

project report

comparison learning vector quantization and naïve bayes algorithm in airline passenger satisfaction

NIKODEMUS GALIH CANDRA WICAKSONO

18.K1.0082

Faculty of Computer Science

Soegijapranata Catholic University

2020

APPROVAL AND RATIFICATION PAGE (Heading plain)

(gunakan style “Approval”)

JUDUL PROJECT ANDA

by

NIKODEMUS GALIH CANDRA WICAKSONO – 18.K1.0082

This project report has been approved and ratified

By the Faculty of Computer Science on January, DD. YYYY

With approval,

|  |  |
| --- | --- |
|  | Supervisor  Dosen Pembimbing Project  NPP: ­ |
| Examiners,  1.)  [NAMA PENGUJI 1]  NPP :  2.)  [NAMA PENGUJI 2]  NPP :  3.)  [NAMA PENGUJI 3]  NPP : |  |
|  | Dean of Faculty of Computer Science  Robertus Setiawan AJi Nugroho, Ph.D  NPP: 058.1.2004.264 |

DECLARATION OF AUTHORSHIP

I, the undersigned:

Name : NIKODEMUS GALIH CANDRA WICAKSONO

ID : 18.K1.0082

declare that this work, titled "JUDUL ANDA", and the work presented in it is my own. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at Soegijapranata Catholic University
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given.
5. Except for such quotations, this work is entirely my own work.
6. I have acknowledged all main sources of help.
7. Where the work is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Semarang, MONTH, DD, YYYY

[sign here]

NIKODEMUS GALIH CANDRA WICAKSONO

18.K1.0082

ACKNOWLEDGMENT

Silahkan tuliskan anda ingin mengucapkan terima kasih atau ucapan persembahan ke siapapun yang anda rasa perlu ditulis disini

I have received a myriad of support, advice, and assistance throughout this document writing. I would like to thank my supervisors XXXXXX for formulating this topic. I would also like to thank my friend XXXXX for guiding with advice to finish this document.

I would like to thank my family and friends for giving me ceaseless love, support, and advices throughout my study in Soegijapranata Catholic University. You gave me great escape to rest my mind from my thesis.

ABSTRACT (Abstract Title)

Isi dari abstract menggunakan syle abstract content. Abstract ditulis dalam 3 paragraf. Semua ditulis dengan huruf italic dan 1 spasi Paragraf pertama berisi tentang permasalahan yang diselesaikan dalam project ini.

Paragraf kedua di sini, membahas tentang proses penyelesaian yang Anda tawarkan.

Sedangkan paragraf ketiga membahas tentang hasil akhir. Setelah itu di bagian paling bawah, sertakan keywords atau kata kunci 3-5 kata.

Keyword: kata\_kunci1, kata\_kunci2, kata\_kunci3, dst

\*Tambahkan informasi mengenai penelitian payung di sini apabila ada (konsultasikan dengan dosen pembimbing).

TABLE OF CONTENTS

COVER i

[APPROVAL AND RATIFICATION PAGE (Heading plain) ii](#_Toc83625616)

[DECLARATION OF AUTHORSHIP iii](#_Toc83625617)

[ACKNOWLEDGMENT iv](#_Toc83625618)

[TABLE OF CONTENTS vi](#_Toc83625619)

[LIST OF FIGURE viii](#_Toc83625620)

[LIST OF TABLE ix](#_Toc83625621)

[CHAPTER 1 INTRODUCTION 10](#_Toc83625622)

[1.1. Background 10](#_Toc83625623)

[1.2. Problem Formulation 11](#_Toc83625624)

[1.3. Scope 11](#_Toc83625625)

[1.4. Objective 11](#_Toc83625626)

[CHAPTER 2 LITERATURE STUDY 12](#_Toc83625627)

[CHAPTER 3 RESEARCH METHODOLOGY 15](#_Toc83625628)

[3.1. Data Collection 15](#_Toc83625629)

[3.2. Algorithm 15](#_Toc83625630)

[3.3. Coding and Design 15](#_Toc83625631)

[3.4. Analysis 16](#_Toc83625632)

[3.5. Make a Report 16](#_Toc83625633)

[CHAPTER 4 ANALYSIS AND DESIGN 17](#_Toc83625634)

[4.1. Analysis 17](#_Toc83625635)

[4.2. Penambahan Gambar, Judul Gambar, dan Penggunaan Gambar 18](#_Toc83625636)

[4.3. Tabel, Posisi, dan Isi tabel 19](#_Toc83625637)

[4.4. Design 20](#_Toc83625638)

[4.5. Function 21](#_Toc83625639)

[CHAPTER 5 IMPLEMENTATION AND RESULTS 23](#_Toc83625640)

[5.1. Implementation 23](#_Toc83625641)

[5.2. Results 23](#_Toc83625642)

[CHAPTER 6 CONCLUSION 24](#_Toc83625643)

[REFERENCES 25](#_Toc83625644)

[APPENDIX a](#_Toc83625645)

LIST OF FIGURE

[**Figure 4.1** Architecture LVQ 18](#_Toc83625542)

[**Figure 4.1** Program Flowchart 20](#_Toc83625543)

LIST OF TABLE

[**Table 5.1.** Tabel Analisis Data (Table caption) 2](#_Toc77239085)

# CHAPTER 1 INTRODUCTION

## Background

Passenger satisfaction is one of the important factors for the improvement of an airline. The airline can find out what things need to be improved. With the hope that more and more airplane passengers use the airline, of course this increase must be done so that income also increases. To improve service, of course, you must know what things make passengers satisfied. This can be done from the data of passengers who have traveled by plane.

In this digital era, data is very easy to store and obtain. Not like in the past, which used paper to record data, but used the help of computers. One of the advantages is that it is easy to store large amounts of data, including passenger satisfaction data. If there are about 130,000 airline passenger satisfaction data, of course it is very difficult to process manually. This will make it difficult for airlines to improve services.

Because data storage uses a computer, we can also use a computer to process it. However, to process the existing data in order to get the results we want, an algorithm is needed. With the algorithm implemented on passenger satisfaction data, we can classify things that can make passengers satisfied with airline flight services. Of course, this is better than processing thousands of data manually.

Therefore, this time I implemented the Learning Vector Quantization (LVQ) and Naïve Bayes algorithms on the airline passenger satisfaction data that I got through Kaggle. It is hoped that this algorithm can process thousands of existing data and classify them. I am using 2 different algorithms so that I can compare the results of each implemented algorithm. And also, to find out which algorithm is better for classifying airline passenger satisfaction data by comparing the accuracy of the two algorithms. The results of this classification algorithm are expected to help airlines know what to do in the future.

## Problem Formulation

From the background above, we can formulate the existing problems.

1. Can the Naïve Bayes algorithm classify airline passenger satisfaction data?
2. Can the Learning Vector Quantization algorithm classify airline passenger satisfaction data?
3. Based on the level of accuracy, which algorithm is better in classifying passenger satisfaction data?

## Scope

In this project, I applied Learning vector quantization and Naive Bayes algorithm only for the data I used from https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction with 129,880 data. The data consists of 23 measuring columns and 1 response column. To find out a better algorithm, I use the accuracy parameters of each algorithm. There will be 5 tests for each algorithm with a percentage of training data of 10%, 25%, 50%, 75%, and finally 90%.

## Objective

The purpose of this project is to find out whether the Learning Vector Quantization and Naive Bayes algorithms can classify aircraft passenger satisfaction from existing data. In addition, to find out from the two implemented algorithms, which algorithm is better based on the level of accuracy.

# LITERATURE STUDY

Gorzalczany et al. [1] explain that a lot of data mining does not provide deeper explanations and justifications than decisions. Therefore, they apply their knowledge discovery technique based on fuzzy rules to the problem of airline passenger satisfaction. They used a dataset from Kaggle of 259,760 records. With 23 variable columns, the dataset is almost the same as the dataset that I will use. The results obtained are that the most significant attribute is Inflight Entertainment with an accuracy of 75.2%. Followed by the attributes of Seat comfort and Inflight Wi-Fi Service. They do not classify, but can determine which classification variables affect airline passenger satisfaction more.

With the US Airlines dataset which is almost the same as before, Hayadi et al. [2] uses several classification algorithms. The algorithms used are KNN, Logistic regression, Gaussian NB, Decision Trees and Random Forest. The author runs using the GridSearchCV algorithm from Scikit-Learn. Of all the algorithms that have been run, Random forest has the best performance with 99% accuracy, 97% precision and 94% recall. From the many simulations carried out, the authors suggest optimizing the in-flight wi-fi service. After that also simplicity about online booking. Unlike before, this time with around 130,000 data that becomes 70,000 after deleting the NaN (Not a Number) value, it doesn't include inflight entertainment as an attribute that needs to be improved.

Different from the previous ones, but still about airline customer satisfaction. Hanif et al. [3] uses a dataset of 152 respondents who have used one of the Indonesian airlines, namely Lion Air. The data is taken and grouped by occupation so that it becomes 100 data and 5 classes of work. The author uses the SPSS tool to get the conclusions. By looking for multiple regression, validity, reliability, T test, F value test and the coefficient of determination and correlation, it is found that there is a positive and significant influence between service quality, passenger satisfaction and passenger behavioral intentions. The disadvantage of this research is that the data used is too little so that it can get different results if there are more datasets.

In the journal written by Wijayanto et al. [4], the Naive Bayes algorithm is also used for the passenger satisfaction dataset taken from Kaggle. The dataset used is most likely the same as that which will be used from this journal. With 129,880 data, the author uses the help of the KNime application for classification with Naive Bayes. The distribution of training data and data testing consists of 4 experiments. The first is training data: testing data is 90:10, the second is 85:15, the third is 80:20 and the last is 75:25. The results obtained that 90% of training data and 10% of testing data have an accuracy of 81.466%.

Religia and Amali [5] also uses Naive Bayes to classify airline passenger satisfaction. The dataset used is also from Kaggle but is different, as many as 25,976 data. In their research, they used Naive Bayes, Naive Bayes optimized particle swarm Optimization (PSO) and finally Naive Bayes optimized Genetic Algorithm (GA). To measure the performance used accuracy, precision and recall. The results obtained are that Naive Bayes optimized by PSO has the best results, namely the accuracy value is 86.13%, the precision value is 87.9% and the recall value is 87.29%.

Similar to this journal, Nugraha et al. [6] compare Naive Bayes with Learning Vector Quantization (LVQ) to classify. But here it is used to classify uterine diseases. In using Naive Bayes, the author uses 2 methods, Naive Bayes by using Laplacian Smoothing and without using it. The data used are 125 data from the medical records of patients at RSUD Dr. Moewardi Solo. The data here is divided into 4 experiments/simulations with the first experiment being training: the data is 20:80, the second is 40:60, the third is 60:40 and the last is 80:20. The results of 4 trials with training 20%, 40%, 60%, 80% got Naive Bayes without Laplacian Smoothing had 32%, 67.8%, 79%, 88.8% accuracy. These results are less good than if Naive Bayes using Laplician smoothing has an accuracy of 88%, 92.4%, 92.8%, 92.4%. The accuracy is said to be stable even though the training data is changed. Compared to LVQ the accuracy is 82.4%, 88.8%, 89.4%, 95.2%. However, the highest accuracy is obtained from LVQ with 80% training.

In another journal, for LVQ signature pattern recognition compared by Prabowo et al. [7] and combined by Ginting et al. [8]. Prabowo et al. compared with the Kohonen Neural Network (KNN), while Ginting et al. combined with Self Organizing Kohonen (SOK).

In the journal Prabowo et al. did 3 tests. Each test with a different number of classes, resolutions and patterns. In the first test with 25 patterns and a resolution of 30x20 Kohonen had 96% success for 1 second while LVQ was 100% for 2 seconds. Second with 40 patterns and 30x20 resolution with 95% Kohonen less than 1 second while LVQ 92.5% less than 1 second. The last test was 9 patterns with 100x100 resolution with 77.78% Kohonen for 2 seconds and LVQ 88.89% for 7 seconds. LVQ does have better accuracy than Kohonen, but it takes longer. While in the journal Ginting et al. can speed up the computational process. The combination of LVQ with SOK increases the processing speed of computing during training or during signature pattern recognition.

Unlike previous comparisons or combinations, Meliawati et al. [9] implement LVQ to predict majors at SMA PGRI 1 Banjarbaru. The data used is obtained from the value of report cards in 2010, 2011 and 2013. The data is used as training data, while the value of report cards in 2014 is used as testing data. It is not known how much of the exact amount of data was used. Researchers get 79.31% accuracy for iterations 60 and 90.

Samsir [10] also implements LVQ. LVQ is used to classify Throat Nose and Ear (ENT) disease at Rantauprapat Hospital Labuhanbatu. The input variable consists of 10 disease symptoms. The dataset used is small, which is only 57 data. Of the 57 data divided into 4 training. With the comparison of training data: Testing data is 60:40, second 70:30, third 80:20 and 90:10. In the results of testing accuracy, it is not found that the more testing data, the accuracy will improve. Maybe it's because there are too few datasets, so you might get different results if you get more datasets.

From the journal Gorzalczany et al. [1] and Hayadi et al. [2], the dataset used is almost the same. But both use different algorithms in classifying them. While Hanif et al. [3] using very different datasets and different algorithms, but it's still about passenger satisfaction. However, Wijayanto et al. [4] using the same dataset and algorithm, namely Naive Bayes only, but not compared to LVQ. Likewise, Religia and Amali [5] use only Naive Bayes to classify airline passenger satisfaction, but the datasets used are different. In the journal Nugraha et al. [6] The algorithms both compare LVQ and Naive Bayes, but they use it to classify obstetrical diseases. Prabowo et al. [7] also compared LVQ but with KNN for the case of signature pattern recognition. While Ginting et al. [8] combines LVQ with SOK for signature pattern recognition cases as well. For Meliawati et al. [9] and Samsir [10], they only implement LVQ with different datasets without comparing them or combining them.

# RESEARCH METHODOLOGY

## Data Collection

In collecting datasets, I use websites that provide various kinds of datasets. For this research I used data from https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction. Data with the file name airline\_passengeer\_satisfaction.csv has a file size of 14.34MB. I downloaded this data on September 20, 2021. To download it you are required to Sign In first (Register if you don't have an account). The downloaded file will be a zip file, so it must be extracted to get the csv file. The total data obtained were 129,880 with 23 measuring columns and 1 response column.

## Algorithm

In choosing the algorithm, I consulted my supervisor. During the consultation, my lecturer informed and suggested the Learning Vector Quantization (LVQ) algorithm. This algorithm has not been used very often. Therefore, I use this LVQ algorithm. After using LVQ I looked for another algorithm to use as a comparison. Then I chose Naive Bayes because this algorithm is an algorithm that is often used, easy and has good accuracy. I use these two algorithms to classify supervised learning data about airline passenger satisfaction that has been obtained previously. In addition to knowing which algorithm is better in accuracy.

## Coding and Design

In this step, the MySql tools will be used. MySql is used because the existing dataset is in the form of 2-dimensional data (columns and rows) the same as the MySql database table. In addition, the installation of Mysql is very easy. By downloading xampp through the website https://www.apachefriends.org/download.html. Xampp already provides several versions for Windows, Linux and OS X operating systems. Here I use Linux. After MySql is installed, the data will be preprocessed. Continuous data such as age and distance will be changed first to make it easier to classify.

## Analysis

In analyzing, I will do 5 tests as follows :

Analysis

|  |  |  |
| --- | --- | --- |
|  | Training Data | Testing Data |
| I | 10% | 90% |
| II | 25% | 75% |
| III | 50% | 50% |
| IV | 75% | 25% |
| V | 90% | 10% |

In this analysis, it is divided into 5 stages to determine whether the amount of training has an effect. Influence on Naive Bayes accuracy and on LVQ accuracy.

## Make a Report

In making the report, I wrote chapters 1-4 first. After chapter 4 finished, I started the coding stage for program development. Then the results that have been carried out during the coding stage will be recorded in the chapter 5 report. And finally, conclusions will be drawn from the results of the coding stage which will be written in chapter 6.

# ANALYSIS AND DESIGN

## Analysis

This part of the analysis is the steps taken to solve the existing problems. The first step is to get the data. The data is obtained by downloading the airline\_passengeer\_satisfaction.csv file at the link https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction. The data has the format as in table 4.1. The data, amounting to 129,880 records, will be preprocessed first. Data from several attributes will be changed or normalized data so that it can be processed by the program to be made. Changed attributes as in table 4.2.

After preprocessing the data, the next step is to create a program to find accuracy using the Naïve Bayes algorithm and Learning Vector Quantization (LVQ). In finding the value of accuracy, the formula as in Function 1. is used. To find the accuracy, it is necessary to implement the Naïve Bayes and LVQ algorithms first. The algorithm implementation process is as shown in Figure 4.1 in the design section.

Naive Bayes is a fairly easy algorithm. This is because the formula used is quite easy. The general form of the Bayes theorem looks like in Function 2. In this Naive Bayes implementation, the steps I use are:

1. Divide the dataset into training datasets and testing datasets. The distribution of the dataset is as shown in table 4.3.

2. Calculate P(Ci) for each class of customer satisfaction as in function 3 of the training dataset.

3. Calculate P(X|Ci) with the predicted attributes of each satisfaction class. Do it for all attributes (gender, age, etc).

4. Calculate the result of multiplying P(X|Ci) all attributes and P(Ci).

5. The biggest results are prediction results.

6. To find the accuracy results later then if :

a. Dataset "satisfied" and prediction "satisfied", TP added 1

b. Dataset "neutral or dissatisfied" and prediction result "satisfied", FP added 1

c. Dataset "neutral or dissatisfied" and prediction result "satisfied", FN added 1

d. Dataset "neutral or dissatisfied" and prediction result "neutral or dissatisfied", TN added 1

7. Repeat steps 3-6 for all testing datasets.

After step 7 is complete then we calculate the accuracy. To calculate accuracy like Function 1 with input in step 6. Then the first test is done. Repeat steps 1-7 for the second to fifth test with the number of training data and datasets as specified. If it has been tested 5 times, then Naive Bayes has been completed in this study.

After doing Naive Bayes, next is the implementation of Learning Vector Quantization (LVQ). LVQ is an algorithm that performs learning where the class is predetermined, just like Naive Bayes. This algorithm is not as easy as Naive Bayes. The steps performed by the LVQ algorithm are:

1. Initialization
   1. The initial weight (W) is randomly selected 1 input data from each class
   2. Maximum iterations (MaxEpoh)
   3. Parameters Learning rate (α)
   4. Minimum error (Eps)
2. Input
   1. Input x(m,n)

* m = data to m
* n =

1. Initial Condition

Learning Vector Quantization merupakan algoritma yang melakukan pembelajaran dimana kelasnya sudah ditentukan sebelumnya. Langkah-langkah yang dilakukan algoritma LVQ adalah:

1. Inisialisasi
   1. Bobot awal (W) dipilih secara acak 1 inputan data dari masing-masing kelas
   2. Maksimum iterasi (MaxEpoh)
   3. Parameter Learning rate (α)
   4. Error minimum (Eps)
2. Input
   1. Input x(m,n)

* m = data ke m
* n = variable input ke-n
  1. Target = T(1,n)

1. Kondisi awal :
   1. Epoh = 0
   2. Eps = 1
2. If Epoh < MaxEpoh or α > eps :
   1. Epoh = epoh+1
   2. Untuk i=1 sampai n

* Tentukan J hingga || Xi -Wj || minimum (Cj)
* Update Wj jika
  + T = Cj maka : Wj(baru) = Wj(lama) + α[Xi-Wj(lama)]
  + T ≠ Cj maka : Wj(baru) = Wj(lama) – α[Xi-Wj(lama)]
  1. Kurangi nilai α = α – α \* 0,1

T = Target

J = Jumlah selisih antara data ke-m dan bobot

C = Kelas selisih bobot terkecil

W = Bobot

α = Rasio Pembelajaran

x = Data

## Penambahan Gambar, Judul Gambar, dan Penggunaan Gambar

Architecture LVQ

The Learning Vector Quantization architecture is divided into 3 parts, namely input, process and output. In the input section, it is necessary to initialize several variables as in chapter 4.1 in LVQ steps no 1-3. After that, step 4 is carried out in the process section. If step 4 has stopped then the output will be obtained.

## Tabel, Posisi, dan Isi tabel

Data Table

|  |  |  |
| --- | --- | --- |
| No. | Attribute Name | Attribute Description |
| 1. | Id | Is the id of the data |
| 2. | Gender | 2 class: “Female” (51%) and “Male” (49%) |
| 3. | Customer Type | 2 class: “Loyal Customer” (82%) and “Disloyal Customer” (18%) |
| 4. | Age | Numbers from 7 to 85 |
| 5. | Type of Travel | 2 class: “Business travel” (69%) and “Personal Travel” (31%) |
| 6. | Customer Class | 3 class: “Business” (48%), “Eco” (45%) and “Eco Plus” (7%) |
| 7. | Flight Distance | Numbers from 31 to 4983 |
| 8. | Inflight Wi-fi Service | ‘0’ (3.02%) ‘1’ (17.19%) ‘2’ (24.88%)  ‘3’ (24.78%) ‘4’ (19.08%) ‘5’ (11.05%) |
| 9. | Departure Arrival Time Convenient | ‘0’ (5.14%) ‘1’ (14.94%) ‘2’ (16.58%)  ‘3’ (17.23%) ‘4’ (24.54%) ‘5’ (21.56%) |
| 10. | Ease of Online Booking | ‘0’ (4.37%) ‘1’ (16.85%) ‘2’ (23.14%)  ‘3’ (23.40%) ‘4’ (18.82%) ‘5’ (13.41%) |
| 11. | Gate Location | ‘0’ (0.00%) ‘1’ (16.93%) ‘2’ (18.71%)  ‘3’ (27.50%) ‘4’ (23.46%) ‘5’ (13.40%) |
| 12. | Food and Drink | ‘0’ (0.10%) ‘1’ (12.36%) ‘2’ (21.08%)  ‘3’ (21.40%) ‘4’ (23.53%) ‘5’ (21.52%) |
| 13. | Online Boarding | ‘0’ (2.37%) ‘1’ (10.21%) ‘2’ (16.89%)  ‘3’ (20.88%) ‘4’ (29.62%) ‘5’ (20.03%) |
| 14. | Seat Comfort | ‘0’ (0.00%) ‘1’ (11.63%) ‘2’ (14.27%)  ‘3’ (17.96%) ‘4’ (30.61%) ‘5’ (25.53%) |
| 15. | Inflight Entertainment | ‘0’ (0.01%) ‘1’ (12.07%) ‘2’ (16.91%)  ‘3’ (18.39%) ‘4’ (28.33%) ‘5’ (24.29%) |
| 16. | Onboard Service | ‘0’ (0.00%) ‘1’ (11.39%) ‘2’ (14.13%)  ‘3’ (21.98%) ‘4’ (29.80%) ‘5’ (22.71%) |
| 17. | Leg Room Service | ‘0’ (0.46%) ‘1’ (9.93%) ‘2’ (18.89%)  ‘3’ (19.29%) ‘4’ (27.63%) ‘5’ (23.79%) |
| 18. | Baggage Handling | ‘0’ (0.00%) ‘1’ (6.95%) ‘2’ (11.06%)  ‘3’ (19.90%) ‘4’ (36.00%) ‘5’ (26.08%) |
| 19. | Check in Service | ‘0’ (0.00%) ‘1’ (12.40%) ‘2’ (12.40%)  ‘3’ (27.30%) ‘4’ (27.97%) ‘5’ (19.93%) |
| 20. | Inflight Service | ‘0’ (0.00%) ‘1’ (6.82%) ‘2’ (11.08%)  ‘3’ (19.49%) ‘4’ (36.43%) ‘5’ (26.23%) |
| 21. | Cleanliness | ‘0’ (0.01%) ‘1’ (12.88%) ‘2’ (15.49%)  ‘3’ (23.59%) ‘4’ (26.15%) ‘5’ (21.88%) |
| 22. | Departure Delay in Minutes | Numbers from 0 to 1592 |
| 23. | Arrival Delay in Minutes | Numbers from 0 to 1584 |
| 24. | Satisfaction | 2 class: “Neutral or dissatisfied” (57%) and “Satisfied” (43%) |

In table 4.1 there are column names of the data to be used. Of the 24 existing columns, 22 columns will be used as input variables, namely columns 2-23. Column 1 is not used because it is an id that has no value as an input variable. While the 24th column is the class column of the data used as the class determining variable.

Modified attribute data table

|  |  |  |
| --- | --- | --- |
| No. | Attribute Name | Attribute Description |
| 1. | Age | “<28” (26.01%), “<52” (50.13%), “>=52” (23.86%) |
| 2. | Flight Distance | “<=414” (25.16%), “<=1744” (50.05%), “>1744” (24.79%) |
| 3. | Departure Delay in Minutes | “<=12” (75.22%) and “>12” (24.78%) |
| 4. | Arrival Delay in Minutes | “<=13” (75,54%) and “>13” (24.46%) |

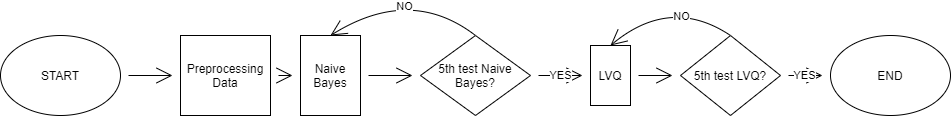
From the existing data, it is necessary to change the data so that it can be processed by the program. The data is changed because it has too many different values. Because it is grouped into data that is easier to process. The converted number is obtained from the quantile value which can be seen via the Kaggle link in the Column section.

Distribution of training and testing data

|  |  |  |
| --- | --- | --- |
|  | Training Data | Testing Data |
| I | 10% | 90% |
| II | 25% | 75% |
| III | 50% | 50% |
| IV | 75% | 25% |
| V | 90% | 10% |

In this study, the distribution of the number of training and testing data was carried out in each test. There will be 5 tests. This division is carried out to see whether the amount of training and testing of different data will affect the accuracy results.

## Design



Program Flowchart

This flowchart is the workflow of making the program. The first is to make changes to the data (normalization) in the data preprocessing step. Then when the implementation of the Naïve Bayes algorithm is carried out 5 times with different comparisons of training and testing data. Then continued with the LVQ algorithm with 5 trials as well, namely the percentage of training data of 10%, 25%, 50%, 75% and 90%.

## Function

True Positive (TP) is the number of data from class satisfied that is correct and classified as class satisfied.

True Negative (TN) is the number of data from class neutral or dissatisfied that is correct and classified as class neutral or dissatisfied.

False Positive (FP) is the number of data from class neutral or dissatisfied that is incorrect and is classified as class satisfied.

False Negative (FN) is the number of data from class satisfied that is incorrect and classified as class neutral or dissatisfied.

x is data whose class is unknown

y is the hypothesis of data x from a specific class.

P(y|x) is the probability of the hypothesis y based on the condition x.

P(y) is the hypothesis probability y.

P(x|y) is the probability of x based on these conditions.

P(x) is the probability of x.

Ci is the airline's passenger satisfaction class (“satisfied” and “neutral or dissatisfied”).

P(Ci) is the probability of each class.

Ci is the airline's passenger satisfaction class (“satisfied” and “neutral or dissatisfied”).

X is class of attributes (in attribute gender, x is “male” and “female”).

P(X | Ci) is the probability of X and Ci.

Total data is the amount of data from training data based on the conditions of testing data.

# IMPLEMENTATION AND RESULTS

## Implementation

Bab implementasi adalah bab tentang narasi pemanfaatan dari data structure dan algoritma dalam bentuk aplikasi terapan.

Bab implementasi menyertakan source code, namun tidak semua source code program disertakan dalam bab ini. Ambil lah penggalan nya saja yang penting dan menjadi inti dari program Anda. Gunakan style “Code” . Jika menyertakan gambar (capture), silahkan ditambahkan caption di gambar tersebut sebagaimana penjelasan pada Chapter 5.

1. create Function sfHelloWorld (vNama varchar(30))
2. returns varchar(100)
3. begin
4. declare vHello varchar(255)
5. select concat(‘Hello ‘, vNama) into vHello;
6. return vHello;
7. End

Setiap source code diberikan nomor urut baris. Jelaskan baris perintah dan untuk apa perintah tersebut. Baris 1-2 kode program berisi perintah untuk membuat function dengan nama sfHelloWorld. Baris 3 dan 7 adalah blok baris untuk function khusus di dalamnya. Inti dari kode program ada pada baris 4 untuk deklarasi variabel vHello, baris 5 untuk menggabungkan karakter “Hello” dengan variabel vNama dan mengembalikan hasilnya pada baris ke 7.

## Results

Sub bab results berisi hasil dari uji coba algoritma dan struktur data yang diterapkan dalam bentuk aplikasi. Hasil disajikan dalam bentuk tabel, narasi atau gambar yang dapat memberikan penjelasan solusi masalah dengan bantuan program sehingga dapat ditarik kesimpulan dari penelitian anda.

# CONCLUSION

Bab ini membahas tentang kesimpulan akhir. Harus menjawab semua pertanyaan yang Anda ajukan sebagai permasalahan yang bab 1 bagian scope. Tidak sekedar menyimpulkan tapi sertakan argumentasi kuat terkait pengambilan kesimpulan tersebut.

Di bagian akhir, sertakan saran untuk penelitian lanjutan. Tidak perlu bertele-tele tapi fokuskan pada saran penelitian Anda saja, apa yang belum dilakukan disertakan di sini.

REFERENCES

[1] M. B. Gorzałczany, F. Rudziński, and J. Piekoszewski, “Business Intelligence in Airline Passenger Satisfaction Study—A Fuzzy-Genetic Approach with Optimized Interpretability-Accuracy Trade-Off,” *Applied Sciences*, vol. 11, no. 11, p. 5098, May 2021, doi: 10.3390/app11115098.

[2] B. H. Hayadi, J.-M. Kim, H. T. Sukmana, and K. Hulliyah, “Predicting Airline Passenger Satisfaction with Classification Algorithms,” *IJIIS : Int. J. Inform. Inform. Systems.*, vol. 4, no. 1, pp. 82–94, Mar. 2021, doi: 10.47738/ijiis.v4i1.80.

[3] M. R. Hanif, R. Chrizara, E. Saribanon, and I. Ozali, “THE EFFECT OF SERVICE QUALITY AND PASSENGER SATISFACTION ON PASSENGER BEHAVIORAL INTENTIONS ON LION AIR,” p. 9.

[4] A. Wijayanto, J. F. A. Bernardo, and S. Pamungkas, “Analisis Klasifikasi Kepuasan Penumpang Maskapai Penerbangan Menggunakan Algoritma Naïve Bayes,” *J-SAKTI*, vol. 3, no. 2, pp. 97–103, May 2021, doi: 10.33084/jsakti.v3i2.2041.

[5] Yoga Religia and A. Amali, “Perbandingan Optimasi Feature Selection pada Naïve Bayes untuk Klasifikasi Kepuasan Airline Passenger,” *RESTI*, vol. 5, no. 3, pp. 527–533, Jun. 2021, doi: 10.29207/resti.v5i3.3086.

[6] P. Aditya Nugroho, R. Saptono, and M. Eko Sulistyo, “Perbandingan Metode Probabilistik Naive Bayesian Classifier dan Jaringan Syaraf Tiruan Learning Vector Quantization dalam Kasus Klasifikasi Penyakit Kandungan,” *ITSmart*, vol. 2, no. 2, p. 21, Mar. 2016, doi: 10.20961/its.v2i2.628.

[7] A. Prabowo and E. A. Sarwoko, “Perbandingan Antara Metode Kohonen Neural Network dengan Metode Learning Vector Quantization Pada Pengenalan Pola Tandatangan,” vol. 14, p. 7.

[8] E. B. Ginting, Prof. Dr. M. Zarlis, and Dr. Z. Situmorang, “KOMBINASI ALGORITMA JARINGAN SYARAF TIRUAN LEARNING VECTOR QUANTIZATION (LVQ) DAN SELF ORGANIZING KOHONEN PADA KECEPATAN PENGENALAN POLA TANDA TANGAN.” 2014.

[9] R. Meliawati, O. Soesanto, and D. Kartini, “PENERAPAN METODE LEARNING VECTOR QUANTIZATION (LVQ) PADA PREDIKSI JURUSAN DI SMA PGRI 1 BANJARBARU,” vol. 04, p. 10, 2016.

[10] Samsir, “Klasifikasi Penyakit Tenggorokan Hidung Telinga (THT) Menggunakan Jaringan Syaraf Tiruan Dengan Metode Learning Vektor Quantization (THT) Di RSUD Rantauprapat Labuhanbatu.”

APPENDIX

Jika Anda punya lampiran dari project, silahkan dilampirkan di bagian ini. Yang wajib Anda lampirkan adalah kode program (coding) lengkap dan diberikan keterangan terlebih dahulu pada bagian atas dari coding tersebut, koding ditulis dengan format font yang berbeda. Contoh:

**CODING PEHITUNGAN**

1. SELECT @a := 5;
2. SELECT @b := 5;
3. SELECT hasil:= @a \* @b;

**PROCEDURE HITUNG PERKALIAN**

1. CREATE PROCEDURE spMaksimal ()
2. BEGIN
3. DECLARE a INT;
4. DECLARE b INT;
5. DECLARE hasil INT;
6. SELECT a \* b INTO hasil;
7. SELECT hasil
8. END

Selain coding, yang dapat dijadikan lampiran adalah: hasil hitungan yang panjang dan tidak mungkin dimuat dalam laporan utama, gambar atau ilustrasi diagram yang cukup panjang namun hanya sebagai penjelasan dari diagram utama yang ada di dalam laporan, foto-foto penunjang, dan dokumen lain yang sifatnya menunjang namun dianggap penting.