UNIVERSITAS
MERCU BUANA
TERAKREDITASI-A

IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSTIAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

THESIS REPORT

FANDY NURRAHMAN
41517010021

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

IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSTIAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

THESIS REPORT

Submitted to Complete Terms
Completed a Computer Bachelor Degree

Created By:

FANDY NURRAHMAN
41517010021

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

ORIGINALITY STATEMENT SHEET

The undersigned below:

Student Number	: 41517010021
Name	: Fandy Nurrahman
Title	: IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

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

Jakarta, 25th March 2021

SEPULUH RIBU RUPIAH 10000 METERAI TEMPEL 4070EAJX101053871

Fandy Nurrahman

FINAL PROJECT PUBLICATION STATEMENT

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

Student Name : FANDY NURRAHMAN

Student Number : 41517010021

Title : IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

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 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 Mercy 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 authot/creator and Copyright owner.

Hereby I made this statement in truthfulness.

Jakarta, 25th March 2020

SEPULUH RIBU RUPIAH
10000
METERAI TEMPEL
D4A9BAJX101053866

Fandy Nurrahman

FINAL PROJECT OUTPUT STATEMENT LETTER

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

Student Name	:	FANDY NURRAHMAN
Student Number	:	41517010021
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

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 Reputeable International Journal		Accepted	
		Reputeable International Journal	√		
	Submitted / Published:	Journal Name	: International Journal of Data Analysis Techniques and Strategies (IJDATS)		
		ISSN	:		
		Journal Link	: bit.ly/JournalIJDATS		
		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

Jakarta, 19 March 2020

10000 METERAI TEMPEL 620FBAJX101053861

Desi Ramayanti, S.Kom., MT

Fandy Nurrahman

EXAMINER APPROVAL SHEET

Student ID	:	41517010021
Student Name	:	FANDY NURRAHMAN
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

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

Jakarta, 19th March 2021

Approved,

(Dr. Mujiono Sadikin, MT)

EXAMINER APPROVAL SHEET

Student ID	:	41517010021
Student Name	:	FANDY NURRAHMAN
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

This thesis has been examined and tried as one of the requirements to obtain a Bachelor’s degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 19th March 2021

Approved,

(Dr. Leonard Goeirmanto)

(Dr. Leonard Goeirmanto, M.Sc)

EXAMINER APPROVAL SHEET

Student ID	:	41517010021
Student Name	:	FANDY NURRAHMAN
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

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

Jakarta, 19th March 2021

Approved

(Dr. Ida Nurhaida, M.T.)

COMMITTEE APPROVAL SHEET

Student Number	:	41517010021
Student Name	:	FANDY NURRAHMAN
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

This thesis has been examined and tried as one of the requirements to obtain a Bachelot's degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, March 9th 2021

Approved,

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

(Diky Firdaus, S.Kom, MM)
Defense Committee 1

(Desi Ramayanti, S.Kom, MT)
Defense Committee 2

VALIDITY SHEET

Student Number	:	41517010021
Student Name	:	FANDY NURRAHMAN
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

This Final Project has been examined and defenced as one of the requirements for obtaining a Bachelor's degree in the Informatics Engineering Study Program, Faculty of Compuer Science, Universitas Mercu Buana.

Jakarta, March 9th 2021

Approved,

(Desi Ramayanti, S.Kom, MT)
Thesis Supervisor

Acknowledged,

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

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

ABSTRAK

Nama	:	FANDY NURRAHMAN
NIM	:	41517010021
Pembimbing TA	:	Desi Ramayanti, S.Kom., MT.
Judul	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSITAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

Abstrak: Untuk meningkatkan kualitas universitas bersamaan dengan akreditasi yang baik ada hal yang perlu diperhatikan dengan memanfaatkan data alumni. Data alumni yang diperoleh 2 tahun setelah mereka lulus dapat dimanfaatkan secara optimal untuk memprediksi berapa lama mahasiswa mendapatkan pekerjaan setelah mereka lulus. Beberapa atribut dari data yang bisa digunakan yaitu tahun yudisium. GPA, dan juga kategori lama, sedang ataupun cepat alumni mendapat kerja bisa menjadi bahan untuk prediksi menggunakan klasifikasi. Penelitian ini menunjukkan hasil dari penggunaan algoritma Naïve Bayes Classifier (NBC) untuk melatih dan menguji data dalam mengklasifikasi lama waktu yang ditempuh mahasiswa berdasarkan data yang telah diambil dari program UMBCTC yaitu Tracer Study 2015, 2016, dan 2017. Dari hasil metode NBC yang diperoleh data tersebut akan divalidasi menggunakan K-Fold Cross Validation. Dengan akurasi yang dihasilkan oleh NBC yaitu 90% dan rata-rata K-Fold Cross Validation adalah 82,81%

Kata kunci:
Data Mining, Naïve Bayes Classifier, K-Fold, Cross Validation

ABSTRACT

Student Name	:	Fandy Nurrahman
Student Number	:	41517010021
Counsellor	:	Desi Ramayanti, S.Kom, MT
Title	:	IMPLEMENTATION OF NAÏVE BAYES ALGORITHM IN PREDICTING THE LENGTH OF TIME FOR UNIVERSTIAS MERCU BUANA ALUMNI TO GET A JOB AFTER GRADUATED

Abstract: Improving the quality of the university to have great accreditation, there are several things that need to be considered, one of them is utilizing alumni data. Alumni data obtained 2 years after they graduate can be used optimally to predict how long students get a job after they graduate. Attributes of the data that are being used are gender, judicial year. GPA, as well as the label of short, mid, or fast alumni getting a job, these attributes could be processed for prediction using classification. This research shows the results of using the Naïve Bayes Classifier (NBC) algorithm to train and test data in classifying the length of time taken by students based on data taken from the UMBCTC program, Tracer Study 2015, 2016, and 2017. From the results of the NBC method, data obtained will be validated using the K-Fold Cross Validation. The accuracy generated by NBC is 90% and the average K-Fold Cross Validation is 82.81%

Keywords:
Data Mining, Naïve Bayes Classifier, K-Fold, Cross Validation

PREFACE

Praise our gratitude for the presence of Allah SWT, because with His grace & guidance author could complete this thesis report, as a condition for completing the Bachelor degree (S1) in Informatika Engineering at 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 psossible to:

1. Dr. Ngadino Surip as a Chancelolor of Universitas Mercu Buana who has provided many changes and positive progress for our university.
2. Dr. Mujiono Sadikin, MT. as Dean of Faculty of Computer Science Universitas Mercu Buana
3. Ibu Desi Ramayanti, S.Kom., M.T. as 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.
4. Mrs. Prastika Indriyanti, S.Kom., M.Cs as Head of International Informatics Department of Universitas Mercu Buana
5. Lecturer of the Informatics Department for the knowledge, dedication, and motivation given during the lecture period.
6. Universitas Mercu Buana staff who have provided invaluable assistance for the author to complete this thesis
7. Parents, Family & Bestfriends who always pray & support the author in completing the final project.
8. Classmates from Informatics English Instructed Class who has been together for these 3 years and keep motivate the author to complete this thesis report.

In writing this thesis, the author realizes that this is not perfect yet, therefore constructive criticism and suggestions from all people are expected. Hopefully this thesis report could increase the knowledge for involved parties. The author would like to thank you very much for the guidance and all the support given, may Allah SWT bestow His mercy and gifts.

Jakarta,
Fandy Nurrahman

TABLE OF CONTENTS

JOURNAL

Implementation of Naïve Bayes Algorithm in Predicting the Length of Time for Universitas Mercu Buana Alumni to Get a Job After Graduated

Fandy Nurrahman

Informatics Department,
Universitas Mercu Buana, Jakarta
Email: 41517010021@student.mercubuana.ac.id

Abstract: Alumni data could be utilized for improving university quality, decision-making, and future research purposes. Alumni data obtained 2 years after they graduate can be used optimally to predict how long students get a job after they graduate using classification. Attributes of the data that are being used are gender, judicial year and GPA. This research using Naïve Bayes Classifier (NBC), a classification algorithm for predicting the value based on the available train data which is decently accurate, and also NBC has been used for other purposes such as spam detection. Results of using the NBC algorithm in this research to train and test data in classifying the length of time taken by alumni to get a job. Data retrieved by Tracer Study 2015 to 2017 program. The results of the NBC validated using the K-Fold Cross-Validation. The accuracy generated by NBC is 90% and the average K-Fold Cross-Validation is 82.81%.

Keywords: data mining, naïve bayes classifier, k-fold, cross validation.

Reference -

Biographical notes: Fandy Nurrahman is an Informatics in Universitas Mercu Buana, Computer Science Department.

1 Introduction

One of the benchmarks for the quality of a University is the alumni who are beneficial to society, as stated in the third point of the tri dharma of higher education, that is "Community Service" (Lian, 2019), in this mission improving quality along with the goal to increase university rankings and accreditation, is a must for producing the best quality of the University's alumni.

With the quality development purpose in this technological era, one thing for certain is the need for data, data holds an important role for development in many various sectors, for University that has more specific data like gender, judicial year, Grade Point Average (GPA), and value that categorize long, medium, and short amount of time of a student that got a job after they graduated. Alumni data that has been mentioned require further analysis in order to find new information and patterns using Data Mining.

Data mining is a Knowledge Discovery in Database (KDD) which has a function to find new patterns from a dataset and it's useful for making decisions based on the result of the new pattern or knowledge.

KDD procedure itself is being done in several steps such as data collection, data transformation, pattern evaluation, and the presentation of the analyzed data result (Agarwal, 2014).

There are several types of concepts and techniques contained in Data Mining, namely generalization, characterization, classification, and association, all of them need to be adjusted based on the need of data that we have (Liao, Chu, & Hsiao, 2012).

Data that has been obtained from the Career & Training Center unit in Universitas Mercu Buana was gathered by the student that has filled the questionnaires. Collected data still in the form of raw data that has not been collected and cleaned. therefore it has to run the cleaning process first

Some of the objectives that want to be achieved in this research are predicting the estimated time or how long alumni of Universitas Mercu Buana got a job after they graduated in an undergraduate program (S1), and wanting to know how accurate the use of Naïve Bayes Classifier (NBC) algorithm dan K-Fold Cross Validation is. Naive Bayes is a classification algorithm that uses the simple probability based on the Bayes theorem that assumes or considers each class in a dataset is independent (Syarli & Muin, 2016). The algorithm will be followed by K-Fold Cross-Validation, CV is a technique to validate or measure the accuracy of a model in a dataset that is being tested by dividing it into 2 parts, test-set, and training-set.

2 Research Methodology

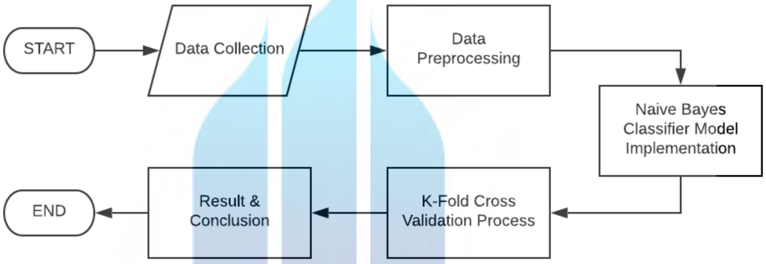

START
Data Collection
Data Preprocessing
Naive Bayes Classifier Model Implementation
K-Fold Cross Validation Process
Result & Conclusion
END

Figure 1. Research Stages

This study uses a classification method, the Naive Bayes classifier, through the process of calculating the frequency from the dataset and focusing on classifications on certain attributes. The result of the classifier is measured by predictive accuracy (Kusumadewi, 2009), a literature study on the naïve bayes classifier implementation has been carried out in several studies such as that has been done at the University of Muhammadiyah Yogyakarta which resulted in accuracy of 71% (Asroni, Maharty Ali, & Riyadi, 2018), and Universitas Islam Indonesia's accuracy of 71,76%. (Amrinda, 2018). Followed by random sampling, random sampling is implemented to set a condition where data is taken from a dataset randomly (Akhmad, Adikara, & Wihandika, 2019) this function is to set a condition in a certain amount of sampling data that later will be divided into training and test sets. After knowing the accuracy results obtained by the naive bayes classifier model, it will be validated using the k-fold cross-validation technique. The process will be carried out on google's cloud tools, namely, google colab.

Data Collection

The data collection used in this study was obtained from the Universitas Mercu Buana Career & Training Center in the Tracer Study program where alumni who have graduated two years after passing the trial must fill out a questionnaire provided by the campus. In this data, there are 141 columns showing details about 7,664 data records taken during the graduation period in 2015, 2016, and 2017. The data that has been collected will be submitted to DIKTI, and for the campus, it will be processed for data collection which will be displayed for accreditation. The selected data will later show several attributes that will be calculated for processing, these attributes are shown in **Table 1**.

Table 1. Tracer Study Data Structure

Attributes	Data Types	Value Range
Gender	Text	Text
Std_program	Varchar	Ordinal
Std_yudisium	Numerical	Continue
GPA	Integer	Continue
Month_job	Numerical	Continue
Label	Numerical	Continue

Description:

- Gender : with a text data type that has a male or female value
- Continue Range Value: contains numerical values
- Ordinal Range Value: integer value that represent study programs
- Std_program : ordinal data type that has 16 study programs
- Std_yudisium : numerical data type that has a range of values "2014, 2015, 2016"
- GPA : have a wide range of integer values
- Month_job : It has a wide range of numerical values
- Label : It has a numerical value range of "0" for the fast category, "1" for the medium category, and "2" for the old category.

Table 2. Study Program List

FACULTY	STUDY PROGRAM
FEB	Manajemen
	Akuntansi
FIKOM	Komunikasi Digital
	Hubungan Masyarakat
	Periklanan
	Penyiaran
FT	Teknik Sipil
	Teknik Industri
	Teknik Arsitektur
	Teknik Elektro
	Teknik Mesin

FASILKOM	Teknik Informatika
	Sistem Informasi
FDSK	Desain Produk
	Desain Komunikasi Visual
	Desain Interior

Data Preprocessing

Data preprocessing is a step that must be done in order to improve the quality of the data which will affect the final result of the accuracy of the implemented model, one of the techniques used is data cleaning, this technique has purposed to clean data, eliminate some data in attributes such as noise, blank data, or data that is inconsistent or irrelevant.

After cleaning and selecting the relevant attributes, the next step followed by removing blanks records, in Tracer Study dataset, has 7664 data records including Regular 2 / Employee Class program data, many of these Employee Class data have worked status before graduation, and blank data. From these data, this data process focuses on students of Regular / Regular Class 1 or full-time classes. After going through this cleaning stage, it produces the required data, 2825 data records and 6 attributes, namely gender, std_program, std_yudisium, gpa, month_job, (how many months after graduating to get a job), and labels that categorize the how length of time, in short, medium, and long with which alumni get jobs, with data from 2825 data. A total of 2172 data are student data with short descriptions, 361 student data with medium descriptions, and 292 student data with long descriptions.

Table 3. Normal Data on Label Attribute

0	2172
1	361
2	272

The result of this cleaning process causes imbalance data which has an impact on the experimental model that being applied so that the data becomes biased, and focuses on 'fast' category only, to overcome the accuracy performance problem it will use oversampling technique using upsampling method (Syukron & Subekti, 2018) which duplicates the data categories from undervalue data; medium and long categories, the data will be duplicated until the it has the same amount of 2172 data. which resulting 6516 the total upsampled data.

Table 4. Upsampled Data

0	2172
1	2172
2	2172

The data cleaning process is also followed by data conversion, a method that changes the format of one data type in an attribute to another data type format, in this stage the data is will be processed manually from the text data type in the gender attribute and the varchar data type on the std_program attribute in the Microsoft Excel worksheet using the filter feature.

Naïve Bayes Algorithm

Naïve Bayes is a statistical classification that can be used to predict the probability of a class. Based on the Bayes classification theorem, the capabilities of statistical classification are used to predict the probability of data that has been used and is shown to have sufficient accuracy to predict a data set (Vembandasamy, Sasipriya, & Deepa, 2015). naive bayes classifier is often used for classification in various machine learning-related studies such as Twitter-based traffic (Dabiri & Heaslip, 2019), *Web-ads detection* (Shaqoor Nengroo & Kuppusamy, 2018), *Predict complications in kid's ingestion* (Berchialla, Foltran, & Gregori, 2013), etc. In the probability formula as follows:

$$\mathrm{P}(\mathrm{C}k \mid \mathrm{X}) = \frac{P(Ck)P(X \mid Ck)}{P(X)} \quad (1)$$

Figure 2. *Bayesian Probability Formula*

Description:
X : Data with unknown classes
Ck : Hypothesis of Data X refers to a specific class
P(Ck|X) : Probability of data Ck based on the hypothetical condition X
P(Ck) : Probability hypothesis of C
P(X|Ck) : Probability of data X based on the hypothetical condition Ck
P(X) : Probability of X

The above probability shows the conditional opportunity attribute X given by class C, while in Naïve Bayes, attribute X can be a qualitative or quantitative attribute. Therefore, when attribute X is quantitative, the probability is very small, making the probability equation unreliable for quantitative attribute type problems. To overcome the quantitative attributes, this research will be using the Gaussian Naïve Bayes normal distribution which is more specific to know the average probability of each supporting attribute against the predicted target attribute.

$$(x = v \mid Ck) = \frac{1}{\sqrt{2\,\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \qquad (2)$$

Figure 3. *Gaussian Naïve Bayes Formula*

Description:
σ^2 = deviation standard
μ = mean (average)

To make the data more accurate in the naïve bayes, naive bayes classifiers will use a splitting method in the initial process that will be carried out and there are 3 types of division that will be done in this research, first with 10%, 20% and 30% sampling scenarios.

Attributes Correlation

In the inter-attribute analysis, the methods will be using the Pearson and Spearman methods, both methods will present a correlation of both discrete and continuous data. The Pearson correlation coefficient is used to evaluate the closeness of the linear correlation between 2 or more variables (Fu et al., 2020) and has a greater range of data volume calculations than Spearman. In the attribute analysis, the Tracer Study (TS) data that had been cleaned resulting correlated attributes with the targeted Label attribute as shown in **Table 5.**

Table 5. Correlation Results Towards Label Attribute

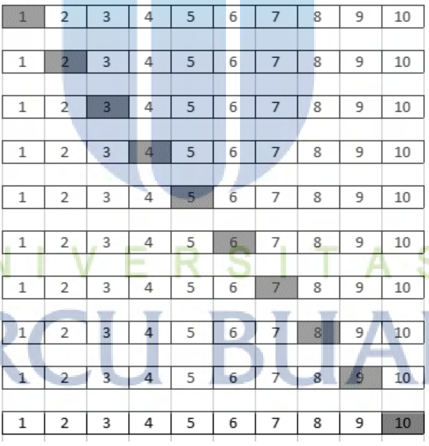
	Pearson	**Spearman**
gender	-0,066	-0,066
std_program	0,037	0,035
std_yudisium	0,032	0.036
gpa	0,017	-0,056
month_job	0,038	0,984

The results of the gender correlation with the label attribute in the Pearson and Spearman correlation method have a low value or only intersect without any significant contribution to the label attribute, in contrast to other attributes which tend to be positive, which means that there is a significant relationship to the label attribute.

Cross Validation

Cross Validation or what is also called rotation estimation is an important parameter because by the way this method works, it will take another sample for the evaluation of the algorithm that has been run, also K-Fold itself is a type of non-exhaustive cross validation. (Mileman, 2001) which the way it works start to split the data to perform a predetermined number of iterations which has been set in the K parameter.

Figure 4. K-Fold 10



One of the K-Fold CVs that is often used is the 10-fold, because it tends to provide unbiased estimation accuracy. In the 10-fold, it means that you will repeat iterations 10 times, which every step of iteration consists 9 part of the data is for training and 1 part for testing. In the process of K-Fold cross validation, the existing data will be tested by 3 types of k-fold with iterations, those are 5-fold, 10-fold, and 15-fold to get results that can be compared to the prediction data that has been carried out by Naïve Bayes.

3 Result & Discussion

In research to find out the predictions of the estimated length of time needed to find a job using the Naïve Bayes Classifier model, where the results of the data classification of the method will be followed by the validation process using K-Fold, the result of both processes will be calculated to find out the average using python programming language on the google colab cloud tools. The first process is the naïve bayes, which divide data into 2 parts of training and testing set will be processed 3 times testing with different data sampling sizes.

The first naïve bayes using a 10% sampling scenario from 6516 data resulting in an accuracy of 90%.

Table 6. Results from 10% data sampling

10		P		
A		0	1	2
	0	228	0	0
	1	0	203	0
	2	0	67	154

Table 7. Classification report table of 10% sampling

	Precision	Recall	F1-Score	Support
0	1,00	1,00	1,00	228
1	0,75	1,00	0,86	203
2	1,00	0,70	0,82	221
Accuuracy			0,90	652
Macro Avg	0,92	0,90	0,89	652
Weighted Avg	0,92	0,90	0,90	652

The second naïve bayes, using a 20% sampling scenario from 6516 data resulting in an accuracy of 90%

Table 8. Results from 20% data sampling

20		P		
A		0	1	2
	0	440	0	0
	1	0	415	0
	2	0	135	314

Table 9. Classification report table of 20% sampling

	Precision	Recall	F1-Score	Support
0	1,00	1,00	1,00	440
1	0,75	1,00	0,86	415
2	1,00	0,70	0,82	449
Accuuracy			0,90	1304
Macro Avg	0,92	0,90	0,89	1304
Weighted Avg	0,92	0,90	0,89	1304

The second naïve bayes, using a 30% sampling scenario from 6516 data resulting in an accuracy of 90%

Table 10. Results from 30% data sampling

30		P		
A		0	1	2
	0	657	0	0
	1	0	639	0
	2	0	188	471

Table 11. Classification report table of 30% sampling

	Precision	Recall	F1-Score	Support
0	1,00	1,00	1,00	228
1	0,75	1,00	0,86	203
2	1,00	0,70	0,82	221
Accuuracy			0,90	1955
Macro Avg	0,92	0,90	0,90	1955
Weighted Avg	0,93	0,90	0,90	1955

As mentioned about the naive bayes theory, each of the attributes towards the target attribute will be processed and get their respective predictive results. Then the results will be validated using k-fold, at this stage k-fold uses three different number of iterations or folds, those are 5-Fold, 10-Fold, and 15-Fold to find the average accuracy results so that the validation is reliable with the results obtained as shown in **Table 12.**

Table 12. Average Accuracy K-Fold 5, 10 & 5 results

Folds	Average Accuracy (%)
5	71,23
10	87,34
15	89,87

From the results of each function that has been collected, the average calculation for the two models can be seen in the following table.

Table 13. Result & Average

	Sampling	Accuracy
NAÏVE BAYES	10%	90%
	20%	90%
	30%	90%
KFOLD CV	5	71,23%
	10	87,34%
	15	89,79%
Average	NB	90%
	KF CV	82,78%

The naive bayes method that uses the different random sampling split data from 10%, 20% and 30% sampling, resulting in 90% system accuracy from all sampling tests, with all three of the different sampling means that naive bayes method is already quite accurate for such dataset.

Author has looking for the relation among the other attribute towards the target attribute (label) in Python using Pearson and Spearman correlation and resulting -0.066 point which means the relation of gender attribute towards the label is not sufficient enough and does not have linear correlation at all, in other words gender attribute should not be added as the supporting attributes.

4 Conclusion & Suggestion

Based on the results of tests carried out with the naïve bayes classifier using the Python programming language on google colab, the average system accuracy is 90% and the validation carried out produces an average value of 82.78%, from the results after the data has been upsampled on this study shows that the use of the naïve Bayes classifier method and k-fold cross validation is reliable enough for a large enough data size.

Table 14. Comparison

UNIVERSITIES	NBC	KFOLD CV
UII	71,76%	
UMY	71%	

UMB	90%	82.81%

Comparison of several research that have been conducted shows that Gaussian naïve Bayes which is implemented through python programming language has fairly accurate results compared to the two-existing research.

Suggestion

This research could be improved by using two or more predictive algorithms to be compared and has more options to use for the best support system for the data.

References

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Akhmad, S., Adikara, P. P., & Wihandika, R. C. (2019). Analisis Sentimen Kebijakan Pemindahan Ibukota Republik Indonesia dengan Menggunakan Algoritma Term-Based Random Sampling dan Metode Klasifikasi Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(10), 10086–10094.

Amrinda G. D., (2018). Analisis Klasifikasi Waktu Tunggu Kerja Dengan Metode Support Vector Machine dan Naïve Bayes Classification. *151*(2), 10–17.

Asroni, A., Maharty Ali, N., & Riyadi, S. (2018). Perkiraan Masa Tunggu Alumni Mendapatkan Pekerjaan Menggunakan Metode Prediksi Data Mining Dengan Algoritma Naive Bayes Classifier. *Semesta Teknika*, *21*(2), 189–197. https://doi.org/10.18196/st.212225

Berchialla, P., Foltran, F., & Gregori, D. (2013). Naïve Bayes classifiers with feature selection to predict hospitalization and complications due to objects swallowing and ingestion among European children. *Safety Science*, *51*(1), 1–5. https://doi.org/10.1016/j.ssci.2012.05.021

Dabiri, S., & Heaslip, K. (2019). Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, *118*, 425–439. https://doi.org/10.1016/j.eswa.2018.10.017

Fu, T., Tang, X., Cai, Z., Zuo, Y., Tang, Y., & Zhao, X. (2020). Correlation research of phase angle variation and coating performance by means of Pearson's correlation coefficient. *Progress in Organic Coatings*, *139*(June 2019), 105459. https://doi.org/10.1016/j.porgcoat.2019.105459

Kusumadewi, S. (2009). *Klasifikasi Status Gizi Menggunakan*. *3*(1), 6–11.

Lian, B. (2019). Tanggung Jawab Tridharma Perguruan Tinggi Menjawab Kebutuhan Masyarakat. *Prosiding Seminar Nasional Pendidikan Program Pascasarjana Universitas PGRI Palembang*, *2*, 999–1015.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311. https://doi.org/10.1016/j.eswa.2012.02.063

Mileman, P. A. (2001). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artifical Intelligence (IJCAI)*, *30*(2), 133–133.

Shaqoor Nengroo, A., & Kuppusamy, K. S. (2018). Machine learning based heterogeneous web advertisements detection using a diverse feature set. *Future Generation Computer Systems*, *89*, 68–77. https://doi.org/10.1016/j.future.2018.06.028

Syarli, S., & Muin, A. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi). *Jurnal Ilmiah Ilmu Komputer*, *2*(1), 22–26.

Syukron, A., & Subekti, A. (2018). Penerapan Metode Random Over-Under Sampling dan Random

Forest Untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika*, *5*(2), 175–185. https://doi.org/10.31311/ji.v5i2.4158

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441–444.

WORKING SHEET

This working paper is a material for completing the journal article entitled "Implementation of Naive Bayes Algorithm in Predicting the Length of Time for Universitas Mercu Buana Alumni to Get a Job After Graduated". This working sheet contain all of the research materials of the Final Project which have not include yet 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 Data Mining, Naïve Bayes Classifier, K-Fold, Cross Validation, the effect of data imbalance, and literature on types of disease.

2. Analysis and design are parts of that consist of an outline and the stages carried out in this study. This stage using the training and testing data for the training using the Naive Bayes Algorithm.

3. The source code in this study is in the form of database processing and the use of the Python programming language in Google Colab. The use of Python in this study is used to train the available dataset and process it using the Naïve Bayes algorithm and cross validation.

4. The dataset explains overall data which in this case using Alumni data from 2015 to 2017.

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, data splitting, Up Sampling Dataset, implementation of the Naïve Bayes Algorithm and Cross Validation.

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, Random split and Cross validation methods.

Universitas Mercu Buana