets. The time needed to process medical abstract data scenario 1 is 3 hours 11 minutes 11 seconds, medical abstract scenario 2 is 1 hour 20 minutes 4 seconds, scenario 1 diagnostic data is 2 hours 15 minutes 55 seconds, diagnosis scenario 2 is 1 hour 39 minutes 57 seconds, the data for the sick group in scenario 1 is 1 hour 23 minutes 33 seconds, and for the sick group in scenario 2 is 1 hour 43 minutes 26 seconds. The purpose of dividing the dataset is to reduce the time to process all text data in a set which takes longer. The following is a representation of the vector form of Medical Abstract, Diagnosis and Disease Group.

A. Implementasi Algorithm

In implementing the algorithm in this research using the Naïve Bayes Algorithm, XGBoost and Support Vector Machine with two scenarios of the Natural language Processing model for each algorithm and using two different types of datasets, namely the imbalance dataset and the sampled dataset so that they are balanced. Naïve Bayes is the simplest form of Bayesian network, where all attributes do not depend on the value of the class variable. Naïve Bayes' advantage is an effective and efficient machine learning algorithm. XGBoost is a tree method which is similar to a gradient booster but with the advantages of faster and more accurate computing. Support Vector Machine is a method that uses a hyperplane to group data based on its distance to the existing hyperplane. In the Support Vector Machine, there are several elements such as parameters and kernels that affect the results of the SVM.

B. Research Result

At the stage of implementing this algorithm, it is carried out in a predetermined scenario and the results can be seen in **Table 8** and **Table 9**. The dataset used is based on the previously processed NLP and the Combined Dataset is a form of combining all NLP results in satu dataset.

C. Comparison of Research Results

In the previous research, it was only limited to the medical abstract dataset, so the desired results might not be optimal and could still be developed again.

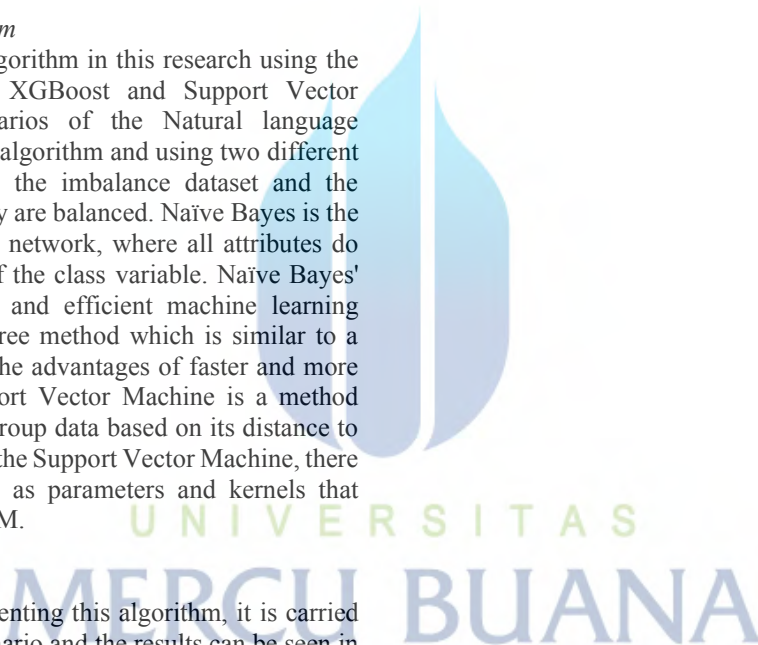


Table 8. Total Random Split Test Results on the Dataset

Dataset	Skenario	Random Split Test AVG Total (%)					
		Naïve Bayes	XGBoost	SVM-Linear	SVM-RBF	SVM-Polynomial	SVM-Sigmoid
Abstrak Medis	Skenario 1	71	84.84	89.33	91	86.67	88
	Skenario 1 Upsample	69	69	75	97.33	92	49.67
	Skenario 2	68.33	85.35	88.33	91	91	87.67
	Skenario 2 Upsampel	68.33	94.13	75	97.33	93	54.33
Diagnosis	Skenario 1	50	84.76	95	91	91	89.33
	Skenario 1 Upsample	67.67	98.01	96.67	97.33	97.33	74
	Skenario 2	49.33	84.52	95.67	81	89.33	89.33
	Skenario 2 Upsampel	70.67	98.01	75	97.33	93	54.33
Golongan sakit	Skenario 1	50	85.79	97.33	91	93	91.33
	Skenario 1 Upsample	80.67	97.76	96.33	98.33	98.33	65.33
	Skenario 2	84.33	84.52	96	91	92.67	91.33
	Skenario 2 Upsampel	80.67	98.16	96.33	98.33	98.33	71
Dataset Gabungan	Skenario 1	71.33	99.11	95.67	91	93.33	93.33
	Skenario 1 Upsample	80	99.31	98.33	100	98.67	49.67
	Skenario 2	71.67	99.11	96.33	91	93.33	93
	Skenario 2 Upsampel	81	99.37	99	100	99	75.33

Table 9. Total K-Fold Test Result on the Dataset

Dataset	Skenario	K-Fold Test AVG Total (%)					
		Naïve Bayes	XGBoost	SVM-Linear	SVM-RBF	SVM-Polynomial	SVM-Sigmoid
Abstrak Medis	Skenario 1	61.44	88.23	88.48	90.82	90.75	86.8
	Skenario 1 Upsample	63.6	75.39	62.2	94.02	86.58	57.17
	Skenario 2	61.98	87.8	88.11	90.82	90.78	86.74
	Skenario 2 Upsampel	64.35	75.84	61.97	94.28	86.92	55.19
Diagnosis	Skenario 1	39.11	88.29	97.47	90.87	90.83	87.37
	Skenario 1 Upsample	68.34	97.04	93.38	96.24	96.24	58.28
	Skenario 2	48.31	87.95	97	90.87	90.83	87.39
	Skenario 2 Upsampel	71.37	96.86	93.26	96.06	95.97	60.05
Golongan sakit	Skenario 1	65.11	87.58	86.84	90.87	90.85	88.47
	Skenario 1 Upsample	79.35	96.29	94.22	97.51	97.02	59.11
	Skenario 2	65.17	88	96.75	90.87	90.84	88.35
	Skenario 2 Upsampel	78.36	96.24	94.25	97.57	96.87	64.01
	Skenario 1	69.26	98.94	96.66	90.82	90.84	89.4

Dataset Gabungan	Skenario 1 Upsample	77.22	99.23	98.58	99.39	98.46	57.17
	Skenario 2	69.26	98.99	96.72	90.82	90.82	89.29
	Skenario 2 Upsampel	80.47	99.31	98.8	99.35	98.55	56.25

Table 10. Comparison of Average Value

Model Scenario	AVG Accuracy Total (%)						
	Naive Bayes	ANN	XGBoost	SVM-Linear	SVM-RBF	SVM-Poly	SVM-Sigmoid
Scenario 1 *)	67.05	89.82	-	-	-	-	-
Scenario 2 *)	67.12	90.27	-	-	-	-	-
Scenario 1 **)	66.22	-	86.54	88.91	90.91	90.8	87.4
Scenario 2 **)	65.16	-	86.58	88.22	90.91	90.89	90.89
Scenario 1 Upsampling	66.3	-	84.5	68.6	95.68	89.29	53.42
Scenario 2 Upsampling	66.34	-	84.99	68.49	95.81	86.96	54.76

information **Table 10**:

*) : Results of Related Experiments

**): Results of the experiments performed

- : No experimental data

In **Table 10** This is the result of managing the same dataset but with different algorithms. In previous studies, the use of ANN had the highest average results in the medical abstract dataset, but in the research that had been done it produced a different comparison with the Upsampling scenario which made the Upsampling results superior to the SVM RBF kernel algorithm with 95.81% better accuracy on the medical abstract dataset. And the research conducted also shows that other attributes such as diagnosis and sickness groups can be used as attributes that can be processed in NLP.

IV. CONCLUSION

This paper presents the results of applying a combination of NLP, Naïve Bayes, XGBoost and SVM learning to predict the type of disease based on a doctor's diagnostic dataset. This research shows the results that unbalanced data can affect the results of each processing method including the total average result of scenario 1 such as XGBoost which has an average of 89.69% and after Up sampling it is 94.59% which shows a 4.9% increase in accuracy, and the SVM kernel RBF the total average result of scenario 1 has an average of 90.92% and after Up sampling it is 97.52% which shows a 6.6% increase in accuracy. Comparing the accuracy of each dataset that has been made has different accuracy, this indicates that the data contained in the dataset such as unstructured data can affect the accuracy of the final result of a method. The average effectiveness of using the up sampling method on this unbalanced data is 2.72%. The use of the XGBoost

method is considered effective because it has the highest average accuracy of 92.13%, compared to the average of other methods such as Naïve Bayes with a total accuracy of 69.22% and SVM with a total accuracy of all kernels of 88.49%. Thus it can be concluded that the use of the XGBoost method is very suitable for use in the word2vector dataset processing in each scenario.

REFERENCE

- [1] E. D. Grober and J. M. A. Bohnen, "Defining medical error," *Can. J. Surg.*, vol. 48, no. 1, pp. 39–44, 2005.
- [2] P. S. Roshanov et al., "Computerized clinical decision support systems for chronic disease management: A decision-maker-researcher partnership systematic review," *Implement. Sci.*, vol. 6, no. 1, pp. 1–17, 2011.
- [3] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 151, 2018.
- [4] F. E. Prabowo and A. Kodar, "Analisis Prediksi Masa Studi Mahasiswa Menggunakan Algoritma Naïve Bayes," *J. Ilmu Tek. dan Komput.*, vol. 3, no. 2, p. 147, 2019.
- [5] B. A. Alyoubi, "Decision Support System and Knowledge-based Strategic Management," *Procedia Comput. Sci.*, vol. 65, no. Iccmit, pp. 278–284, 2015.
- [6] Guardian Y. Sanjaya, S. Harry, L. Lazuardi, and N. Faizah, "Datamining pereseapan elektronik di pelayanan kesehatan primer: potensi pengembangan sistem pendukung keputusan klinis," *Semin. Nas. Inform. Medis*, no. September, pp. 26–30, 2012.
- [7] J. K. Abdul Aziz Priatna, Rani Megasari, "Penerapan Association Rules Menggunakan Algoritma Apriori

- Pada Sistem Rekomendasi Pemilihan Resep Obat Berdasarkan Data Rekam Medis,” ... J. Apl. dan ..., vol. 1, no. 2, pp. 55–60, 2018.
- [8] C. I. Ratnasari, S. Kusumadewi, and L. Rosita, “Model Natural Language Processing untuk Perumusan Keluhan Pasien,” *Semin. Nas. Inform. Medis V*, pp. 11–18, 2014.
- [9] N. Indrawati, “Natural Language Processing (NLP) Bahasa Indonesia Sebagai Preprocessing pada Text mining ;,” no. 1, 2010.
- [10] T. F. M. Raj and S. Prasanna, “Implementation of ML using naïve bayes algorithm for identifying disease-treatment relation in bio-science text,” *Res. J. Appl. Sci. Eng. Technol.*, vol. 5, no. 2, pp. 421–426, 2013.
- [11] I. Faisal Hamzah, E. Kumala Dewi, and Suparno, “Makna Sakit Pada Penderita Penyakit Jantung Koroner : Studi Fenomenologis,” *J. Psikol. Undip Vol.13*, vol. 13, no. 1, pp. 1–10, 2014.
- [12] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, “Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifier Dengan Seleksi Fitur Dan Boosting,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 227–232, 2019.
- [13] Y. D. Prabowo, T. L. Marselino, and M. Suryawiguna, “Pembentukan Vector Space Model Bahasa Indonesia Menggunakan Metode Word to Vector,” *J. Buana Inform.*, vol. 10, no. 1, p. 29, 2019.
- [14] W. I and S. S. U. Rahman S, “Treatment of Missing Values in Data Mining,” *J. Comput. Sci. Syst. Biol.*, vol. 09, no. 02, pp. 51–53, 2015.
- [15] P. Liu, E. El-Darzi, L. Lei, C. Vasilakis, P. Chountas, and W. Huang, “An analysis of missing data treatment methods and their application to health care dataset,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3584 LNAI, pp. 583–590, 2005.
- [16] Irwan budiman, M. R. Faisal, and D. T. Nugrahadi, “Studi Ekstraksi Fitur Berbasis Vektor Word2Vec pada Pembentukan Fitur Berdimensi Rendah,” *J. Komputasi*, vol. 8, no. 1, pp. 62–69, 2020.
- [17] X. Rong, “word2vec Parameter Learning Explained,” pp. 1–21, 2014.
- [18] B. Jang, I. Kim, and J. W. Kim, “Word2vec convolutional neural networks for classification of news articles and tweets,” *PLoS One*, vol. 14, no. 8, pp. 1–20, 2019.
- [19] J. Luengo, A. Fernández, S. García, and F. Herrera, “Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling,” *Soft Comput.*, vol. 15, no. 10, pp. 1909–1936, 2011.
- [20] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A review,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 176–204, 2015.
- [21] I. A. M. SUPARTINI, I. K. G. SUKARSA, and I. G. A. M. SRINADI, “Analisis Diskriminan Pada Klasifikasi Desa Di Kabupaten Tabanan Menggunakan Metode K-Fold Cross Validation,” *E-Jurnal Mat.*, vol. 6, no. 2, p. 106, 2017.
- [22] F. Tempola, M. Muhammad, and A. Khairan, “Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018.
- [23] G. Parthiban, A. S.K.Srivatsa, and A. Rajesh, “Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method,” *Int. J. Comput. Appl.*, vol. 24, no. 3, pp. 7–11, 2011.
- [24] K. Vembandasamy, R. Sasipriya, and E. Deepa, “Heart Diseases Detection Using Naive Bayes Algorithm,” *Int. J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, pp. 441–444, 2015.
- [25] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016.
- [26] W. Li, Y. Yin, X. Quan, and H. Zhang, “Gene Expression Value Prediction Based on XGBoost Algorithm,” *Front. Genet.*, vol. 10, no. November, pp. 1–7, 2019.
- [27] K. Polat and S. Güneş, “A new feature selection method on classification of medical datasets: Kernel F-score feature selection,” *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10367–10373, 2009.
- [28] I. B. Aydilek, “Examining Effects of the Support Vector Machines Kernel Types on Biomedical Data Classification,” 2018 *Int. Conf. Artif. Intell. Data Process. IDAP 2018*, 2019.
- [29] A. Goel and S. K. Srivastava, “Role of kernel parameters in performance evaluation of SVM,” *Proc. - 2016 2nd Int. Conf. Comput. Intell. Commun. Technol. CICT 2016*, pp. 166–169, 2016.
- [30] J. Novakovic and A. Veljovic, “C-support vector classification: Selection of kernel and parameters in medical diagnosis,” *SISY 2011 - 9th Int. Symp. Intell. Syst. Informatics, Proc.*, pp. 465–470, 2011.

WORKING PAPER

Summary

This working paper is a material for completing the journal article entitled “The Influence of Using *Up Sampling* Method in Predicting Patients' Disease Types using a Combination of *Natural Language Processing*, *Naive Bayes* Algorithm, *XGBoost* and *Support Vector Machine*”. The working paper contains all the research materials of the Final Project which are not published / or included in journal articles. In this paper, the following sections are presented :

1. Literature Review is a section that contains the results of literature studies carried out related to the experiments carried out. Broadly speaking, the literature review conducted on the concept of Natural Language Processing, Data mining, Naïve Bayes Algorithm, XGBoost, and Support Vector Machine, the effect of data imbalance, and literature on types of disease.
2. Analysis and design is a part that consists of an outline and the stages carried out in this study. At this stage it is explained that the research is carried out using 2 scenarios. In the first scenario the Word2vec model training process in the first scenario only uses the Wikipedia corpus as the dataset. The second scenario is the Word2vec model training process. The second scenario does not only use the Wikipedia corpus as a dataset. But the wikipedia corpus is combined with a medical abstract as a dataset.
3. The source code in this study is in the form of database processing and the use of the Python programming language. Database processing here is the process of retrieving and cleaning data so that it can be used in this study. The use of Python in this study is used to train the Word2Vec model, process data from modeling to training and implement the Naïve Bayes algorithm, XGBoost and the Support Vector Machine.
4. Dataset is a part that explains what data is used in the experiment. This section describes the structure of the initial dataset, the treatments performed and the results of the representation of medical abstracts, diagnoses, and sickness groups into vector form.
5. Experimental Stages is a section that contains all experimental stages that are not included in the journal. This section outlines the overall technical flow of the research. The stages described in this section include the stages of data collection, data cleaning and treatment, conversion of text data, data splitting, Up Sampling Dataset, implementation of the Naïve Bayes

algorithm, XGBoost, Support Vector Machine and evaluation and comparison of scenarios.

6. Results All Experiments is a part consisting of the results of the experiment carried out, the comparison of the results of each scenario. The experiments carried out included the Naïve Bayes Algorithm, XGBoost, Support Vector Machine with four kernels, namely Linear, RBF, Polynomial and Sigmoid with Random split and Cross validation methods.

