**Vietnam General Confederation of Labor**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**MIDTERM REPORT**

**INTRODUCTION TO**

**MACHINE LEARNING**

*Instructor*: **Ph.D LE ANH CUONG**

*Student*: **Ho Huu An – 521H0489**

**Tran Nhut Anh – 521H0491**

**Do Minh Quan – 521HH0290**

*Class* **: 21H50301**

*Year* **: 25**

**HO CHI MINH CITY, 2023**

Vietnam General Confederation of Labor

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**MIDTERM REPORT**

**INTRODUCTION TO**

**MACHINE LEARNING**

*Instructor*: **Ph.D LE ANH CUONG**

*Student*: **Ho Huu An – 521H0489**

**Tran Nhut Anh – 521H0491**

**Do Minh Quan – 521HH0290**

*Class* **: 21H50301**

*Year* **: 25**

**HO CHI MINH CITY, 2023**

ACKNOWLEDGEMENT

We would like to express my sincere appreciation to Le Anh Cuong for his exceptional teaching on machine learning. His expertise and guidance have provided us with a comprehensive understanding of various models, metrics, and practical applications. His dedication to our growth, patient approach, and real-world examples have been truly inspiring. I am grateful for his unwavering support and belief in our abilities. Thank you, Le Anh Cuong, for being an outstanding educator and empowering us in the field of machine learning..

*Ho Chi Minh city, 14th October, 2023*

*Author*

*(Sign and write full name)*

**THIS PROJECT WAS COMPLETED AT**

**TON DUC THANG UNIVERSIY**

We fully declare that this is our own project and is guided by Mr. Le Anh Cuong; The research contents and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the author himself from different sources, clearly stated in the reference section.

Besides that, the project also uses a number of comments, assessments as well as data from other authors, other agencies and organizations, with citations and source annotations.

**Should any frauds were found, I will take full responsibility for the content of my report.** Ton Duc Thang University is not related to copyright and copyright violations caused by me during the implementation process (if any).

*Ho Chi Minh city, 14th January, 2022*

*Author*

*(Sign and write full name)*

CONFIRMATION AND ASSESSMENT SECTION

**Instructor confirmation section**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh January, 2022*

*(Sign and write full name)*

**Evaluation section for grading instructor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh January 2022*

*(Sign and write full name)*

SUMMARY

This is a report on Introduction to Machine Learning by Faculty of Information Technology of Ton Duc Thang University.

INDEX

[ACKNOWLEDGEMENT i](#_Toc148875251)

[CONFIRMATION AND ASSESSMENT SECTION iii](#_Toc148875252)

[SUMMARY iv](#_Toc148875253)

[INDEX 1](#_Toc148875254)

[CHAPTER 1 – INTRODUCTION 3](#_Toc148875255)

[Introduce the problem 3](#_Toc148875256)

[Dataset Description 3](#_Toc148875257)

[CHAPTER 2 – DATA PREPROCESSING 5](#_Toc148875258)

[1. Deleting unuseful columns 5](#_Toc148875259)

[2. Finding and cleaning null values 5](#_Toc148875260)

[3. Delete duplicate data 5](#_Toc148875261)

[4. Handling Outliers 5](#_Toc148875262)

[5. Encoding Target Column 5](#_Toc148875263)

[6. Unnecessary Features 5](#_Toc148875264)

[7. Data encoding 5](#_Toc148875265)

[8. Standardization 5](#_Toc148875266)

[9. Data Splitting 6](#_Toc148875267)

[10. Imbalanced Data 6](#_Toc148875268)

[CHAPTER 3 – MODEL BUILDING 7](#_Toc148875269)

CHAPTER 1 – INTRODUCTION

Introduce the problem

In this document, we will address a classification problem related to airline passengers. The goal is to predict whether a passenger is likely to cancel their flight or not based on various features available in the dataset. This problem is of significant importance to airline companies as it can help them understand customer behavior and optimize their operations accordingly.

Dataset Description

The dataset consists of information about airline passengers and their travel experiences. Each row represents a passenger and includes the following features:

id: A unique identifier for each passenger.

Gender: The gender of the passenger.

Customer Type: Indicates whether the passenger is a loyal customer or a disloyal customer.

Age: The age of the passenger.

Type of Travel: Specifies whether the passenger is traveling for personal or business reasons.

Class: The class of the flight booked by the passenger.

Flight Distance: The distance of the flight in miles.

Inflight wifi service: The rating of inflight wifi service on a scale of 1 to 5.

Departure/Arrival time convenient: The rating of the convenience of departure/arrival times on a scale of 1 to 5.

Ease of Online booking: The rating of the ease of online booking on a scale of 1 to 5.

Gate location: The rating of gate location on a scale of 1 to 5.

Food and drink: The rating of food and drink quality on a scale of 1 to 5.

Online boarding: The rating of the online boarding process on a scale of 1 to 5.

Seat comfort: The rating of seat comfort on a scale of 1 to 5.

Inflight entertainment: The rating of inflight entertainment on a scale of 1 to 5.

On-board service: The rating of on-board service on a scale of 1 to 5.

Leg room service: The rating of leg room service on a scale of 1 to 5.

Baggage handling: The rating of baggage handling on a scale of 1 to 5.

Check-in service: The rating of check-in service on a scale of 1 to 5.

Inflight service: The rating of inflight service on a scale of 1 to 5.

Cleanliness: The rating of cleanliness on a scale of 1 to 5.

Departure Delay in Minutes: The number of minutes of departure delay.

Arrival Delay in Minutes: The number of minutes of arrival delay.

Satisfaction: Indicates whether the passenger was satisfied or (neutral or dissatisfied) with their travel experience.

The data includes various features that capture different aspects of the passenger's journey, such as ratings for services and amenities, flight details, and customer characteristics. These attributes can be used to build a machine learning model to predict passenger satisfaction or make other relevant predictions related to airline passengers.

CHAPTER 2 – DATA PREPROCESSING

Data preprocessing refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models

1. Deleting unuseful columns

Observations: There are two columns Unnamed: 0 and id,These two columns are of no use in the classification

Decisions: It is better to delete these two columns

1. Finding and cleaning null values

Train Test

Observations

* Arrival Delay in Minutes column contain null values.
* For training data 310 null values ~ (0.298%)
* For testing data 83 null values ~ (0.32%)

Decisions

* Drop Nan values in Arrival Delay in Minutes column

1. Delete duplicate data

A group of white text

Description automatically generated

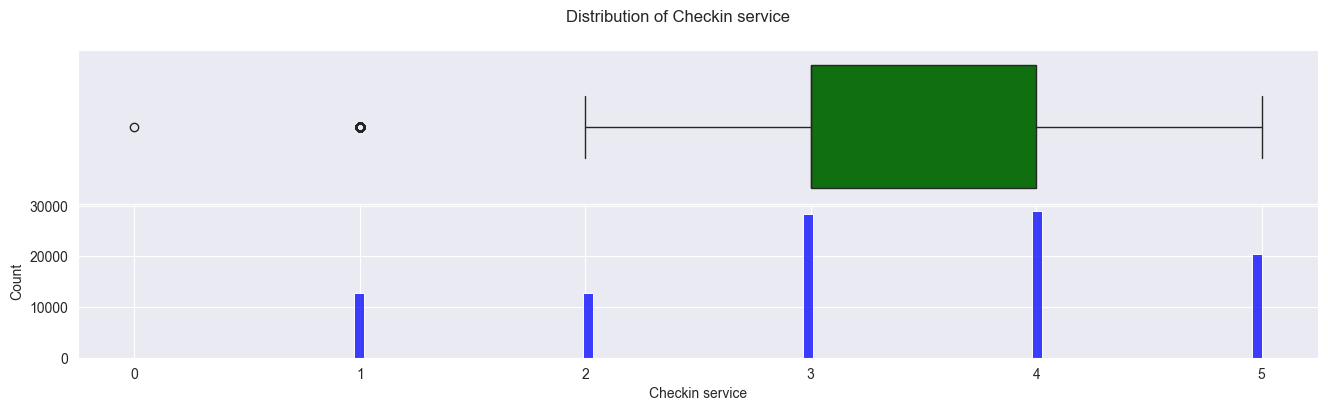
Observations: There is no duplicate data in both training and testing data

1. Handling Outliers

Train

A graph with a green square and blue bars

Description automatically generated



A screenshot of a computer

Description automatically generated

A white paper with black text

Description automatically generated

Test

A graph with a green square and blue bars

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A white paper with black text

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

Observations

There are outliers in Train data

* 'Arrival Delay in Minutes' ~ 13.42%
* 'Checkin service' ~ 12.40%
* 'Departure Delay in Minutes'~ 13.92%
* 'Flight Distance' ~ 2.20%

There are outliers in Test data. In

* 'Arrival Delay in Minutes' ~ 13.62%
* 'Checkin service' ~ 12.38%
* 'Departure Delay in Minutes' ~ 13.73%
* 'Flight Distance' ~ 2.24%

Decisions

* First drop outliers in 'Flight Distance' column in train and test data
* Second let's impute outliers in 'Checkin service' columns
* Third let's handle outliers in 'Departure Delay in Minutes' column with KNN Algorithm and don't care about outliers in 'Arrival Delay in Minutes' it will be droped

1. Encoding Target Column

Using LabelEncoder for target column

Space efficiency: LabelEncoder represents categorical labels using integers, which typically requires less memory compared to storing the original categorical labels as strings or objects. This can be advantageous when dealing with large datasets

Compatibility with certain algorithms: Some machine learning algorithms, such as decision trees or random forests, can handle categorical variables directly when they are encoded as integers. LabelEncoder provides a convenient way to convert categorical labels into integers, making them compatible with these algorithms

1. A blue and white chart with numbers

   Description automatically generatedUnnecessary Features

Observation: The correlation coefficient between the two columns is 0.97, which is very high. This means that the two columns are highly related to each other.

Decision: Dropping the "Arrival Delay in Minutes" column would make the dataset smaller and easier to manage

1. Data encoding

OneHot Encoding is a process used in data analysis and machine learning to convert categorical variables into numerical format. It creates binary vectors where each category is encoded as a binary feature. In this encoding scheme, each category is represented by a vector of zeros with only one element being 1, known as the "hot" element. This encoding helps in representing categorical data in a way that can be easily understood and processed by machine learning algorithms

Decision: encoding ['Gender','Customer Type', 'Type of Travel', 'Class']

1. Normalization

MinMax scaling, also known as min-max normalization, rescales the range of features to a given range, typically [0, 1]. It does this by subtracting the minimum value and then dividing by the range (maximum value - minimum value). This means that all features will have a minimum of 0 and a maximum of 1, regardless of whether they were originally centered around 0.

The formula for MinMax scaling: scaled\_feature = (feature - min\_feature) / (max\_feature - min\_feature)

Decision: normalize columns ['Age', 'Flight Distance', 'Inflight wifi service',

'Ease of Online booking', 'Food and drink', 'Online boarding',

'Seat comfort', 'Inflight entertainment', 'On-board service',

'Leg room service', 'Baggage handling', 'Checkin service',

'Inflight service', 'Cleanliness', 'Departure Delay in Minutes']

1. Data Splitting

Splitting training data: Create X\_train by dropping the "satisfaction" column from the train dataset, and assign "satisfaction" column to y\_train.

Splitting test data: Create X\_test by dropping the "satisfaction" column from the test dataset, and assign "satisfaction" column to y\_test.

Encoding target variable: Convert categorical values in y\_test ("satisfied" and "neutral or dissatisfied") to numerical values (1 and 0, respectively).

K-fold cross-validation: Create k\_fold using KFold with 10 folds, shuffling the data with a random state of 42.

1. Imbalanced Data

The RandomOverSampler randomly duplicates samples from the minority class until it is balanced with the majority class. This helps to increase the representation of the minority class in the training data.

CHAPTER 3 – MODEL BUILDING

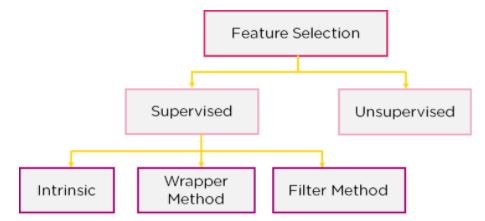
K-Nearest Neighbors (KNN): A simple classification algorithm that assigns a class label based on the majority vote of its nearest neighbors.

Logistic Regression: A linear classification algorithm that models the relationship between input features and the probability of belonging to a class.

Decision Tree: A hierarchical model that recursively partitions the feature space to make decisions based on feature values.

Naive Bayes: A probabilistic classification algorithm based on Bayes' theorem, assuming independence between features.

XGBoost: An ensemble learning method that combines weak predictive models (decision trees) to create a strong model through gradient boosting.

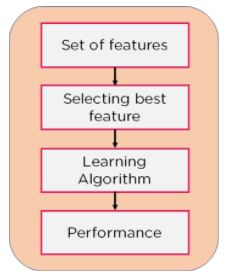
CHAPTER 4 – FEATURE SELECTION

Feature selection models are of two types:

* Supervised Models: Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model
* Unsupervised Models: Unsupervised feature selection refers to the method which does not need the output label class for feature selection. Using them for unlabelled data.

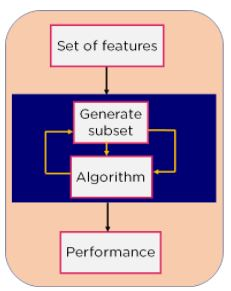
1. ****Filter Method****

features are dropped based on their relation to the output, or how they are correlating to the output. Using correlation to check if the features are positively or negatively correlated to the output labels and drop features accordingly



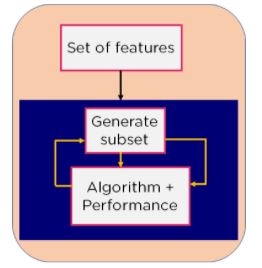
1. Wrapper Method

Split our data into subsets and train a model using this. Based on the output of the model, we add and subtract features and train the model again



1. Intrinsic Method

Combines the qualities of both the Filter and Wrapper method to create the best subset.



Method takes care of the machine training iterative process while maintaining the computation cost to be minimum. Eg: Lasso and Ridge Regression.

1. Remove features with low variance

VarianceThreshold is a basic method for performing feature selection. This method removes all features that have a variance lower than a certain threshold.

Naturally, VarianceThreshold will remove features that have a variance of 0, meaning features that have only a single unique value.

Let's consider the following example: suppose we have a dataset with boolean features, and we want to remove columns that have more than 80% of the records with the same value. Boolean variables can be seen as random discrete Bernoulli variables. And the formula to calculate variance for this type of variable is: Var[X] = p(1 - p).

**Features are retained:** ['Flight Distance', 'Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness', 'Customer Type\_Loyal Customer', 'Customer Type\_disloyal Customer', 'Type of Travel\_Business travel', 'Type of Travel\_Personal Travel', 'Class\_Business', 'Class\_Eco', 'Class\_Eco Plus']

1. Eliminate features with low variance

Univariate feature selection is a technique for filtering the best features based on statistical tests with univariates. Feature selection can be used as a preprocessing step for a machine learning model. Scikit-learn organizes feature selection functions as objects that contain transform functions.

SelectKBest only retains the k features with the highest score.

SelectPercentile retains the percentage of features with the highest score.

Use common univariate statistical tests on each feature, such as false positive rate (false positive rate), SelectFpr, false discovery rate (false discovery rate,

FP/(FP+TP)) SelectFdr, or family wise error (system error) SelectFWE.

GenericUnivariateSelect allows you to use univariate feature selection methods in a customizable way. Using this method helps us choose the best single variable selection strategy, using estimation models to choose the best hyper-parameters.

The above objs are required to receive an evaluation algorithm, passed in as param of the objs. These evaluation algorithms return univariate scores and p-values ​​of features.

For regression problems, we use f\_regression, mutual\_info\_regression

For the classification problem, we use chi2, f\_classif, mutual\_info\_classif.

**Features are retained:** ['Age', 'Flight Distance', 'Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness', 'Departure Delay in Minutes', 'Gender\_Female', 'Gender\_Male', 'Customer Type\_Loyal Customer', 'Customer Type\_disloyal Customer', 'Type of Travel\_Business travel', 'Type of Travel\_Personal Travel', 'Class\_Business', 'Class\_Eco', 'Class\_Eco Plus']

1. Tree-based feature selection

Tree-based feature selection is a technique that uses decision tree algorithms to identify the most important features in a dataset. It involves training a tree-based model, such as Random Forest or Gradient Boosting, and evaluating the contribution of each feature to the model's performance. The most important features are then selected for further analysis or modeling.

**Features are retained:** ['Inflight wifi service', 'Ease of Online booking', 'Online boarding', 'Inflight entertainment', 'Baggage handling', 'Type of Travel\_Business travel', 'Type of Travel\_Personal Travel', 'Class\_Business', 'Class\_Eco']

1. A screenshot of a computer

   Description automatically generatedCompare between models

K-Nearest Neighbors

A screenshot of a computer

Description automatically generated  
Logistic regression

A screenshot of a computer

Description automatically generatedDecision Tree

Naive Bayes

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generatedXgboost

CHAPTER 5 – CONCLUSION

* Feature selection methods can increase accuracy for some models by reducing noise and improving generalization, but they can decrease accuracy for other models that rely on a larger set of features or have built-in mechanisms to handle irrelevant features. The impact of feature selection depends on the characteristics of the model and the specific data set.
* Compatibility between features and models: The suitability of a feature selection method depends on factors like data type, feature space size, and model requirements. It's important to choose a method that aligns well with these factors.
* There is no perfect choice of method features for every problem
* Pay attention to overfitting: When applying feature selection, we need to ensure that the model is not overfitting. Some feature selection methods can cause overfitting if not used appropriately. Therefore, evaluating and monitoring overfitting during the feature selection process is very important.