A MACHINE LEARNING APPROACH FOR PREDICTING THE SMOKING BEHAVIOR AMONG ADOLESCENTS

A PROJECT REPORT

Submitted by

GANPATH T	**(311616104023)**
AKSHATH D PATNI	**(311616104005)**
DHUMAVAT SHUBHAM	**(311616104021)**

in partial fulfilment for the award of degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

MISRIMAL NAVAJEE MUNOTH JAIN ENGINEERING COLLEGE

THORAIPAKKAM, CHENNAI-600097

ANNA UNIVERSITY: CHENNAI 600025

APRIL 2020

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report "**A MACHINE LEARNING APPROACH FOR PREDICTING THE SMOKING BEHAVIOR AMONG ADOLESCENTS"** is the bonafide work of "**GANPATH T, AKSHATH D PATNI & DHUMAVAT SHUBHAM "** who carried out the project work under my supervision.

SIGNATURE
Dr. P.Indira Priya B.E., ME., Ph.D
HEAD OF THE DEPARTMENT

Department of Computer Science
and Engineering
Misrimal Navajee Munoth Jain
Engineering College,
Thoraipakkam,
Chennai-600097

SIGNATURE
Ms. X.ANITHA SARAFIN, M.E., (Ph.D)
SUPERVISOR
ASSOCIATE PROFESSOR
Department of Computer Science
and Engineering
Misrimal Navajee Munoth Jain
Engineering College,
Thoraipakkam,
Chennai-600097

Submitted for the Project Viva-Voce Examination held on 22-09-2020

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We express our sincere gratitude to our honourable Secretary (Administration) **Dr. Harish L Metha** and Secretary (Academic) **Shri L. Jaswant S Munoth** for providing the infrastructure facilities to do this project during our course period.

We thank our **Principal**, **Dr. C. C. CHRISTOPHER, M.E., Ph.D,** for his valuable suggestions and guidance for the development and completion of this project.

We express profound sense of gratitude and thanks to our **HEAD OF THE DEPARTMENT, Dr. P.Indira Priya, B.E., M.E., Ph.D** for her valuable suggestions and guidance for the development and completion of this project.

We thank our project coordinator, **Mr. S. Saravanan, M.E., (Ph.D),** for his valuable suggestions for the development and completion of this project.

We thank our supervisor **Mrs. X. Anitha Sarafin, M.E., (Ph.D), Associate Professor** for her constant help and valuable suggestions. We thank her for her whole hearted guidance right from the inception to the completion of our project.

Finally, we thank all the **Teaching and Non-Teaching Staff members of our Department who helped us to complete our project.**

Above all we thank the Almighty, our parents and Siblings for their constant support and encouragement for completing this project.

ABSTRACT

The use of electronic cigarette (e-cigarette) is increasing among adolescents. This is problematic since consuming nicotine at an early age can cause harmful effects in developing teenager's brain and health. Additionally, the use of e-cigarette has a possibility of leading to the use of cigarettes, which is more severe. There were many researches about e-cigarette and cigarette that mostly focused on finding and analyzing causes of smoking using conventional statistics. However, there is a lack of research on developing prediction models, which is more applicable to anti-smoking campaign, about e-cigarette and cigarette.

Our aim is to develop a prediction models that can be used to predict an individual e-cigarette user's intention to smoke cigarettes, so that one can be early informed about the risk of going down the path of smoking cigarettes.

To construct the prediction models, four machine learning (ML) algorithms are exploited and tested for their accuracy in predicting the intention to smoke cigarettes among never smokers using data from the 2018 National Youth Tobacco Survey (NYTS).

In our investigation, the Gradient Boosting Classifier, one of the prediction models, shows the highest accuracy out of all the other models. Also, with the best prediction model, we made a public website that enables users to input information to predict their intentions of smoking cigarettes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

S.NO	ABBREVIATIONS	EXPANSION
1	AI	ARTIFICIAL INTELLIGENCE
2	ML	MACHINE LEARNING
3	KNHANES	KOREA NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY
4	NYTS	NATIONAL YOUTH TOBACCO SURVEY
5	CSV	COMMA SEPARATED VALUE
6	GNB	GAUSSIAN NAÏVE BAYES
7	DTC	DECISION TREE CLASSIFIER
8	RFC	RANDOM FOREST CLASSIFIER
9	GBC	GRADIENT BOOSTING CLASSIFIER
10	TP	TRUE POSITIVE
11	TN	TRUE NEGATIVE
12	FP	FALSE POSITIVE
13	FN	FALSE NEGATIVE

CHAPTER 1

INTRODUCTION

1.1 BIG DATA:

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with datasets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis, searching, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. When we handle big data, we may not sample but simply observe and track what happens.

1.2 DATA ANALYSIS:

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today’s business world, data analysis plays a role in making decisions more scientific and helping and helping businesses operate more effectively.

1.3 ARTIFICIAL INTELLIGENCE (AI):

Artificial intelligence is a branch of computer science that aims to create intelligent machines. It is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions) and self-correction. Artificial intelligence (AI) makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks. Most AI examples that you hear about today – from chess-playing computers to self-health evaluation system – rely heavily on machine learning and deep learning. Using these technologies, computers can be trained to accomplish specific tasks by processing large amount of data and recognizing patterns in the data.

1.4 MACHINE LEARNING (ML):

Machine learning (ML) is the scientific study of algorithms and statistical models that computer use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning (ML) algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine Learning (ML) is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

1.5 DATA MINING:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information from a data set and transform the information from a data set and transform the information into a comprehensible structure for further use. The primary aim is to allow extraction of patterns and knowledge from large amounts of data, not the extraction of data itself.

CHAPTER 2

LITERATURE SURVEY

2.1 DATA MINING TECHNIQUE FOR THE ENHANCED SMOKING CESSATION MANAGEMENT SYSTEM

ABDULLAH ALSHARIF, H et al has stated, that data mining can be defined as the use of complex tools of data analysis to discover previously unknown relationships and patterns in large datasets. Health care databases have huge amounts of data, and with effective analysis tools, a great deal of hidden knowledge in huge amounts of data such as those obtained from smokers. It also can help to discover new healthcare knowledge for clinical and administrative decision making, as well as producing scientific hypotheses from large sets of experimental data and clinical databases. This study aim to detect smoker behavior, and behavior change therapy to determine the smokers quit plan. This system is based on a continuous acquisition of data, thereby improving its results regularly [1].

2.2 SMOKING PROBABILITY AND CIGARETTES CONSUMPTION OF CHINESE PEOPLE

CHEN KE et al has stated, has calculated smoking behavior using:

- Personal smoking probability estimation based on age, income, price where a logits functions is used calculate cross function entropy.
- Estimated cigarette demand in each state of china based on income, price, age, smoking time using regression technique [2].

2.3 A MACHINE LEARNING APPROACH FOR PREDICTING SUCCESS IN SMOKING CESSATION INTERVENTION

KISHHIGSUREN DAVAGDORJ et al has stated, to analyze the smoking cessation intervention dataset conducted from the Korea National Health and

Nutrition Examination Survey (KNHANES) 2009 to 2017. Using chi-square test to filter relevant and significant features, thus the multivariate analysis was used with logistics regression. Age, education and frequent alcohol use are important predictors in smoking cessation success, furthermore, the lowest level of subjective health status has increased the likelihood of unsuccessful smoking cessation. This study does not solves the prediction problem by dealing with imbalanced data. Although five different algorithms are use but highest accuracy of 86.03% is obtained [3].

2.4 BIG DATA ANALYSIS OF YOUTH TOBACCO SMOKING TRENDS IN THE UNITED STATES

SHILPA BALAN et al has stated, to understand the smoking trends among patients and found out the top states of US where a youth consume more tobacco and showed the comparison of youth tobacco smokers between men and women based on data visualization such as histogram but this does not provide a clear information on how each parameters are taken for calculation and any preventive actions are not provided [4].

2.5 PREDICTION OF DAILY SMOKING BEHAVIOR BASED ON DECISION TREE MACHINE LEARNING ALGORITHM

ZHANG YUPU et al has stated, with the accumulation of smoking data and the development of the algorithm, precise analysis become possible and this can benefit smoking cessation a lot. They used the decision tree machine learning algorithm to predict daily smoking time. Data they used are from the Chinese center for disease control and prevention. In order to solve the problem of too little feature information, they used a feature information extraction module. This prediction produce with a accuracy of 84.11% and its training time was faster than any of other model [5].

CHAPTER 3

SYSTEM OVERVIEW

3.1 EXISTING SYSTEM

3.1.1 OVERVIEW

An important component of the smoking cessation program is understanding the factors and predicting success for quitting which is an effective way for public health benefit.

Existing systems have highlighted the significant factors associated with smoking cessations using relationships such as sociodemographic behavioral and environmental characteristics, comorbid conditions, and quitting method using bivariate analysis.

Existing system estimated weight gain would associated with early that post-cessation weight gain would relapse over and influence of other variable such as treatment condition, baseline, nicotine consumption, gender. They use time-invariant regression to subgroups smokers.

One-sided focus on individual’s problems and not on their existing strengths. Since each person has different weakness and strengths, using their strength to their advantage is not employed in any of the available system.

Existing system have accuracy and speed of prediction very much less. Stream processing of data was not available in the system.

3.1.2 DRAWBACK

In the existing system, while computing classifiers, class imbalanced problem occurs in many real-words applications where the class distributions of data are highly imbalanced.

Existing system doesn’t provide correlation between the predictor variable to that of target variable.

Data pattern recognition in early system was not accurate and speed which tends to provide user incorrect information about smoking behavior, because when we don’t have higher volume and variety of data it is not possible to have greater accuracy of prediction.

Simple features that are already available can be embedded to the applications in a more appealing and useful manner but aren’t used due to the accessibility level for the youth smokers. This can be solved by making the applications follow universal design principles.

Personalized Evaluation of smoker intention to smoke in future can be made based on available health care data set so that user can evaluate or change its behavior based on various predictions.

3.2 PROPOSED SYSTEM

3.2.1 OVERVIEW

We use the National Youth Tobacco Survey (NYST) results from 2018 and construct multiple prediction models (Gradient Boosting Classifier and Decision Tree Classifier) that can predict whether a person will have an intention to smoke cigarettes or not. After data analysis, Gradient Boosting Classifier, one of the prediction models, had the highest accuracy of 93% out of all the models tested. We divide the data into two sets: never smokers and smokers of cigarette. The group of never-smokers were analyzed to find the best fitting model to predict the intention to smoke cigarettes for both e-cigarette smokers and non-e-cigarette smokers. In addition, we create a website involving the Gradient Boosting Classifier model in order to allow the public to input factors (e.g., sex, age, and various habits) and receive a prediction of whether or not they will have a high intention to smoke cigarettes or not. This will give the general public more awareness of their position as to whether or not they will smoke cigarettes and possibly steer away from the path of smoking cigarettes.

Consequently, two important aspects of this system includes:

- Find the best-fitting model to predict smoking intention from the NYST data
- Create a website to help students, especially e-cigarette smokers, be able to prevent e-cigarette use due to possible chance of smoking cigarette.

3.2.2 ADVANTAGES

This public web that will allow adolescents to know whether they will have the intention to smoke or not.

An adaptive user interface will allow youths to input their behavior and based on input a machine learning algorithm predict the best-fitting behavior for that user.

Features used in this system offers higher dimensionality because length and breadth of data are largely considered.

Speed and accuracy of this system is far better than previous models used.

Feature selection is done based on the mutual information classification which will largely help us to assign weightage to each parameter which is used in model.

3.3 REQUIREMENT ANALYSIS

The requirement specification is a technical specification of requirements for the software products. The purpose of the software requirement specification is to provide a detailed overview of the software project, its parameter and goal. It describes the project target audience and its user interface, hardware and software requirements.

3.4 SOFTWARE AND HARDWARE REQUIREMENTS

3.4.1 SOFTWARE REQUIREMENT

The software requirements give a detailed description of the system and all its features.

- MS Excel
- Python 3
- Jupyter Notebook
- Flask
- SKlearn

3.4.2 HARDWARE REQUIREMENT

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete engineer as the starting point for the system design.

Ram	:	2GB Ram and more
Processor	:	Any Intel Processor
Hard Disk	:	6GB and more
Speed	:	1GHZ and more

3.5 TECHNOLOGIES USED

3.5.1 MS Excel

Microsoft Excel is a spreadsheet developed by Microsoft for windows. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letters named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It can be also be used to create a comma separated value (CSV) file which can be used in python for further processing the data.

MS Excel can also be used for data processing, and can be used for removing all those attribute which has missing attribute in it.

3.5.2 PYTHON 3

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Typical uses includes math and computation, algorithm development, modelling, simulation, prototyping, data analysis, exploration, visualization and application development including graphical user interface building.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python interpreters are available for many operating systems. A global community of programmers develops and maintains CPython, an open source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and CPython development. object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

3.5.3 JUPYTER NOTEBOOK

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, and containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

Jupyter notebook also provide us with cell which can be used as code editor for python based editor.

A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell.

3.5.4 FLASK

Flask was created by Armin Ronacher of Pocoo, an international group of Python enthusiasts formed in 2004. Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, and upload handling, various open authentication technologies and several common framework related tools. Extensions are updated far more frequently than the core Flask program.

CHAPTER 4

SYSTEM DESIGN

4.1 OVERVIEW OF ARCHITECTURE

The diagram shows the overall architecture of the project which consists of the web pages and server block. Each block shows their structural and functional components of the project.

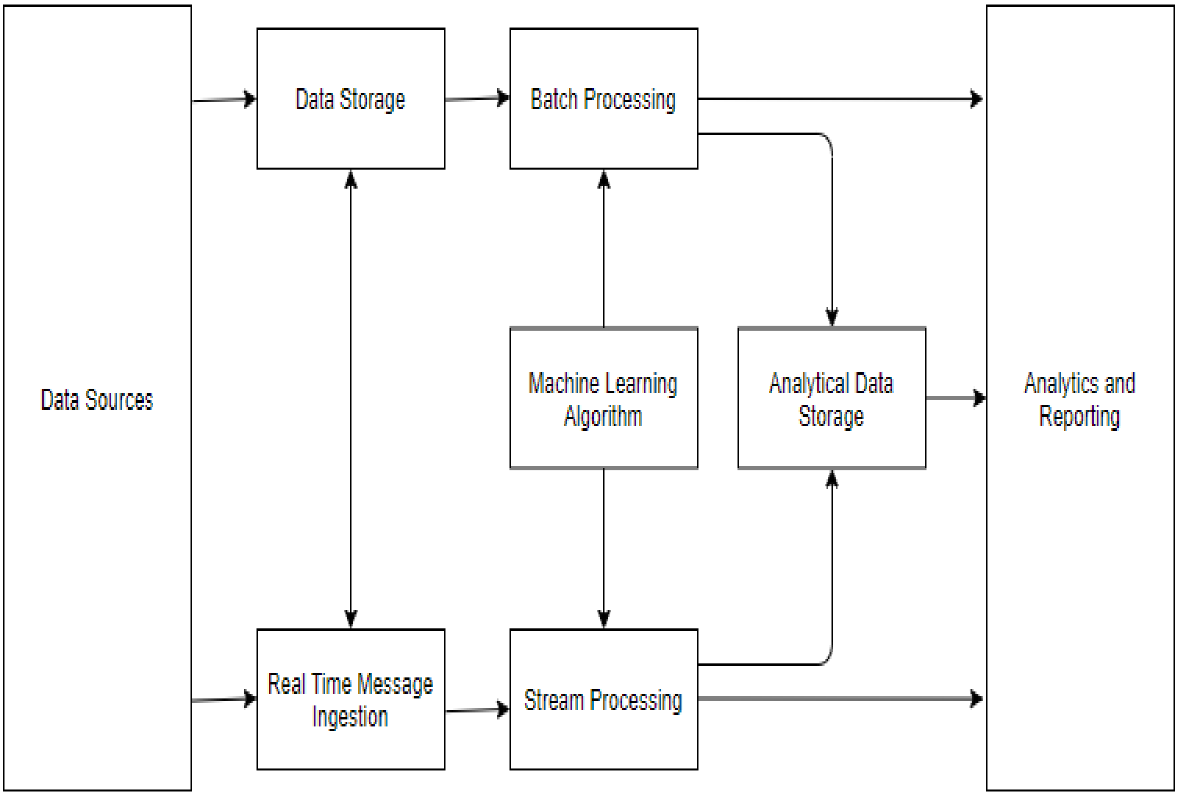


Figure 4.1 OVERVIEW OF ARCHITECTURE

4.2 USE CASE DIAGRAM

This diagram shows the user such as actor, editors and the role of developer in this project. This behaviour diagram models the functionality of the system using use cases.

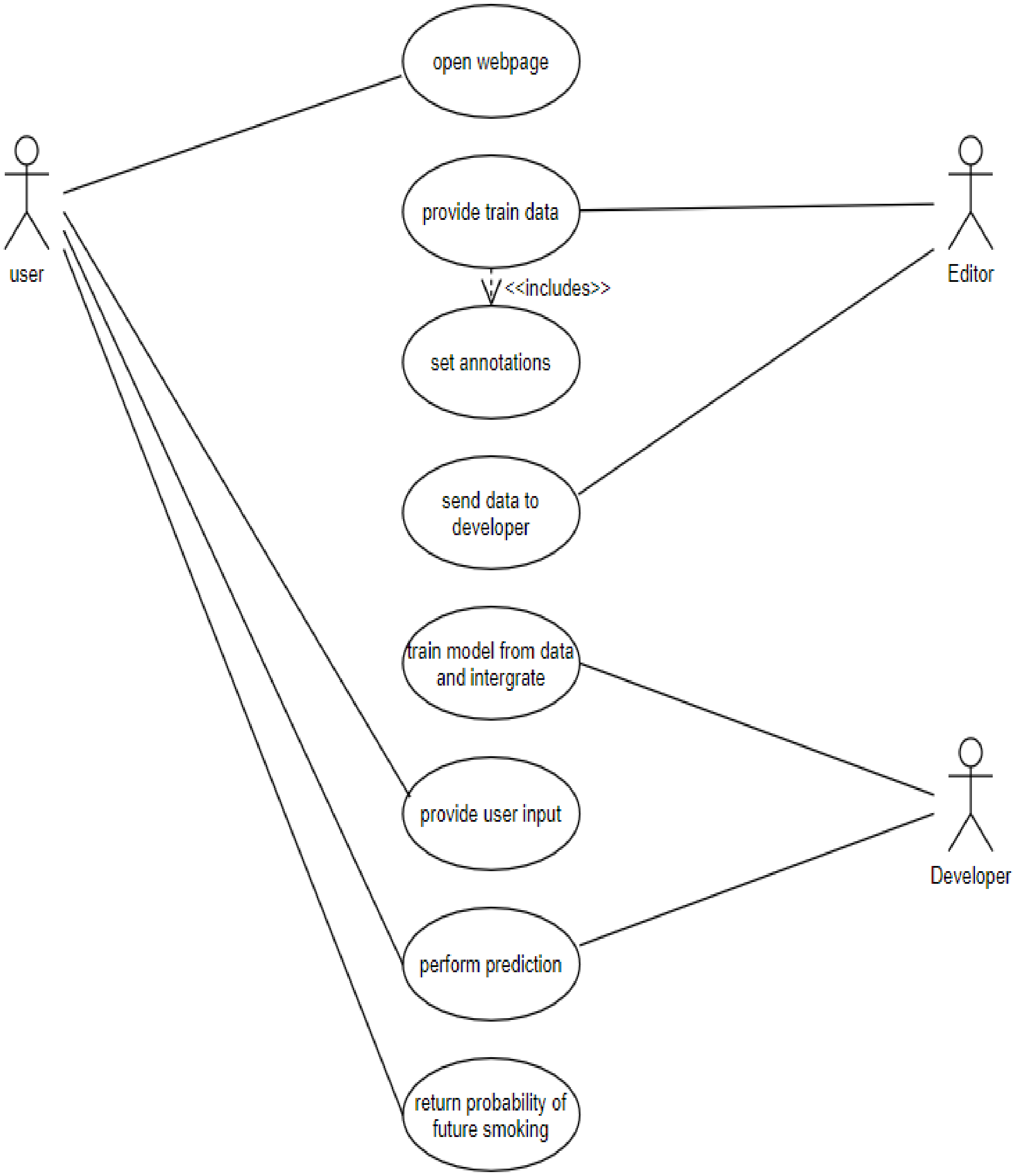


Figure 4.2 USE CASE DIAGRAM

4.3 CLASS DIAGRAM

The structure of the application id described in the class diagram by showing the system’s classes, their attributes, operations (or methods), and the relationships among objects. Each class has their attributes and the operations defined with the relationship between the classes.

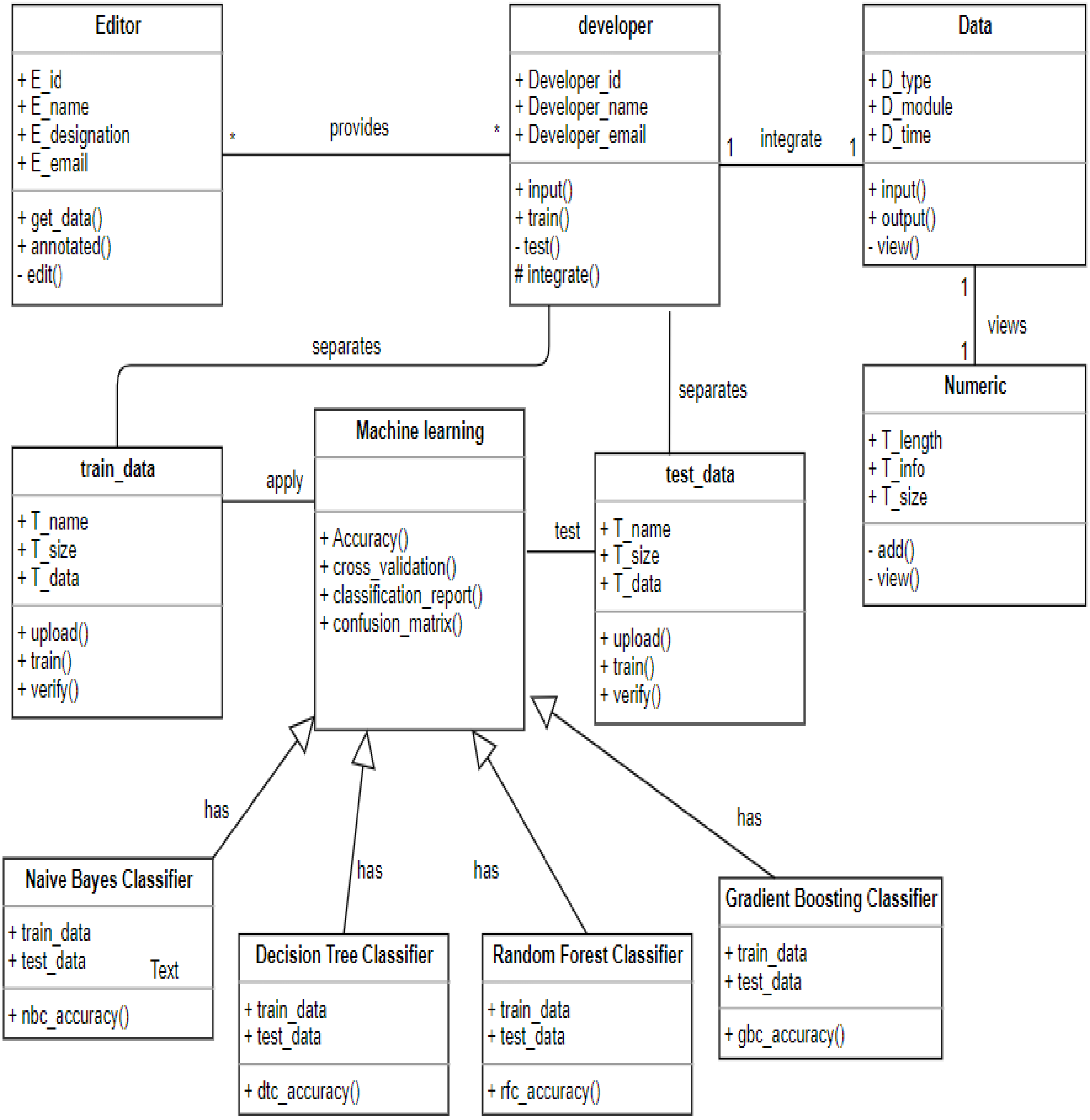


Figure 4.3 CLASS DIAGRAM

4.4 MODULE DIAGRAM

4.4.1 MODULE 1/2/3 – BUSINESS UNDERSTANDING

This module diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interaction take place. The objects in this module are Developer, server, User and Application.

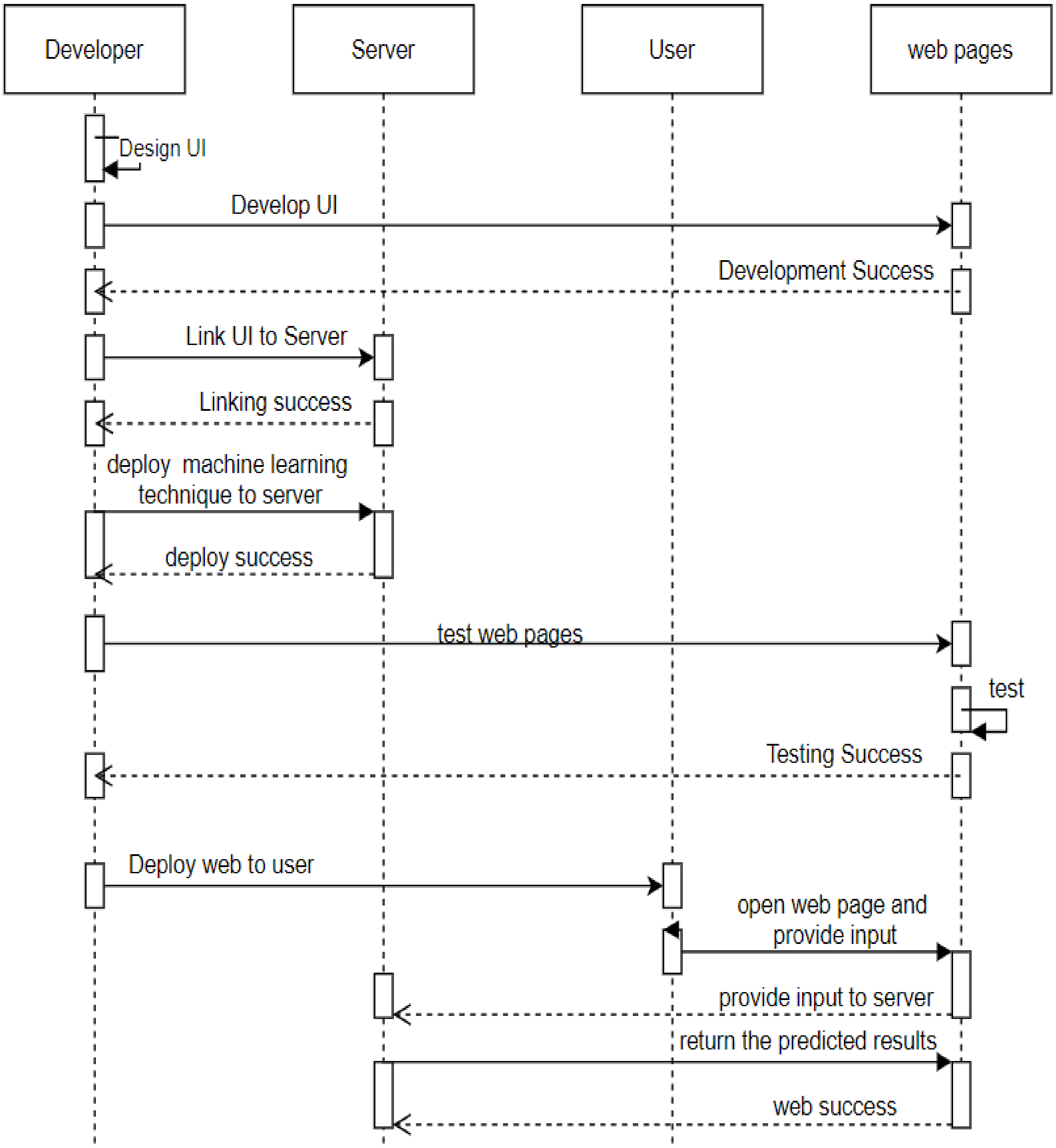


Figure 4.4.1 MODULE 1/2/3 – Sequence Diagram for Business Understanding

4.4.2 MODULE 4/5 – MODELLING

This module diagram represents the flow from one activity to another activity. The activity can be described as an operation of the system. Some activities are based on conditions satisfied by the actor/object.

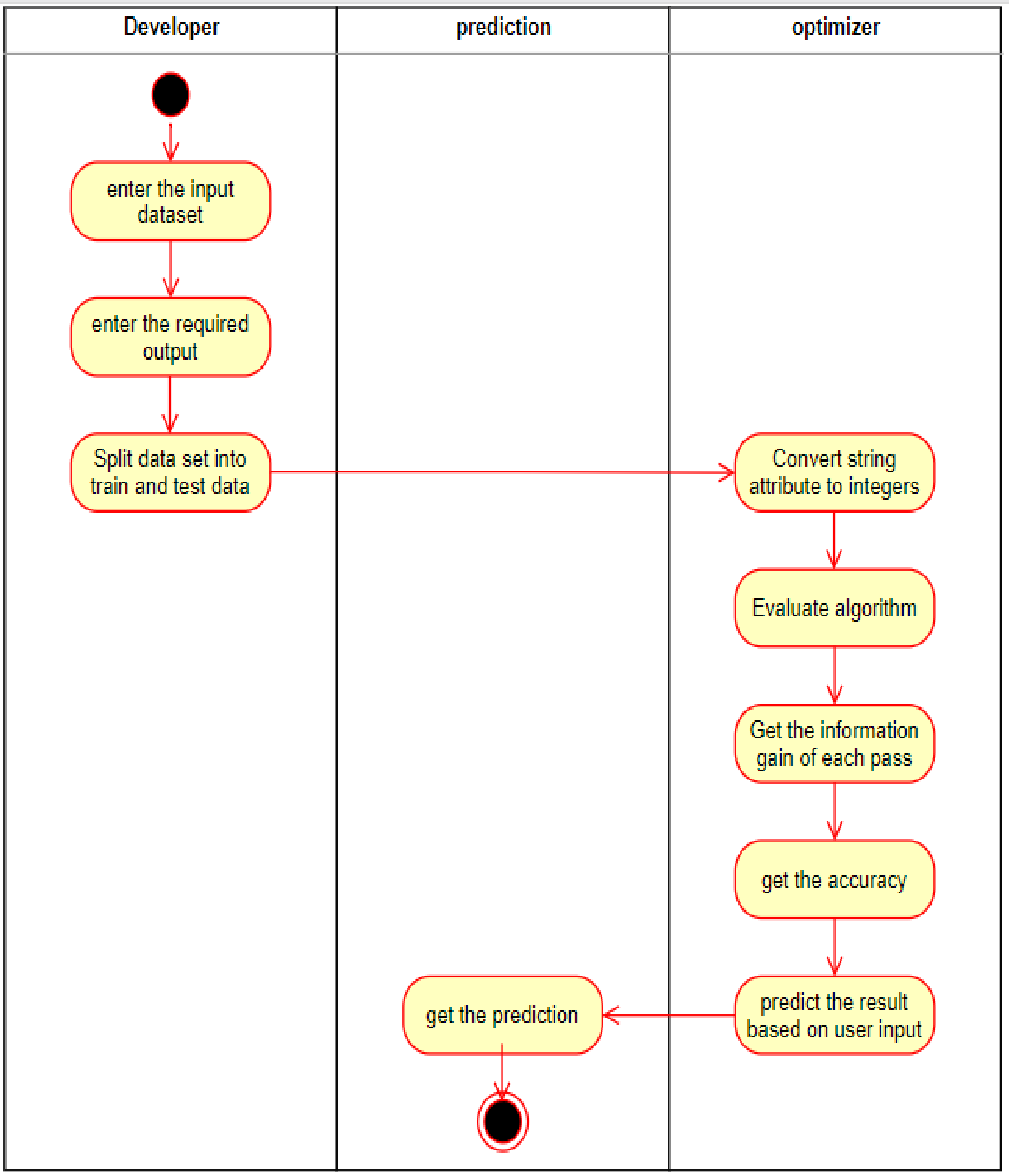

Developer
prediction
optimizer
enter the input dataset
enter the required output
Split data set into train and test data
Convert string attribute to integers
Evaluate algorithm
Get the information gain of each pass
get the accuracy
predict the result based on user input
get the prediction

Figure 4.4.3 MODULE 4/5 Activity diagram for Modelling

4.4.3 MODULE 6 – DEPLOYMENT

This module diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. The objects in this module are User, Web Server, ML model Reasoner.

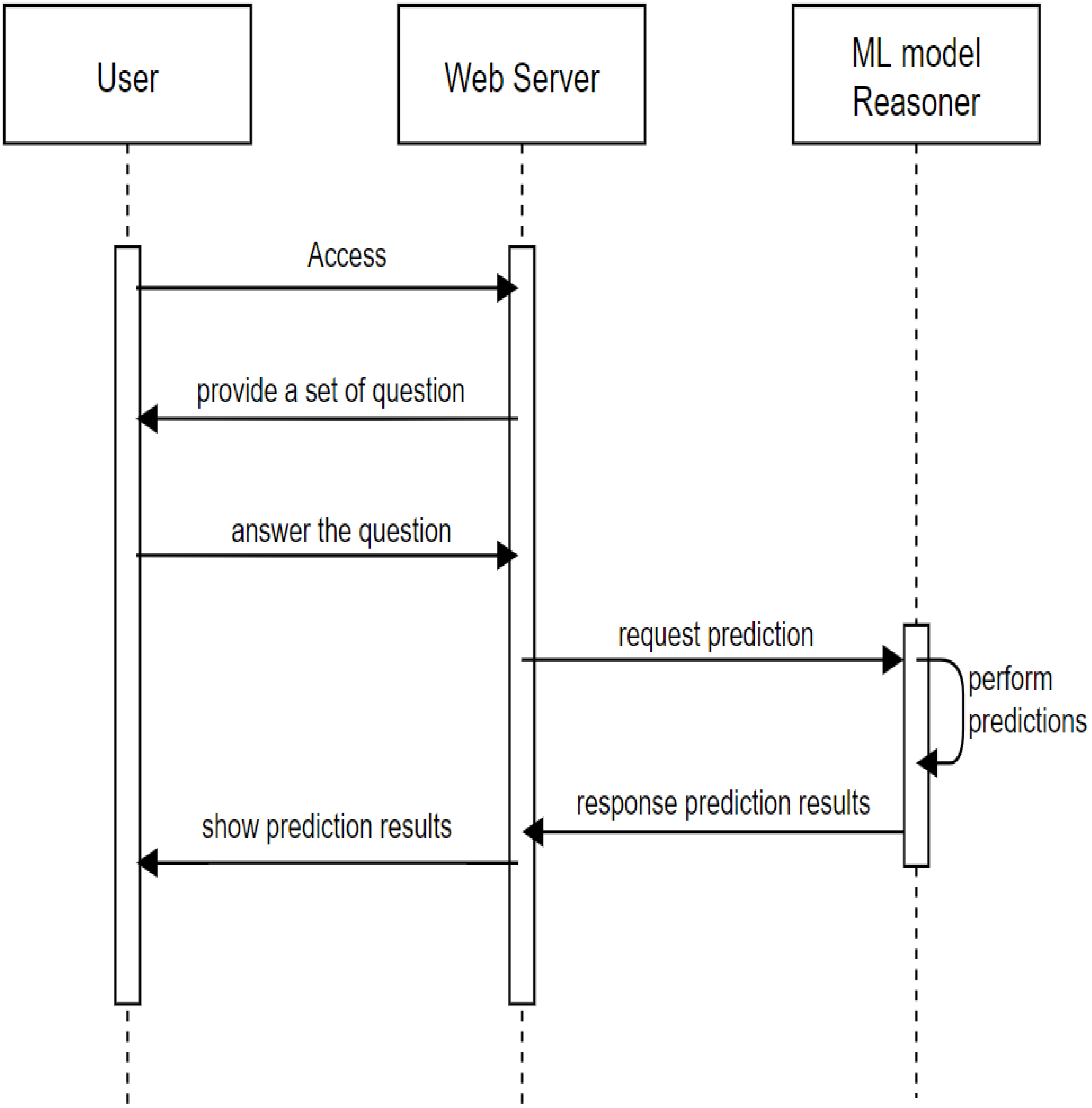


Figure 4.4.3 MODULE 6 – Sequence Diagram for Deployment

CHAPTER 5

IMPLEMENTATION

5.1 MODULES

- Business Understanding
- Data Understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

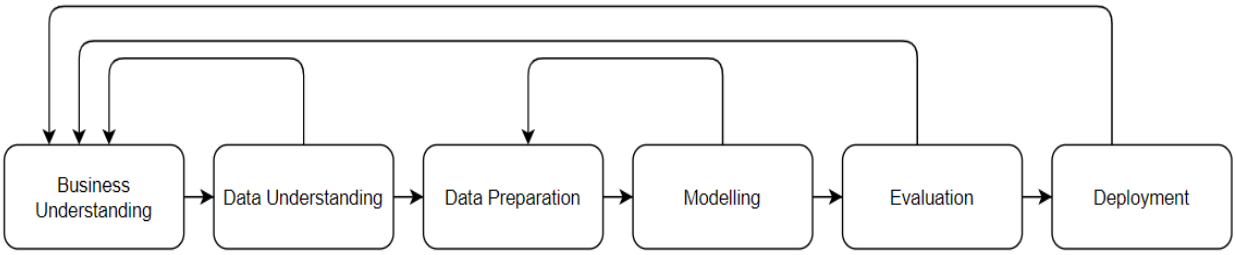


Figure 5.1 phases of implementation

5.1.1 BUSINESS UNDERSTANDING

In this project, our goal has two folds of objectives.

- The first goal is to determine the most accurate machine learning model between Gaussian Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting.
- The second is to construct a public web that will allow adolescents to know whether they will have the intention to smoke or not.

We utilize the data from NYTS for 2018 in order to construct ML models, and the models are analysed to choose the most accurate one in predicting the intention for a non-smoker to smoke cigarettes.

5.1.2 DATA UNDERSTANDING

Data Understanding includes finding an initial data set, familiarizing with the contents in the data set, and making observations from the data set in order to create a hypothesis. The initial data set must be reliable and accurate, which means that it is not outdated and has correlation within itself.

Question Number	Question	Answers
Q1	How old are you?	9, 10, ..., 19 years old
Q2	What is your sex?	Male/Female
Q3	What grade are you in?	6, 7, …, 12, ungraded or other grade
…	…	…
Q88	Because of a physical…making decision?	No/Yes

5.1.3 DATA PREPARATION

- Our original data was taken from the National Youth Tobacco Surveys 2018. This data set was compiled of questions represented as questions Q and answers represented by numbers (e.g., 1, 2, 3, and 4) and words (e.g., Yes and No).
- We filtered every question to prepare for machine learning. For example, any null in the answer meant that the question was not answered. We went through the process of replacing all the nulls with 0’s, which represents the unanswered choices.
- The next process involved classifying related data. When we first found the data, it was a set of 20189 rows x 195 columns.

- Our goal is to construct a prediction model construct a public web. In order to achieve this goal, we needed to divide the data into two groups: never smoked cigarette users and ever smoked cigarette users.
- Then, all the rows containing individuals who have ever smoked in their lives were deleted, because we only need to analyse the youths who never smoked cigarettes. The purpose of the first split of the cigarette and non-cigarette users is to form a predictive model.
- We extracted our target question and the questions pertaining to it. There are 88 questions within the survey we used, but not every question was related to our prediction.
- After examining all the survey questions, we chose specific questions to be used, since not all the questions were applicable to our goal and redundancies and indirect correlations were present. Out of 88 questions, 41 questions were used.

5.1.4 MODELLING

We used five ML algorithms to generate each ML model using a training data and evaluated the accuracy for each one of the models. The models include

- Gaussian NB Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

Out of the whole data set, we assigned the training data set, which is 80 percent of the original data, and the test data set, which is 20 percent of the original data. The training data set is used to learn the prediction model while the test data set is used to test the learned model by evaluating the accuracy for each model. After examining the predicted accuracy, we can evaluate the model that showed the highest accuracy.

NAÏVE BAYES CLASSIFIER

- Naive Bayes (NB) Classifier is a classification method based on probability theory, by which one can classify class labels using inputs. Basically, NB represents a joint distribution for input random variables(X) and a class random variables(Y) on the assumption of conditionally independence of X given Y = c.

$$P(\mathrm{X}, \mathrm{Y} = \mathrm{c}) = P(\mathrm{Y} = \mathrm{C}) \prod_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{P}(\mathrm{X}_{\mathrm{i}} | \mathrm{Y} = \mathrm{c}) \qquad (1)$$

- Where K is the number of input random variables and c denotes a class label in Y.
- It uses continuous inputs under the Gaussian (or Normal) distribution assumption.

$$P(X, Y = c) = P(Y = C) \prod_{i=1}^{k} N(X_i | \mu_i \text{ ,c}, \sigma_i, \text{c}) \qquad (2)$$

- Where μ_i, c and σ_i, c denote the mean and standard deviation of the input i for the class label c, respectively.

DECISION TREE CLASSIFIER

- Decision tree Classifier consists of a tree structure containing a set of hierarchical nodes. The root node and the internal nodes represent features or variables, while the leaf nodes denote values of a target variable. The main challenge of machine learning for decision tree is to construct these nodes and their hierarchy in a decision tree, so that it can effectively classify classes using input data of predictor variables.

$$I(p, n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n} \qquad (3)$$

- Where p and n denote the numbers of positive and negative cases, respectively. The expected information E(.) for a parent node A of the target variable can be derived as the weighted average.

$$E(A) = \sum_{i=1}^{v} \frac{p_{i+} n_i}{p+n} I(p_i|n_i) \quad (4)$$

- Where v denotes the number of the parent node values and I(p_i, n_i) denotes the expected information for the i-th value of the parent node. The information gain(A) for the node A can be obtained as the following equation.

$$gain(A) = I(p, n) - E(A) \quad (5)$$

RANDOM FOREST CLASSIFIER

- A set of ML models can often have a better performance than the use of a simple ML model. Such integration of ML models is called an ensemble learning.
- Random Forest Classifier uses the ensemble learning by forming a set of decision trees and resulting in an output which are voted from each decision tree. Random Forest draw random samples from training data and learn a decision tree model from the sample data, so that it can have a set of decision trees (i.e., forest).
- After machine learning, in the prediction (or application) stage, the class voted by the majority of learned decision trees is chosen as the final result. The following shows an equation for such majority voting.

$$\hat{y} = \text{mode}\{a_1(x), a_2(x), \dots, a_n(x)\} \quad (6)$$

- Where $a_i(x)$ is a single decision tree and the function *mode*(.) yields the output as the class label that is the most frequent class among the set of classification results.

GRADIENT BOOSTING CLASSIFIER

- Gradient Boosting Classifier uses an ensemble model consisting of a set of simple models.

- By adding such simple models, the result ensemble model can be sequentially improved and finally fitted to data.
- In other words, after applying a simple model, samples which are classified by it are reused to another simple model. And then this process is repeated until convergence (or achieving better predictive performance). Gradient Boosting Classifier is a generalized method of boosting by using gradient of a loss function.

5.1.5 EVALUATION

Evaluation is reviewing performance (e.g., accuracy) of models and deciding a best model. This is the process where the models are improved based a goal performance. Success criteria can be based on speed of algorithm, memory usage, or prediction accuracy. The sum of correct classification divided by the total number of classifications, can be used.

Since this project uses classification, some criteria for classification are introduced in the following.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{The\ number\ of\ correct\ classification}}{\boldsymbol{The\ total\ number\ of\ classification}} = \frac{\sum_{i=1}^{N} x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}} \qquad (7)$$

Where N denotes the number of class x_{ij} and denotes the total number of the case in which values of i-th prediction and j-th observation are identical. In this study, we used the four – performance metrics Accuracy (Equation1), Precision (Equation2), Recall (Equation 3), and F1-score (Equation 4). These metrics can be easily calculated by using the following four indicators.

- True Positive (TP): The amount of the observed positive values which were correctly predicted.
- False Positive (FP): The amount of the observed positive values which were wrongly predicted.

- False Negative (FN): The amount of the observed negative values which were wrongly predicted.
- True Negative (TN): The amount of the observed negative values which were correctly predicted.

These four indicators can be used to define the equations of Precision and Recall as shown.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{8}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{9}$$

Precision is commonly used to measure the influence of False Positive, while Recall is used to measure the influence of False Negative. F1-score is defined as the weighted average of Precision and Recall.

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2 * (Precision *Recall)}}{\boldsymbol{Precision+Recall}} \tag{10}$$

Precision, Recall, and F1-score have a score of one when the prediction is perfect. For the total prediction failure, they yield a score of zero.

5.1.6 DEPLOYMENT

Deployment is the process of organizing the information gained, such as the model, so that it is understandable for the model user. This step is carried out by the user rather than the analyst, so it is essential for the user to understand how to use the models. The result of this step is a final report.

CHAPTER 6

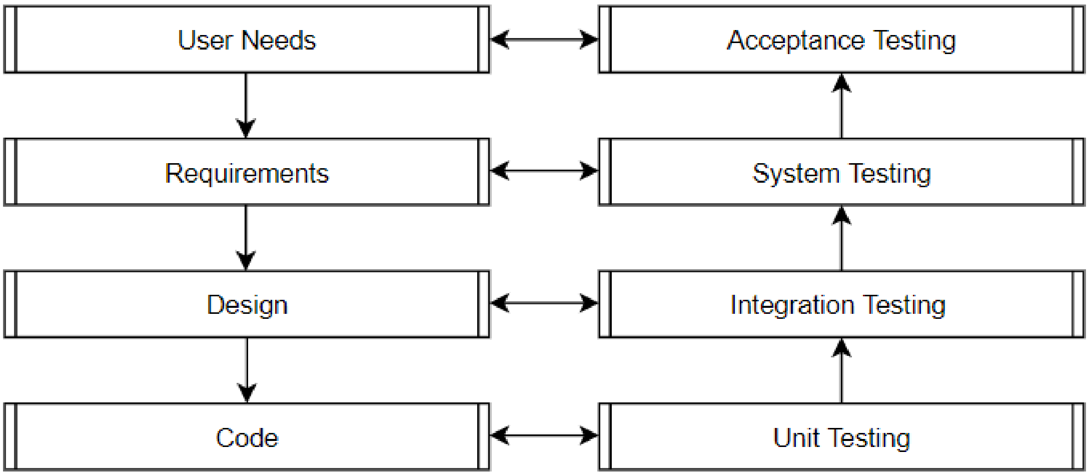
SYSTEM TESTING

6.1 TESTING OBJECTIVES

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6.2 TYPES OF TESTS

In order to uncover the errors present in different phases we have the concept of levels of testing. The basic levels of testing are



6.2.1 UNIT TEST CASES

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

In this project the unit testing validates the program logic of all modules in range of machine learning and big data.

The testing takes place as:

Input: Module select, Text, drop down boxes

Output: Adapt user interface based on module select. Show information on getting input required for each operations, to show required output.

6.2.2 FUNCTIONAL TEST CASE

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centred on the following items:

Valid Input: Select Text, Image, Data.

Functions: Inputs must be processed based on the user requirements and Machine Learning algorithm must be executed.

Output: Probability of data is carried out.

Systems/Procedure: Health care system or algorithm must be invoked.

In this project Functional testing is applied to the functionality of all modules (input and output of the modules) and hence it is verified.

Input: Selection and user inputs

Output: Change of user interfaces according to user problems, conversion of sender input to recipient's mode, translation of input, recognition, detection.

6.2.3 INTEGRATION TEST CASES

Integration tests are designed to test integrated software and hardware components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing. The combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

In this project the integration takes place in hardware components and software testing is verified.

Input: Proper functional testing of the system (Device), processing speed, Memory, Data usage.

Output: System working well, instantaneous output, less memory consumption and Data usage.

Front End: HTML, CSS, JavaScript.

Back End: Machine Learning.

6.2.4 SYSTEM TEST CASES

System testing manages that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process, identity errors, links, and integration points.

In this project, the system testing is verified.

- Coding.
- Machine learning Algorithms.
- Debugging coding errors.

6.2.5 ACCEPTANCE TEST CASES

User acceptance testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

The main types of software testing are

Components: Hardware is tested.

Interface: Functionality of the algorithm is tested.

System: step by step procedure of working is tested.

Acceptance: Input is accepted as per the functionality.

Release: Output is produced according to the input given.

In this project the acceptance testing is verified.

6.2.6 WHITE BOX TEST CASES

White box testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

In this project, white box testing takes place as follows

- Testing of Back end through the database provided.
- Testing of Front end through the input provided and relevant output produced for the inputs through the coding logic and verified.

6.2.7 BLACK BOX TEST CASES

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, must be written from a definitive source document, like specification or requirements documents, such as specification or requirements document. It is attesting in which the software under test is treated, as a black box. The test provides inputs and responds to outputs without considering how the software works.

In this project Black box testing is the final step which is the documentation process with the theoretical information and does not involve the practical step by step procedure, hence the testing is performed and verified.

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENT

7.1 CONCLUSION

cigarette use has increased among adolescents. This is a worldwide problem, because it has been stated in many researches mentioned in the introduction that e-cigarette use can cause future use of cigarettes. Since e-cigarette is a recent rising issue, there is little research done on this topic, compared to smoking cigarettes. Even among the researches done, there is a lack of researches implementing prediction models, which are more practical in preventing adolescents from using cigarettes. Thus, we researched using the 2018 NYTS data and developed multiple prediction models to predict an adolescent's intention to smoke cigarette. The most accurate prediction model was Gradient Boosting Classifier with an overall accuracy of 93%. This model was applied in the website we designed to allow the public to input their information in respect to tobacco products, including e-cigarette, cigarette, and cigar. With this information, the algorithm can predict the respondee's probability of future of smoking. This will help the public become more aware about certain factors in their lives and be attentive about their drug use or how their environment can affect their intention to smoke cigarettes.

7.2 FUTURE ENHANCEMENT

This project can be further improved by including a wider range of ages, since our project is mainly focused on adolescents rather than adults. In order to improve the accuracy of the prediction model, it is essential to increase the amount of data or choose better, more fitting, variables.

APPENDICES

APPENDIX 1

SAMPLE CODING

NAÏVE BAYES CLASSIFIER:

```
from sklearn.preprocessing import MinMaxScaler

from sklearn.naive_bayes import GaussianNB

from sklearn.externals import joblib

import numpy as np

import pandas as pd

import pickle

import os

from sklearn.metrics import accuracy_score

from sklearn.model_selection import train_test_split,KFold

#root = os.path.dirname(__file__)

#path_df = os.path.join(root, 'recons_dataset/combined_dataset.csv')

data = pd.read_csv('dataset7.csv')

scaler = MinMaxScaler()

train, test = train_test_split(data, test_size=0.10)

X_train = train.drop('Q7', axis=1)

Y_train = train['Q7']

X_test = test.drop('Q7', axis=1)

Y_test = test['Q7']
```

```
# We don't scale targets: Y_test, Y_train as SVC returns the class labels not probability values

X_train = scaler.fit_transform(X_train)

X_test = scaler.fit_transform(X_test)

clf1 = GaussianNB()

# Training the classifier

clf1.fit(X_train, Y_train)

# Testing model accuracy. Average is taken as test set is very small hence accuracy varies a lot everytime the model is trained

acc = 0

acc_binary = 0

for i in range(0, 20):

    Y_hat = clf1.predict(X_test)

    Y_hat_bin = Y_hat>0

    Y_test_bin = Y_test>0

    acc = acc + accuracy_score(Y_hat, Y_test)

    acc_binary = acc_binary + accuracy_score(Y_hat_bin, Y_test_bin)

print("Average test Accuracy:{}".format(acc/20))

print("Average binary accuracy:{}".format(acc_binary/20))

# Saving the trained model for inference

#model_path = os.path.join(root, 'models/rfc.sav')

#joblib.dump(clf, model_path)
```

```
# Saving the scaler object
#scaler_path = os.path.join(root, 'models/scaler.pkl')
#with open(scaler_path, 'wb') as scaler_file:
#    pickle.dump(scaler, scaler_file)
x = np.array([5,1,2,2,2,2,1,5,2,1,3,1,1,1,2,1,1,1,1,4,3,5,1,3,4,5,2,2,3,1,4,
3,3,2,4,3,1,1,1,1,1,2]).reshape(1, -1)
x = scaler.transform(x)
y = clf1.predict(x)
predictions = clf1.predict_proba(x)
print(predictions)
print(y)
```

DECISION TREE CLASSIFIER:

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.externals import joblib
import numpy as np
import pandas as pd
import pickle
import os
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn import tree
```

```
#root = os.path.dirname(__file__)

#path_df = os.path.join(root, 'recons_dataset/combined_dataset.csv')

data = pd.read_csv('dataset1.csv')

scaler = MinMaxScaler()

train, test = train_test_split(data, test_size=0.10)

X_train = train.drop('Q7', axis=1)

Y_train = train['Q7']

X_test = test.drop('Q7', axis=1)

Y_test = test['Q7']

# We don't scale targets: Y_test, Y_train as SVC returns the class labels not
probability values

X_train = scaler.fit_transform(X_train)

X_test = scaler.fit_transform(X_test)

clf1 = tree.DecisionTreeClassifier()

# Training the classifier

clf1.fit(X_train, Y_train)

# Testing model accuracy. Average is taken as test set is very small hence
accuracy varies a lot everytime the model is trained

acc = 0

acc_binary = 0

for i in range(0, 20):

    Y_hat = clf1.predict(X_test)
    Y_hat_bin = Y_hat>0
    Y_test_bin = Y_test>0
    acc = acc + accuracy_score(Y_hat, Y_test)
    acc_binary = acc_binary + accuracy_score(Y_hat_bin, Y_test_bin)
print("Average test Accuracy:{}".format(acc/20))
print("Average binary accuracy:{}".format(acc_binary/20))
# Saving the trained model for inference
#model_path = os.path.join(root, 'models/rfc.sav')
#joblib.dump(clf, model_path)
# Saving the scaler object
#scaler_path = os.path.join(root, 'models/scaler.pkl')
#with open(scaler_path, 'wb') as scaler_file:
#    pickle.dump(scaler, scaler_file)
x = 
np.array([5,1,2,2,2,2,1,5,2,1,3,1,1,1,2,1,1,1,1,4,3,5,1,3,4,5,2,2,3,1,4,3,3,2,4,3,1,1
,1,1,1,2]).reshape(1, -1)
x = scaler.transform(x)
y = clf1.predict(x)
predictions = clf1.predict_proba(x)
print(predictions)
print(y)
```

RANDOM FOREST CLASSIFIER:

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.externals import joblib
import numpy as np
import pandas as pd
import pickle
import os
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
root = os.path.dirname(__file__)
path_df = os.path.join(root, 'recons_dataset/combined_dataset.csv')
data = pd.read_csv(path_df)
scaler = MinMaxScaler()
train, test = train_test_split(data, test_size=0.25)
X_train = train.drop('num', axis=1)
Y_train = train['num']
X_test = test.drop('num', axis=1)
Y_test = test['num']
# We don't scale targets: Y_test, Y_train as SVC returns the class labels not probability values
X_train = scaler.fit_transform(X_train)
X_test = scaler.fit_transform(X_test)
```

```
clf = RandomForestClassifier()
# Training the classifier
clf.fit(X_train, Y_train)
# Testing model accuracy. Average is taken as test set is very small hence accuracy varies a lot everytime the model is trained
acc = 0
acc_binary = 0
for i in range(0, 20):
    Y_hat = clf.predict(X_test)
    Y_hat_bin = Y_hat>0
    Y_test_bin = Y_test>0
    acc = acc + accuracy_score(Y_hat, Y_test)
    acc_binary = acc_binary +accuracy_score(Y_hat_bin, Y_test_bin)
print("Average test Accuracy:{}".format(acc/20))
print("Average binary accuracy:{}".format(acc_binary/20))
# Saving the trained model for inference
model_path = os.path.join(root, 'models/rfc.sav')
joblib.dump(clf, model_path)
# Saving the scaler object
scaler_path = os.path.join(root, 'models/scaler.pkl')
with open(scaler_path, 'wb') as scaler_file:
    pickle.dump(scaler, scaler_file)
```

GRADIENT BOOSTING CLASSIFIER

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.externals import joblib
from sklearn.ensemble import GradientBoostingClassifier
import numpy as np
import pandas as pd
import pickle
import os
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
root = os.path.dirname(__file__)
path_df = os.path.join(root, 'recons_dataset/dataset1.csv')
data = pd.read_csv(path_df)
scaler = MinMaxScaler()
train, test = train_test_split(data, test_size=0.10)
X_train = train.drop('Q15', axis=1)
Y_train = train['Q15']
X_test = test.drop('Q15', axis=1)
Y_test = test['Q15']
# We don't scale targets: Y_test, Y_train as SVC returns the class labels not probability values
X_train = scaler.fit_transform(X_train)
```

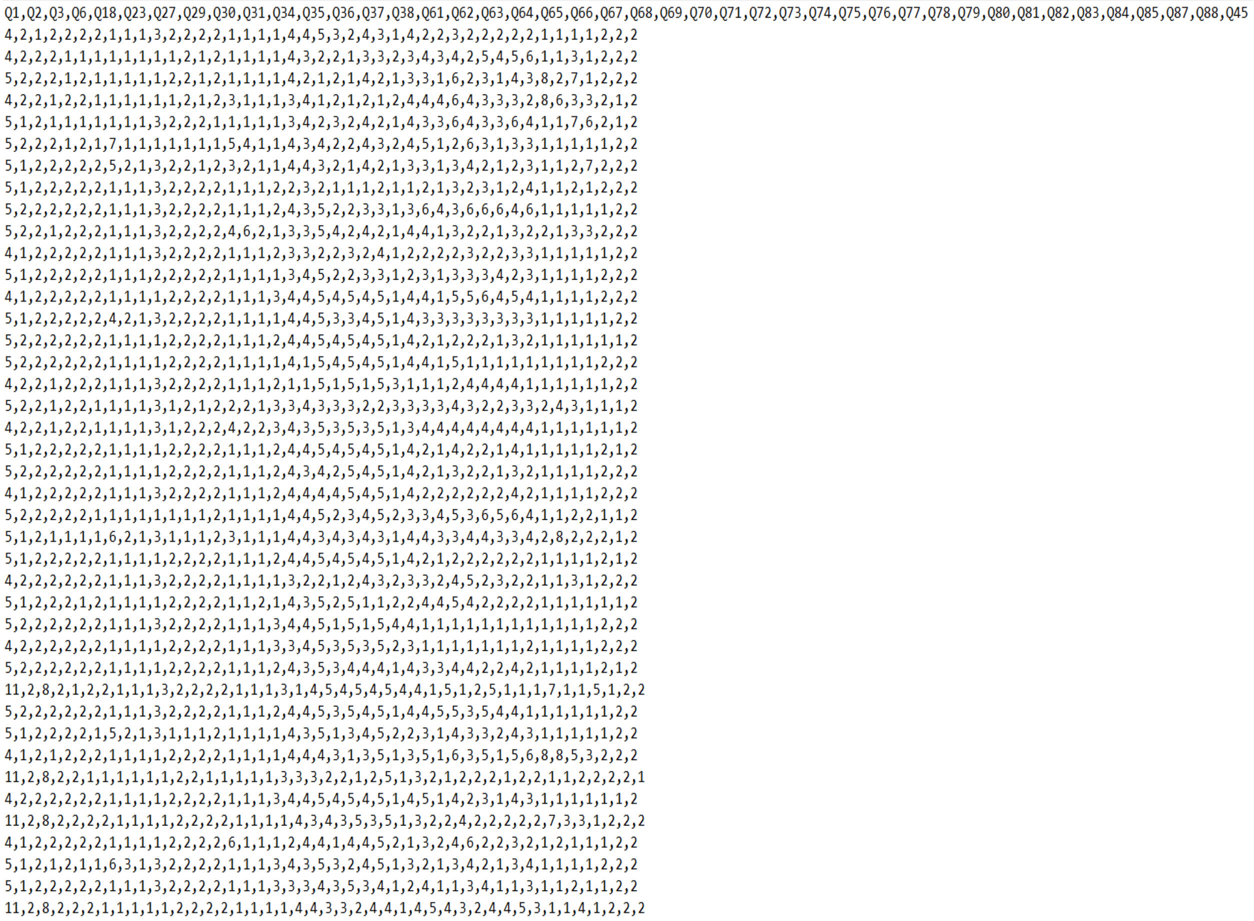
```
X_test = scaler.fit_transform(X_test)
clf1 = GradientBoostingClassifier()
clf1.fit(X_train, Y_train)
# Testing model accuracy. Average is taken as test set is very small hence accuracy varies a lot everytime the model is trained
acc = 0
acc_binary = 0
for i in range(0, 200):
    Y_hat = clf1.predict(X_test)
    Y_hat_bin = Y_hat>0
    Y_test_bin = Y_test>0
    acc = acc + accuracy_score(Y_hat, Y_test)
    acc_binary = acc_binary + accuracy_score(Y_hat_bin, Y_test_bin)
print("Average test Accuracy:{}".format(acc/200))
print("Average binary accuracy:{}".format(acc_binary/200))
# Saving the trained model for inference
model_path = os.path.join(root, 'models/rfc1.sav')
joblib.dump(clf1, model_path)
# Saving the scaler object
scaler_path = os.path.join(root, 'models/scaler1.pkl')
with open(scaler_path, 'wb') as scaler_file:
    pickle.dump(scaler, scaler_file)
```

APPENDIX 2

SCREEN SHOTS

MODULE-1: BUSINESS UNDERSTANDING AND DATA UNDERSTANDING

- Business Understanding is introducing the objectives of the project and understanding what the goals will be. Knowing the goals and the objectives, a data mining problem is formed and a plan to achieve these goals is made.
- Data Understanding includes finding an initial data set, familiarizing with the contents in the data set, and making observations from the data set in order to create a hypothesis. The initial data set must be reliable and accurate, which means that it is not outdated and has correlation within itself.

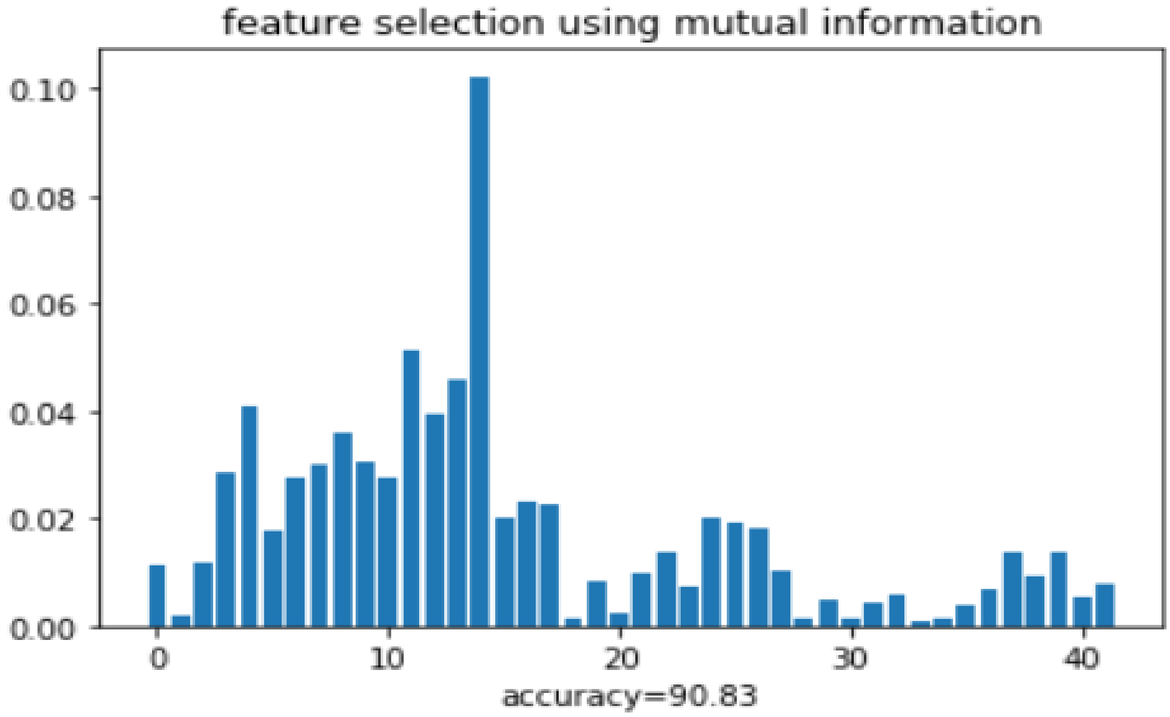


```
Q1,Q2,Q3,Q6,Q18,Q23,Q27,Q29,Q30,Q31,Q34,Q35,Q36,Q37,Q38,Q61,Q62,Q63,Q64,Q65,Q66,Q67,Q68,Q69,Q70,Q71,Q72,Q73,Q74,Q75,Q76,Q77,Q78,Q79,Q80,Q81,Q82,Q83,Q84,Q85,Q87,Q88,Q45
4,2,1,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,1,4,4,5,3,2,4,3,1,4,2,2,3,2,2,2,2,2,1,1,1,1,2,2,2
4,2,2,2,1,1,1,1,1,1,1,1,2,1,2,1,1,1,1,4,3,2,2,1,3,3,2,3,4,3,4,2,5,4,5,6,1,1,3,1,2,2,2
5,2,2,2,1,2,1,1,1,1,1,2,2,1,2,1,1,1,1,4,2,1,2,1,4,2,1,3,3,1,6,2,3,1,4,3,8,2,7,1,2,2,2
4,2,2,1,2,2,1,1,1,1,1,1,2,1,2,3,1,1,1,3,4,1,2,1,2,1,2,4,4,4,6,4,3,3,3,2,8,6,3,3,2,1,2
5,1,2,1,1,1,1,1,1,1,3,2,2,2,1,1,1,1,1,3,4,2,3,2,4,2,1,4,3,3,6,4,3,3,6,4,1,1,7,6,2,1,2
5,2,2,2,1,2,1,7,1,1,1,1,1,1,1,5,4,1,1,4,3,4,2,2,4,3,2,4,5,1,2,6,3,1,3,3,1,1,1,1,1,2,2
5,1,2,2,2,2,2,5,2,1,3,2,2,1,2,3,2,1,1,4,4,3,2,1,4,2,1,3,3,1,3,4,2,1,2,3,1,1,2,7,2,2,2
5,1,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,2,2,3,2,1,1,1,2,1,1,2,1,3,2,3,1,2,4,1,1,2,1,2,2,2
5,2,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,2,4,3,5,2,2,3,3,1,3,6,4,3,6,6,6,4,6,1,1,1,1,1,2,2
5,2,2,1,2,2,2,1,1,1,3,2,2,2,2,4,6,2,1,3,3,5,4,2,4,2,1,4,4,1,3,2,2,1,3,2,2,1,3,3,2,2,2
4,1,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,2,3,3,2,2,3,2,4,1,2,2,2,2,3,2,2,3,3,1,1,1,1,1,2,2
5,1,2,2,2,2,2,1,1,1,2,2,2,2,2,1,1,1,1,3,4,5,2,2,3,3,1,2,3,1,3,3,3,4,2,3,1,1,1,1,2,2,2
4,1,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,3,4,4,5,4,5,4,5,1,4,4,1,5,5,6,4,5,4,1,1,1,1,2,2,2
5,1,2,2,2,2,2,4,2,1,3,2,2,2,2,1,1,1,1,4,4,5,3,3,4,5,1,4,3,3,3,3,3,3,3,3,1,1,1,1,1,2,2
5,2,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,2,4,4,5,4,5,4,5,1,4,2,1,2,2,2,1,3,2,1,1,1,1,1,1,2
5,2,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,1,4,1,5,4,5,4,5,1,4,4,1,5,1,1,1,1,1,1,1,1,1,2,2,2
4,2,2,1,2,2,2,1,1,1,3,2,2,2,2,1,1,1,2,1,1,5,1,5,1,5,3,1,1,1,2,4,4,4,4,1,1,1,1,1,1,2,2
5,2,2,1,2,2,1,1,1,1,3,1,2,1,2,2,2,1,3,3,4,3,3,3,2,2,3,3,3,3,4,3,2,2,3,3,2,4,3,1,1,1,2
4,2,2,1,2,2,1,1,1,1,3,1,2,2,2,4,2,2,3,4,3,5,3,5,3,5,1,3,4,4,4,4,4,4,4,4,1,1,1,1,1,1,2
5,1,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,2,4,4,5,4,5,4,5,1,4,2,1,4,2,2,1,4,1,1,1,1,1,2,1,2
5,2,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,2,4,3,4,2,5,4,5,1,4,2,1,3,2,2,1,3,2,1,1,1,1,2,2,2
4,1,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,2,4,4,4,4,5,4,5,1,4,2,2,2,2,2,2,4,2,1,1,1,1,2,2,2
5,2,2,2,2,2,1,1,1,1,1,1,1,1,2,1,1,1,1,4,4,5,2,3,4,5,2,3,3,4,5,3,6,5,6,4,1,1,2,2,1,1,2
5,1,2,1,1,1,1,6,2,1,3,1,1,1,2,3,1,1,1,4,4,3,4,3,4,3,1,4,4,3,3,4,4,3,3,4,2,8,2,2,2,1,2
5,1,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,2,4,4,5,4,5,4,5,1,4,2,1,2,2,2,2,2,2,1,1,1,1,2,1,2
4,2,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,1,3,2,2,1,2,4,3,2,3,3,2,4,5,2,3,2,2,1,1,3,1,2,2,2
5,1,2,2,2,1,2,1,1,1,1,2,2,2,2,1,1,2,1,4,3,5,2,5,1,1,2,2,4,4,5,4,2,2,2,2,1,1,1,1,1,1,2
5,2,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,3,4,4,5,1,5,1,5,4,4,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2
4,2,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,3,3,4,5,3,5,3,5,2,3,1,1,1,1,1,1,1,2,1,1,1,1,2,2,2
5,2,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,2,4,3,5,3,4,4,4,1,4,3,3,4,4,2,2,4,2,1,1,1,1,2,1,2
11,2,8,2,1,2,2,1,1,1,3,2,2,2,2,1,1,1,3,1,4,5,4,5,4,5,4,4,1,5,1,2,5,1,1,1,7,1,1,5,1,2,2
5,2,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,2,4,4,5,3,5,4,5,1,4,4,5,5,3,5,4,4,1,1,1,1,1,1,2,2
5,1,2,2,2,2,1,5,2,1,3,1,1,1,2,1,1,1,1,4,3,5,1,3,4,5,2,2,3,1,4,3,3,2,4,3,1,1,1,1,1,2,2
4,1,2,1,2,2,2,1,1,1,1,2,2,2,2,1,1,1,1,4,4,4,3,1,3,5,1,3,5,1,6,3,5,1,5,6,8,8,5,3,2,2,2
11,2,8,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,3,3,3,2,2,1,2,5,1,3,2,1,2,2,2,1,2,2,1,1,2,2,2,2,1
4,2,2,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,3,4,4,5,4,5,4,5,1,4,5,1,4,2,3,1,4,3,1,1,1,1,1,1,2
11,2,8,2,2,2,2,1,1,1,1,2,2,2,2,1,1,1,1,4,3,4,3,5,3,5,1,3,2,2,4,2,2,2,2,2,7,3,3,1,2,2,2
4,1,2,2,2,2,2,1,1,1,1,2,2,2,2,6,1,1,1,2,4,4,1,4,4,5,2,1,3,2,4,6,2,2,3,2,1,2,1,1,1,2,2
5,1,2,1,2,1,1,6,3,1,3,2,2,2,2,1,1,1,3,4,3,5,3,2,4,5,1,3,2,1,3,4,2,1,3,4,1,1,1,1,2,2,2
5,1,2,2,2,2,2,1,1,1,3,2,2,2,2,1,1,1,3,3,3,4,3,5,3,4,1,2,4,1,1,3,4,1,1,3,1,1,2,1,1,2,2
11,2,8,2,2,2,1,1,1,1,1,2,2,2,2,1,1,1,1,4,4,3,3,2,4,4,1,4,5,4,3,2,4,4,5,3,1,1,4,1,2,2,2
```

MODULE-2: DATA PREPARATION

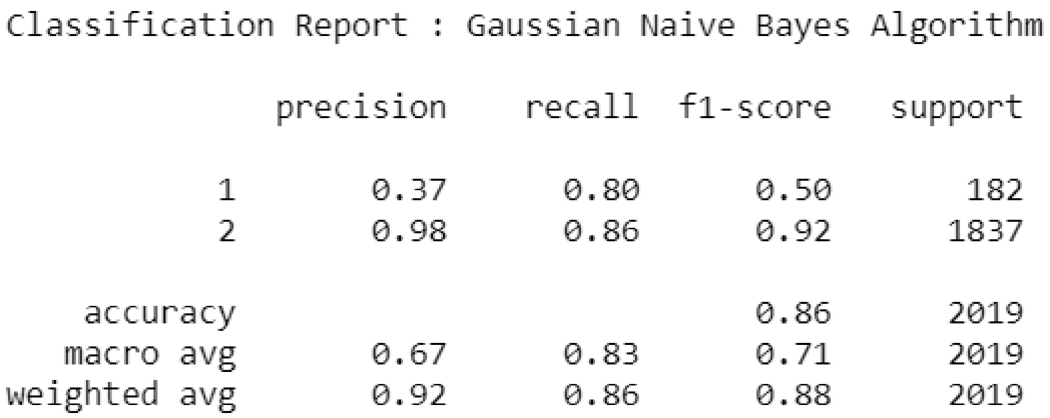
Data Preparation is the process of converting the initial, or raw, data set into the final data set. This step could alter the data to become more applicable to achieve the goal.

Input data sets are correlated with the final output using the mutual information classifier algorithm.

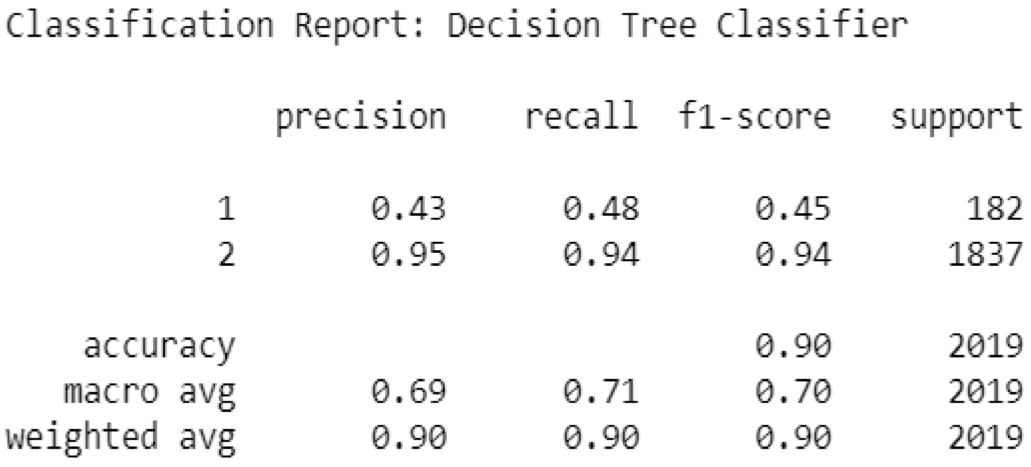


MODULE-3 MODELLING AND EVALUATION PHASE

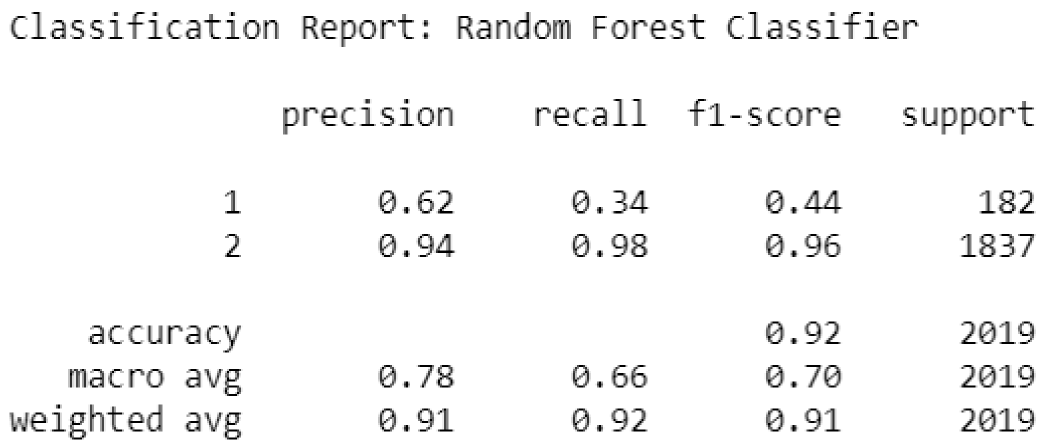
- In Modeling, artificial intelligence models are created using various machine learning algorithms. The data set prepared from the data preparation step are applied to this modeling step.
- Evaluation is reviewing performance (e.g., accuracy) of models and deciding a best model. This is the process where the models are improved based a goal performance. Success criteria can be based on speed of algorithm, memory usage, or prediction accuracy.



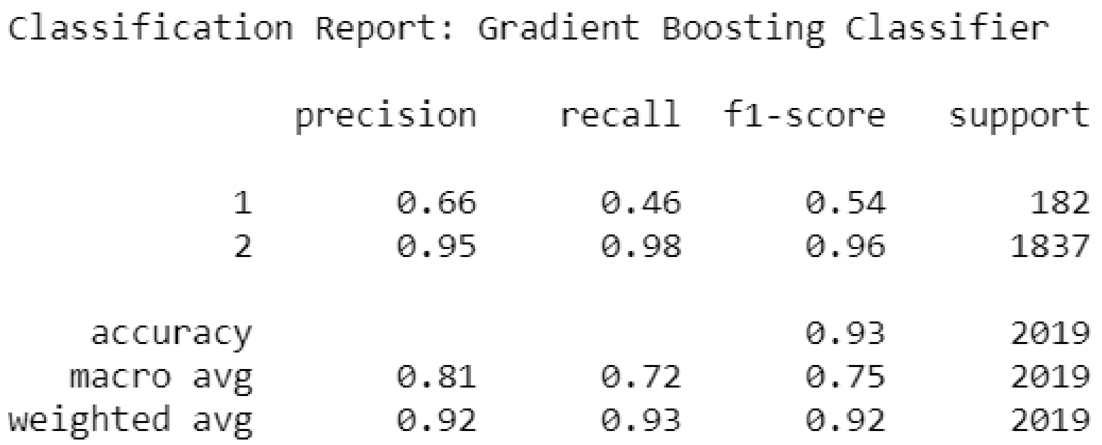
```
Classification Report : Gaussian Naive Bayes Algorithm

              precision    recall  f1-score   support

           1       0.37      0.80      0.50       182
           2       0.98      0.86      0.92      1837

    accuracy                           0.86      2019
   macro avg       0.67      0.83      0.71      2019
weighted avg       0.92      0.86      0.88      2019
```



```
Classification Report: Decision Tree Classifier

              precision    recall  f1-score   support

           1       0.43      0.48      0.45       182
           2       0.95      0.94      0.94      1837

    accuracy                           0.90      2019
   macro avg       0.69      0.71      0.70      2019
weighted avg       0.90      0.90      0.90      2019
```

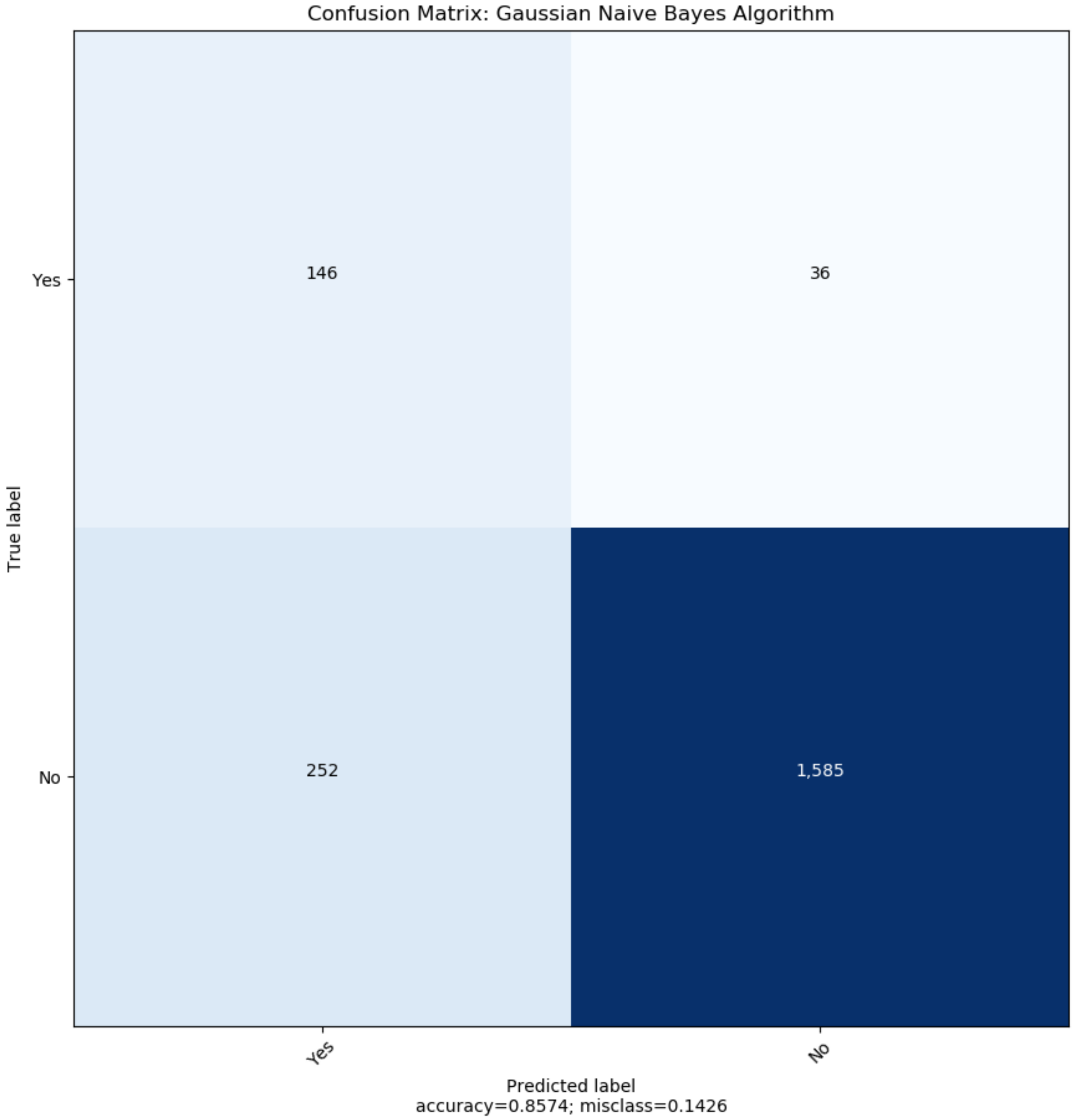


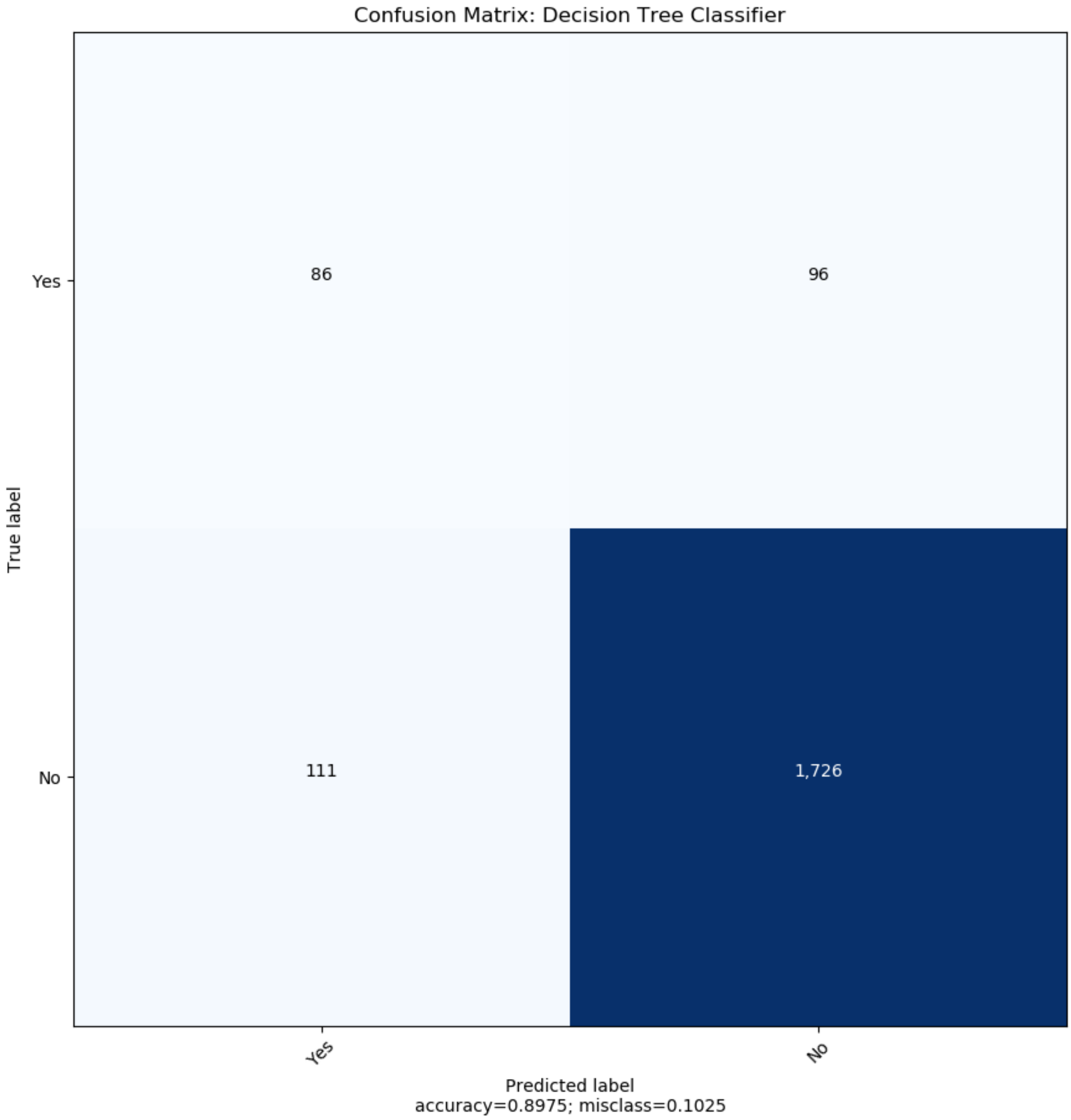
```
Classification Report: Random Forest Classifier

              precision    recall  f1-score   support

           1       0.62      0.34      0.44       182
           2       0.94      0.98      0.96      1837

    accuracy                           0.92      2019
   macro avg       0.78      0.66      0.70      2019
weighted avg       0.91      0.92      0.91      2019
```

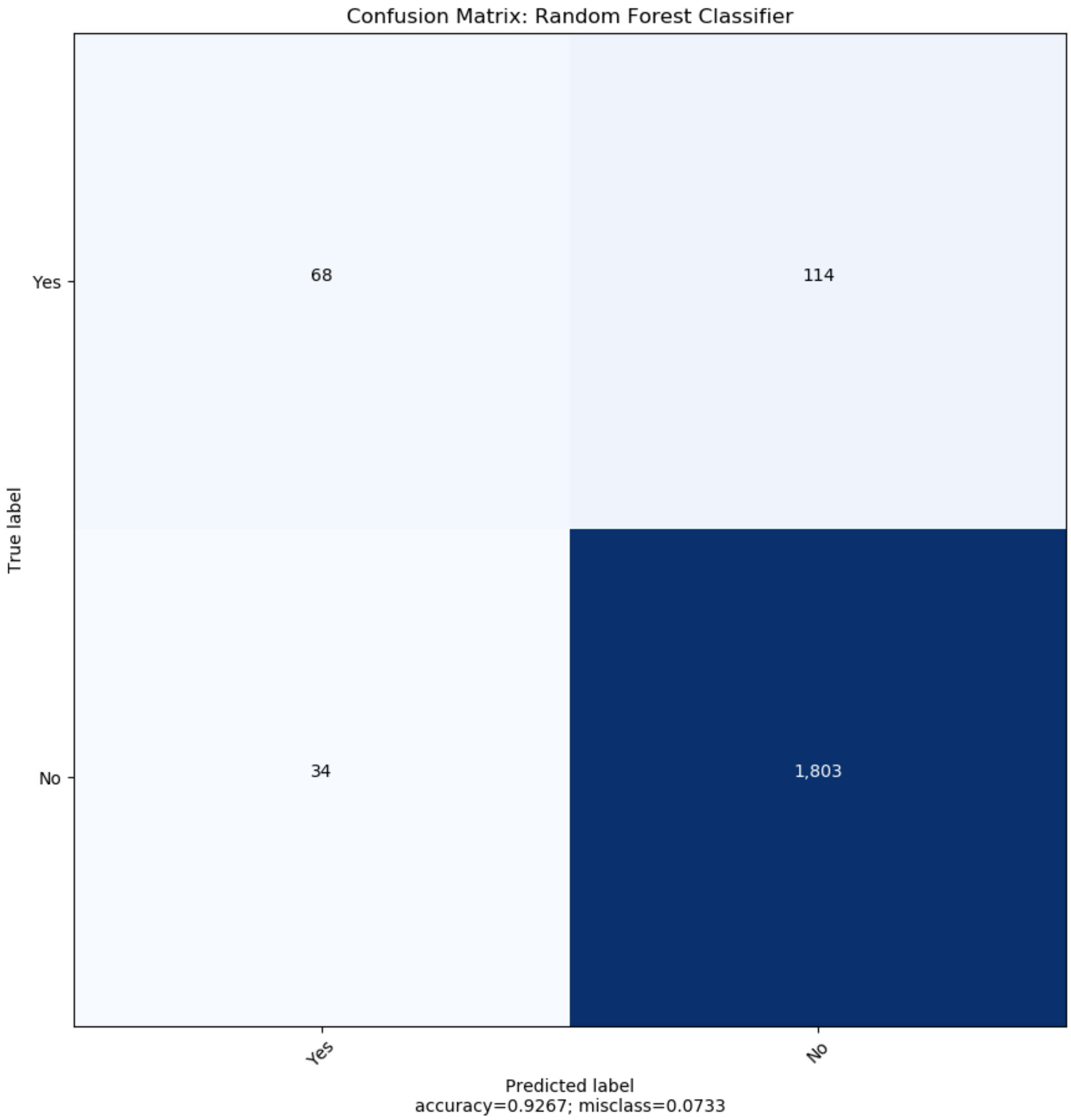


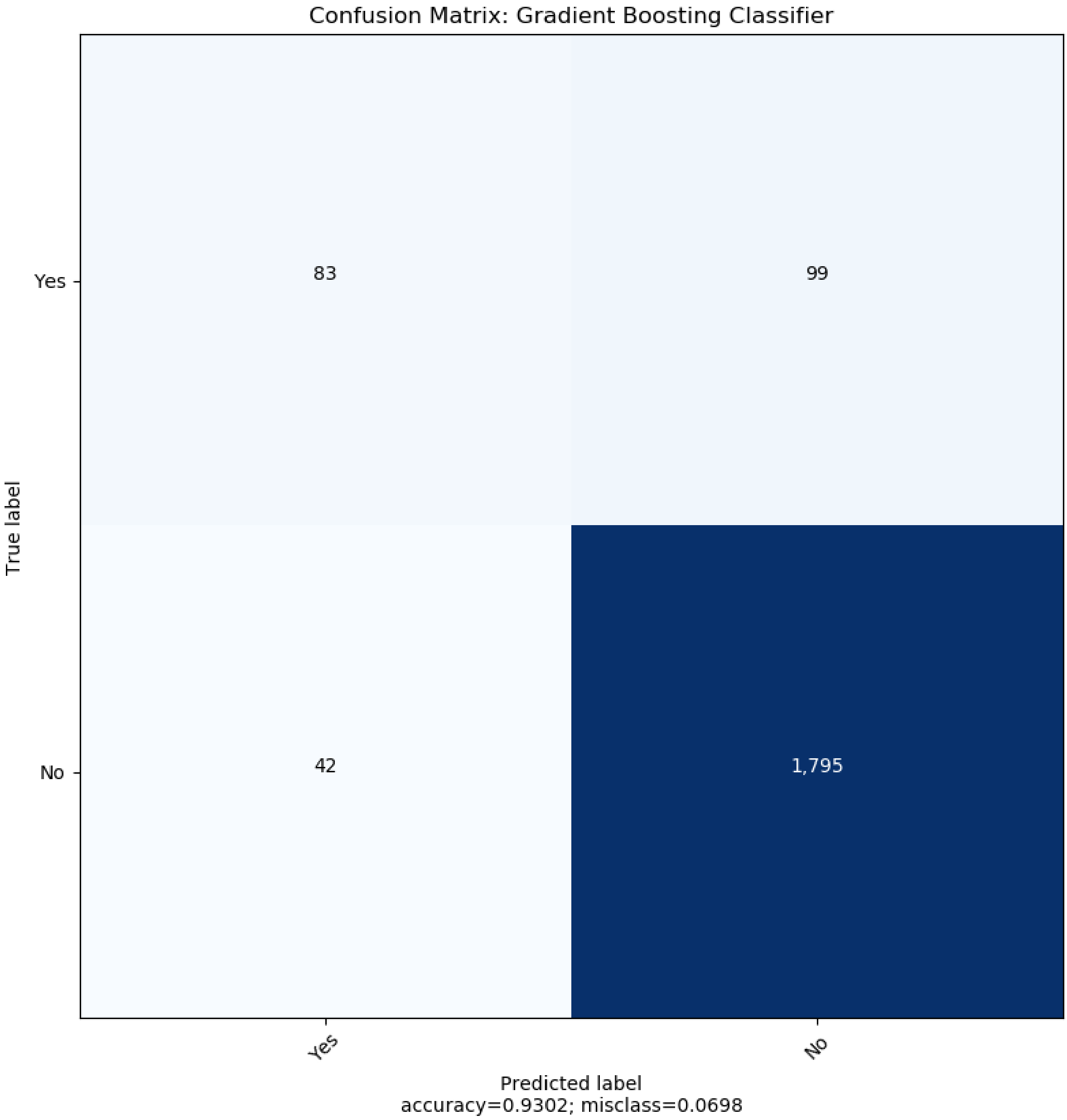
```
Classification Report: Gradient Boosting Classifier

              precision    recall  f1-score   support

           1       0.66      0.46      0.54       182
           2       0.95      0.98      0.96      1837

    accuracy                           0.93      2019
   macro avg       0.81      0.72      0.75      2019
weighted avg       0.92      0.93      0.92      2019
```

CONFUSION MATRIX



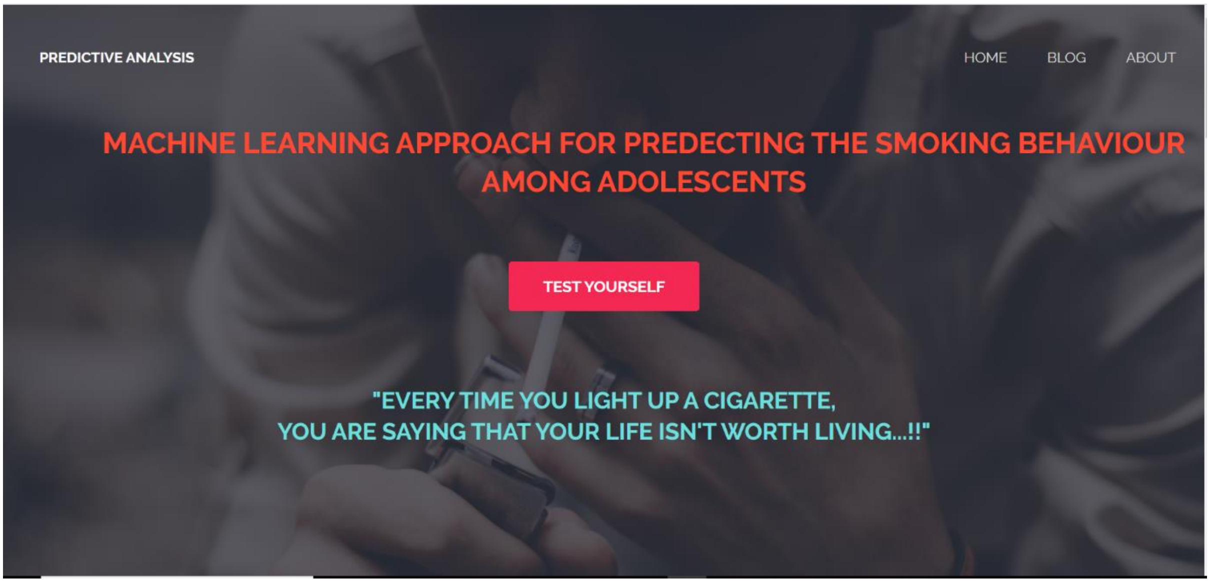

Confusion Matrix: Decision Tree Classifier
True label
Yes
No
86
96
111
1,726
Yes
No
Predicted label
accuracy=0.8975; misclass=0.1025


Confusion Matrix: Random Forest Classifier
True label
Yes
No
68
114
34
1,803
Yes
No
Predicted label
accuracy=0.9267; misclass=0.0733


Confusion Matrix: Gradient Boosting Classifier
True label
Yes
No
83
99
42
1,795
Yes
No
Predicted label
accuracy=0.9302; misclass=0.0698

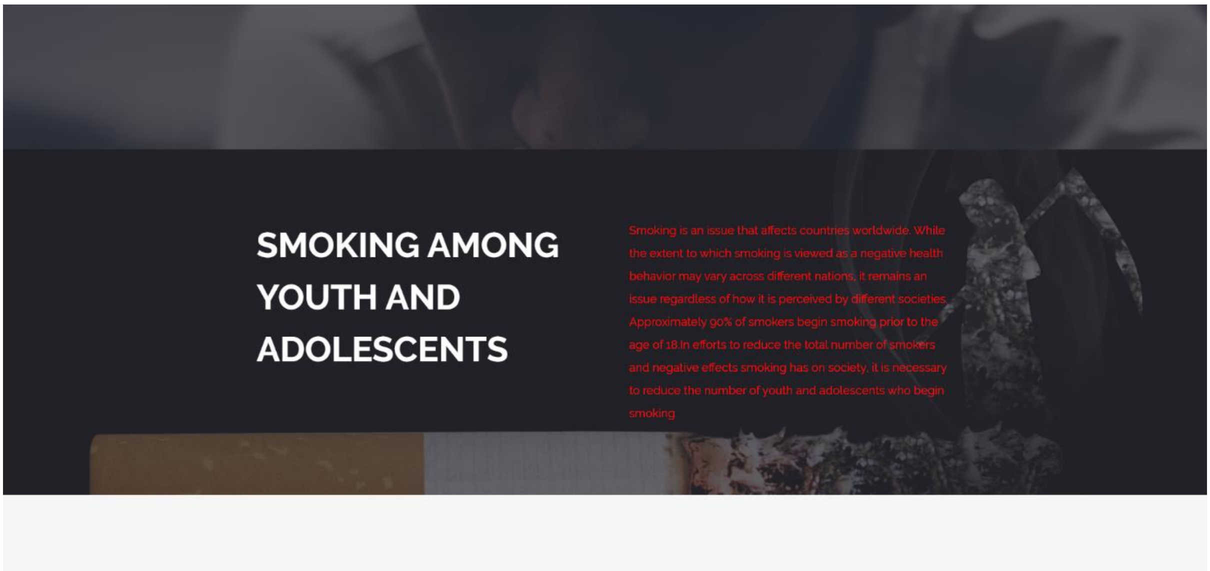
MODULE-4: DEPLOYMENT

Deployment is the process of organizing the information gained, such as the model, so that it is understandable for the model user. This step is carried out by the user rather than the analyst, so it is essential for the user to understand how to use the models.



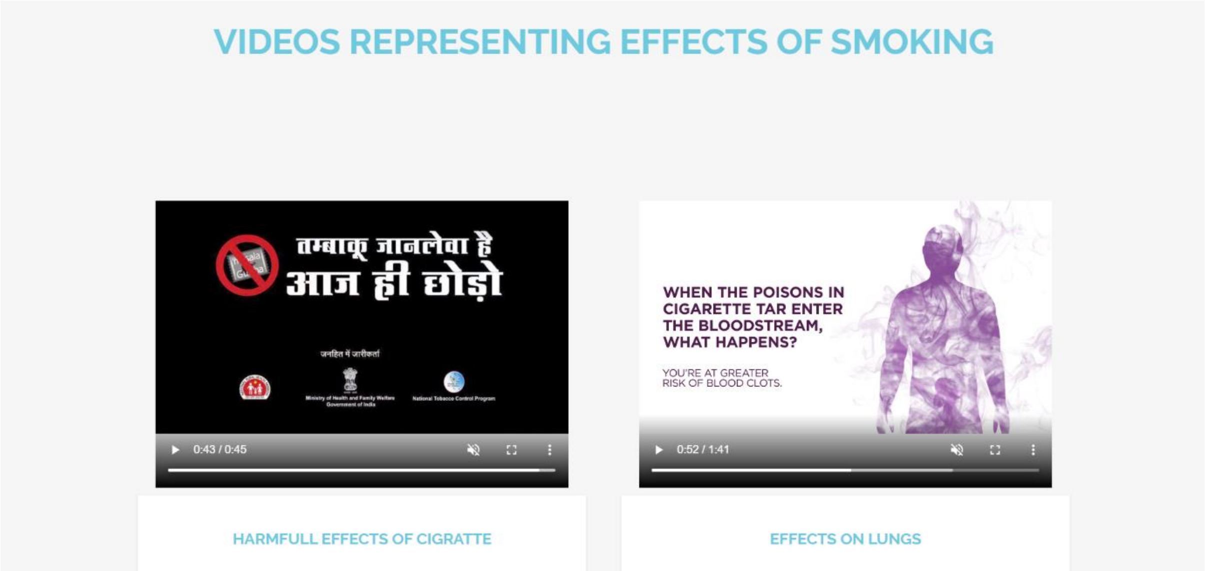
MACHINE LEARNING APPROACH FOR PREDECTING THE SMOKING BEHAVIOUR AMONG ADOLESCENTS

"EVERY TIME YOU LIGHT UP A CIGARETTE,
YOU ARE SAYING THAT YOUR LIFE ISN'T WORTH LIVING...!!"



SMOKING AMONG YOUTH AND ADOLESCENTS

Smoking is an issue that affects countries worldwide. While the extent to which smoking is viewed as a negative health behavior may vary across different nations, it remains an issue regardless of how it is perceived by different societies. Approximately 90% of smokers begin smoking prior to the age of 18.In efforts to reduce the total number of smokers and negative effects smoking has on society, it is necessary to reduce the number of youth and adolescents who begin smoking

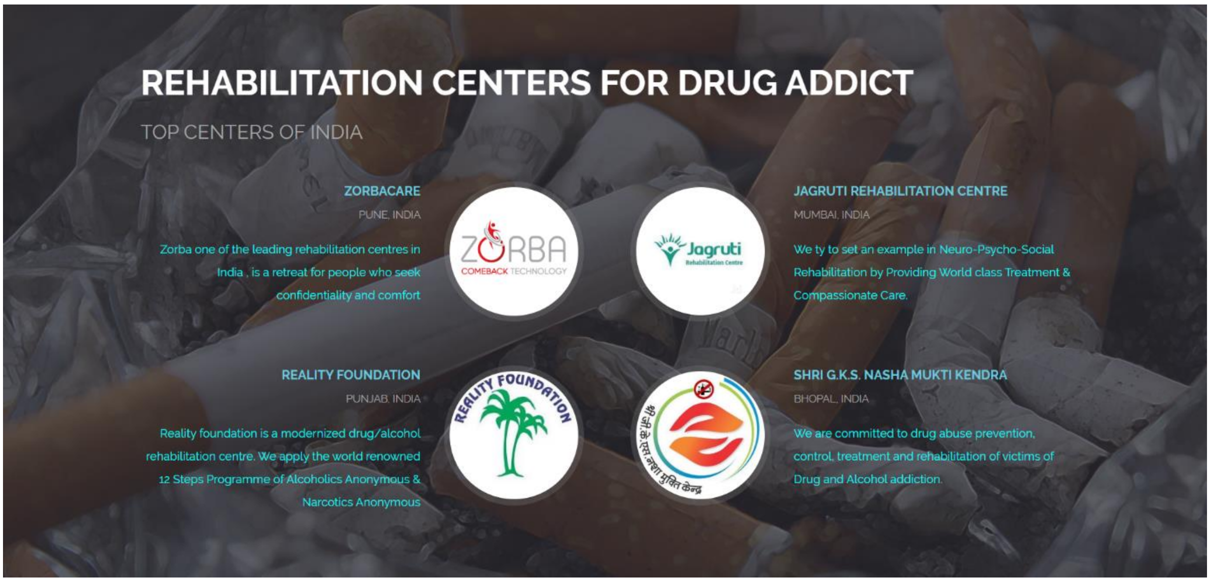


VIDEOS REPRESENTING EFFECTS OF SMOKING

तम्बाकू जानलेवा है
आज ही छोड़ो
जनहित में जारीकर्ता
Ministry of Health and Family Welfare Government of India
National Tobacco Control Program
0:43 / 0:45
WHEN THE POISONS IN CIGARETTE TAR ENTER THE BLOODSTREAM, WHAT HAPPENS?
YOU'RE AT GREATER RISK OF BLOOD CLOTS.
0:52 / 1:41

HARMFULL EFFECTS OF CIGRATTE

EFFECTS ON LUNGS



REHABILITATION CENTERS FOR DRUG ADDICT

TOP CENTERS OF INDIA

ZORBACARE
PUNE, INDIA

Zorba one of the leading rehabilitation centres in India , is a retreat for people who seek confidentiality and comfort

JAGRUTI REHABILITATION CENTRE
MUMBAI, INDIA

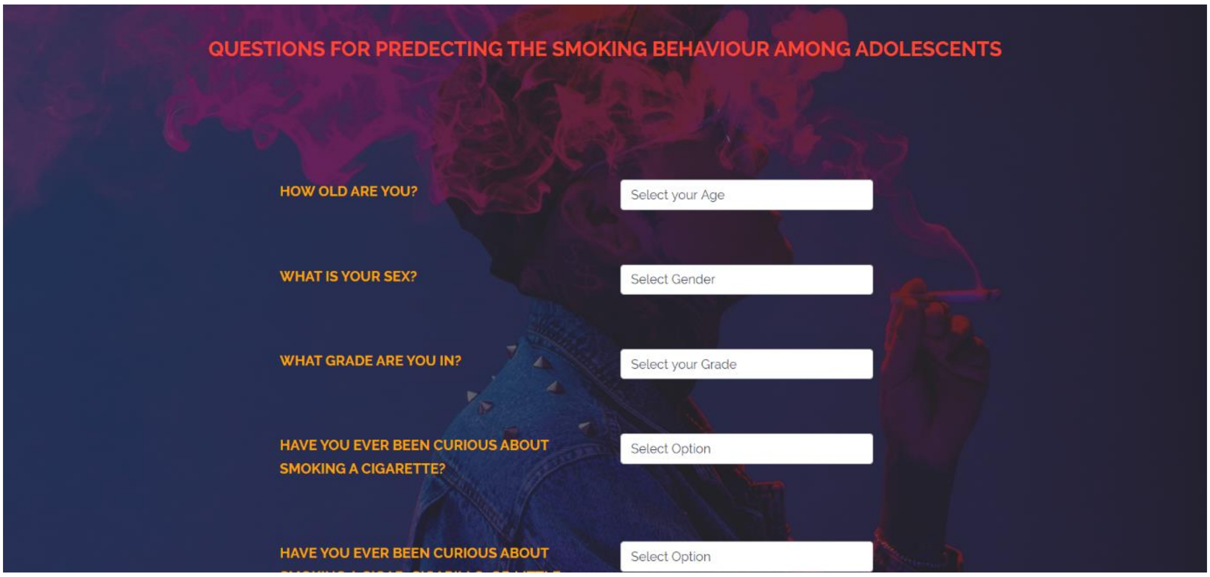
We ty to set an example in Neuro-Psycho-Social Rehabilitation by Providing World class Treatment & Compassionate Care.

REALITY FOUNDATION
PUNJAB, INDIA

Reality foundation is a modernized drug/alcohol rehabilitation centre. We apply the world renowned 12 Steps Programme of Alcoholics Anonymous & Narcotics Anonymous

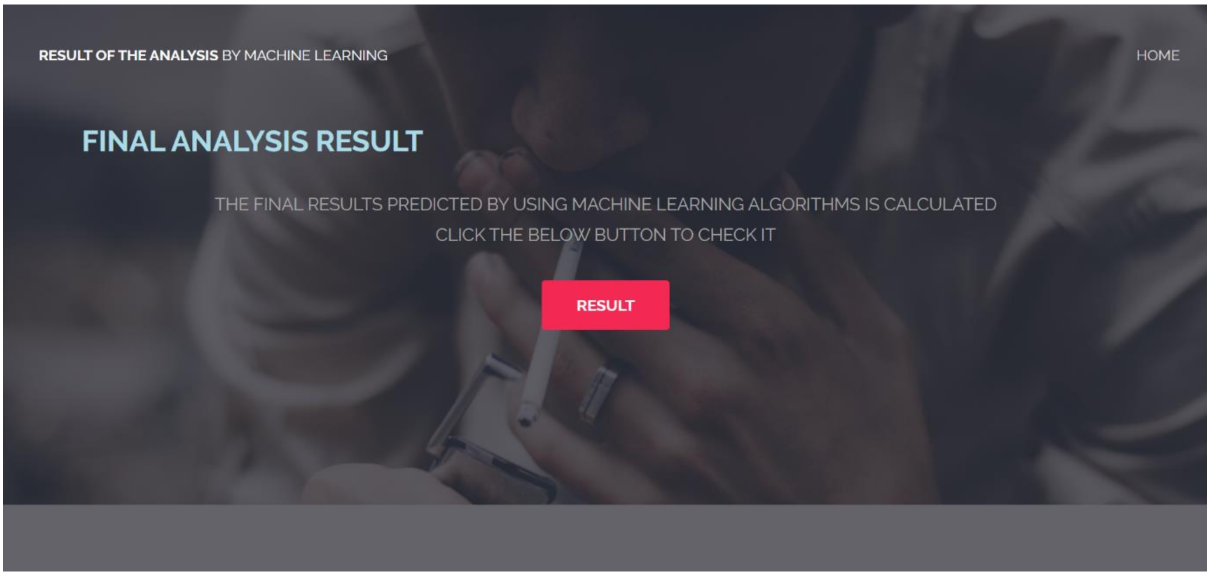
SHRI G.K.S. NASHA MUKTI KENDRA
BHOPAL, INDIA

We are committed to drug abuse prevention, control, treatment and rehabilitation of victims of Drug and Alcohol addiction.



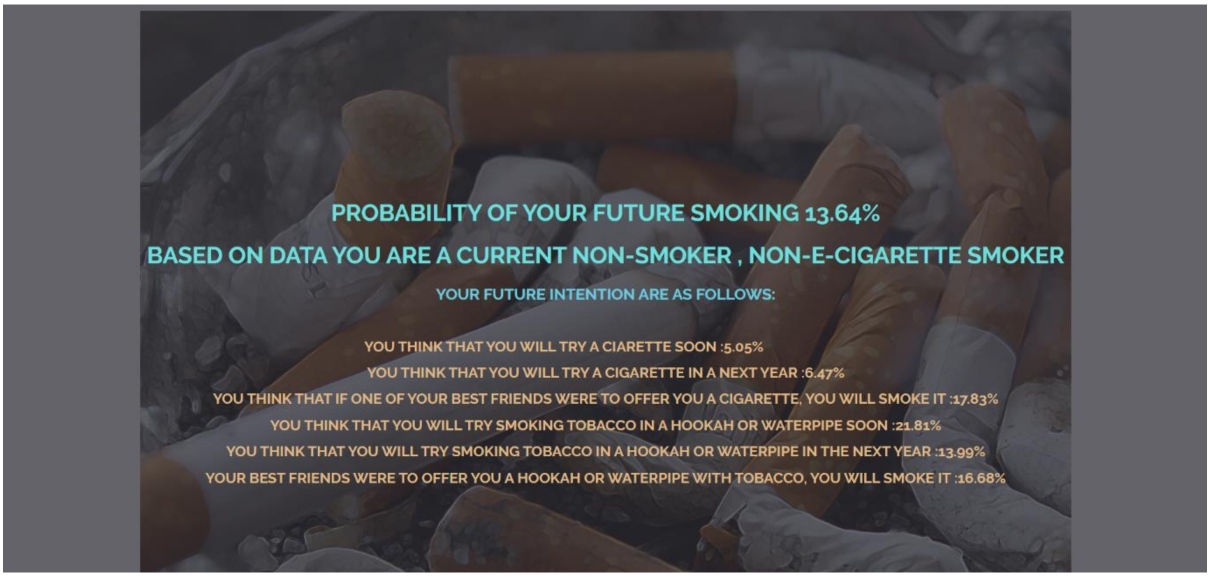
QUESTIONS FOR PREDECTING THE SMOKING BEHAVIOUR AMONG ADOLESCENTS

HOW OLD ARE YOU?	Select your Age
WHAT IS YOUR SEX?	Select Gender
WHAT GRADE ARE YOU IN?	Select your Grade
HAVE YOU EVER BEEN CURIOUS ABOUT SMOKING A CIGARETTE?	Select Option
HAVE YOU EVER BEEN CURIOUS ABOUT	Select Option



FINAL ANALYSIS RESULT

THE FINAL RESULTS PREDICTED BY USING MACHINE LEARNING ALGORITHMS IS CALCULATED
CLICK THE BELOW BUTTON TO CHECK IT

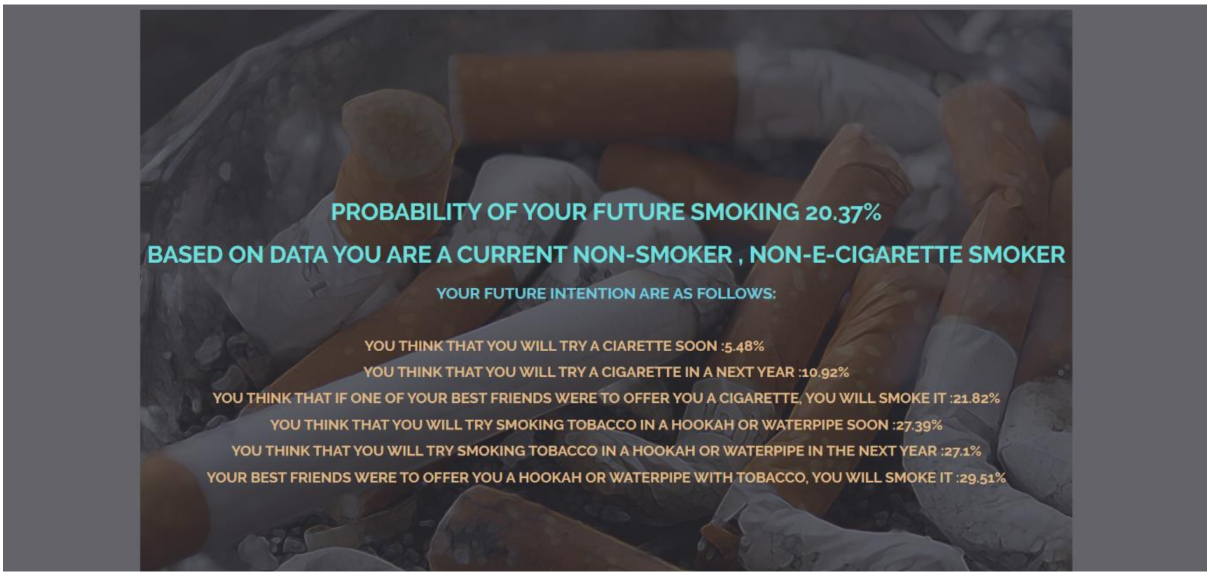


PROBABILITY OF YOUR FUTURE SMOKING 13.64%

BASED ON DATA YOU ARE A CURRENT NON-SMOKER , NON-E-CIGARETTE SMOKER

YOUR FUTURE INTENTION ARE AS FOLLOWS:

YOU THINK THAT YOU WILL TRY A CIARETTE SOON :5.05%
YOU THINK THAT YOU WILL TRY A CIGARETTE IN A NEXT YEAR :6.47%
YOU THINK THAT IF ONE OF YOUR BEST FRIENDS WERE TO OFFER YOU A CIGARETTE, YOU WILL SMOKE IT :17.83%
YOU THINK THAT YOU WILL TRY SMOKING TOBACCO IN A HOOKAH OR WATERPIPE SOON :21.81%
YOU THINK THAT YOU WILL TRY SMOKING TOBACCO IN A HOOKAH OR WATERPIPE IN THE NEXT YEAR :13.99%
YOUR BEST FRIENDS WERE TO OFFER YOU A HOOKAH OR WATERPIPE WITH TOBACCO, YOU WILL SMOKE IT :16.68%



PROBABILITY OF YOUR FUTURE SMOKING 20.37%

BASED ON DATA YOU ARE A CURRENT NON-SMOKER , NON-E-CIGARETTE SMOKER

YOUR FUTURE INTENTION ARE AS FOLLOWS:

YOU THINK THAT YOU WILL TRY A CIARETTE SOON :5.48%

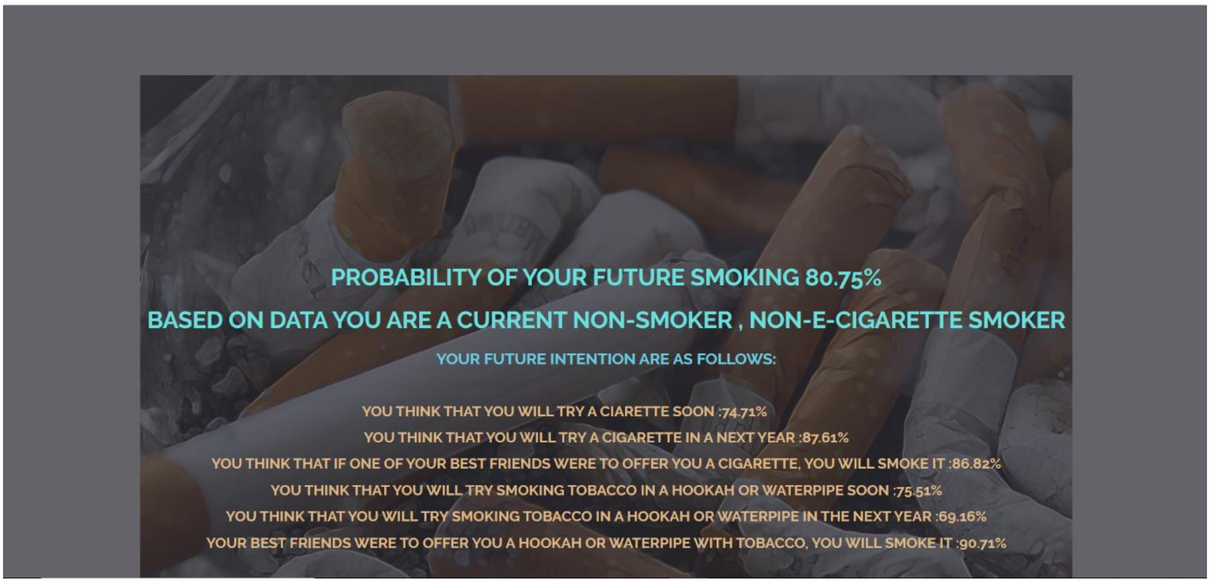
YOU THINK THAT YOU WILL TRY A CIGARETTE IN A NEXT YEAR :10.92%

YOU THINK THAT IF ONE OF YOUR BEST FRIENDS WERE TO OFFER YOU A CIGARETTE, YOU WILL SMOKE IT :21.82%

YOU THINK THAT YOU WILL TRY SMOKING TOBACCO IN A HOOKAH OR WATERPIPE SOON :27.39%

YOU THINK THAT YOU WILL TRY SMOKING TOBACCO IN A HOOKAH OR WATERPIPE IN THE NEXT YEAR :27.1%

YOUR BEST FRIENDS WERE TO OFFER YOU A HOOKAH OR WATERPIPE WITH TOBACCO, YOU WILL SMOKE IT :29.51%



PROBABILITY OF YOUR FUTURE SMOKING 80.75%

BASED ON DATA YOU ARE A CURRENT NON-SMOKER , NON-E-CIGARETTE SMOKER

YOUR FUTURE INTENTION ARE AS FOLLOWS:

YOU THINK THAT YOU WILL TRY A CIARETTE SOON :74.71%

YOU THINK THAT YOU WILL TRY A CIGARETTE IN A NEXT YEAR :87.61%

YOU THINK THAT IF ONE OF YOUR BEST FRIENDS WERE TO OFFER YOU A CIGARETTE, YOU WILL SMOKE IT :86.82%

YOU THINK THAT YOU WILL TRY SMOKING TOBACCO IN A HOOKAH OR WATERPIPE SOON :75.51%

YOU THINK THAT YOU WILL TRY SMOKING TOBACCO IN A HOOKAH OR WATERPIPE IN THE NEXT YEAR :69.16%

YOUR BEST FRIENDS WERE TO OFFER YOU A HOOKAH OR WATERPIPE WITH TOBACCO, YOU WILL SMOKE IT :90.71%

REFERENCES

1. Abdullah H. Alsharif, Nada Philip. (2016) ‘Data mining technique for the Enhanced smoking cessation management system (smoke mind)’ International Conference on Big Data.
2. Chen ke, GAO Ying. (2018) ‘Smoking Probability and Cigarettes Consumption of Chinese People’ International Journal of Machine Learning and Computing.
3. Khishigsuren Davagdorj, Keun Ho Ryu. (2019) ‘A Machine Learning Approach for Predicting Success in Smoking Cessation Intervention’ International Journal of Machine Learning and Computing.
4. Shilpa Balan, Nishant Shristiraj, Vrunda Shah and Anusha Manjappa. (2017) ‘Big Data Analysis of Youth Tobacco Smoking Trends in the US’ International Conference on Big Data.
5. Yupu Zhang, Jinhai Liu, Zhihang Zhang, Junnan Huang. (2019) ‘Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm’ International Conference on Big Data.
6. **Website:** machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning
7. **Website:** towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
8. National Youth Tobacco Survey(NYTS). https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm. Accessed: 2019-04-20.
9. National Youth Tobacco Survey. https://www.healthypeople.gov/2020/data-source/national-youth-tobacco-survey. Accessed: 2019-08-10.