STUDENT PERFORMANCE ANALYZER

A Project Report Submitted in the
Partial Fulfillment of the Requirements
for the Award of the Degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Pranith Reddy Sankepally	**17881A0524**
Srricharan Bolisetti	**17881A0548**
K.Sai Rohith Reddy	**17881A0536**
Lakkireddy Rohith Reddy	**17881A0530**

SUPERVISOR

Mr.Manoj Kumar Vemula

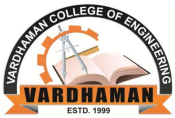
Assistant Professor

Department of Computer Science and Engineering

VARDHAMAN COLLEGE OF ENGINEERING, HYDERABAD

An Autonomous Institute, Affiliated to JNTUH

May, 2021



VARDHAMAN COLLEGE OF ENGINEERING, HYDERABAD

An Autonomous Institute, Affiliated to JNTUH

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project titled **STUDENT PERFORMANCE ANALYZER** is carried out by

Pranith Reddy Sankepally	**17881A0524**
Srricharan Bolisetti	**17881A0548**
K.Sai Rohith Reddy	**17881A0536**
Lakkireddy Rohith Reddy	**17881A0530**

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** during the year 2020-21.

Signature of the Supervisor
Mr.Manoj Kumar Vemula
Assistant Professor

Signature of the HOD
Dr. Rajanikanth Aluvalu
Professor and Head, CSE

Kacharam (V), Shamshabad (M), Ranga Reddy (Dist.)–501218, Hyderabad, T.S.
Ph: 08413-253335, 253201, Fax: 08413-253482, www.vardhaman.org

Acknowledgement

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **Mr.Manoj Kumar Vemula**, Assistant Professorand Project Supervisor, Department of Computer Science and Engineering, Vardhaman College of Engineering, for his able guidance and useful suggestions, which helped us in completing the project in time.

We are particularly thankful to **Dr. Rajanikanth Aluvalu**, the Head of the Department, Department of Computer Science and Engineering, his guidance, intense support and encouragement, which helped us to mould our project into a successful one.

We show gratitude to our honorable Principal **Dr. J.V.R. Ravindra**, for providing all facilities and support.

We avail this opportunity to express our deep sense of gratitude and heartful thanks to **Dr. Teegala Vijender Reddy**, Chairman and **Sri Teegala Upender Reddy**, Secretary of VCE, for providing a congenial atmosphere to complete this project successfully.

We also thank all the staff members of Electronics and Communication Engineering department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

Pranith Reddy Sankepally

Srricharan Bolisetti

K.Sai Rohith Reddy

Lakkireddy Rohith Reddy

Abstract

Creating tools to help students and learning in a customary or web-based setting is a significant undertaking in the present educational environment. The initial steps towards empowering such advances utilizing machine learning procedures focused on anticipating the student's performance in terms of accomplished grades. The hindrance of these methodologies is that they don't proceed to perform well in predicting poor-performing students. The goal of our work is two-overlay. To begin with, to beat this limitation, we investigate if ineffectively performing students can be more precisely predicted by figuring the issue as binary classification. Second, to acquire experiences concerning which are the factors that can prompt poor performance, we designed various human-interpretable features that measure these factors. These factors were gotten from the student's evaluations from the University of Minnesota, an undergrad public organization. In view of these factors, perform a study to distinguish diverse student groups of interest, while simultaneously, recognize their significance.

Keywords: Decision Tree Algorithm, Random Forest Algorithm, Feature Extraction, Confusion Matrix, Precision, Accuracy.

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Description
VCE	Vardhaman College of Engineering
CMOS	Complementary Metal Oxide Semiconductor

CHAPTER 1

Introduction

1.1 Problem statement

Higher educational institutions constantly try to improve the retention and success of their enrolled students. According to the US National Center for Education Statistics [8], 60undergraduate students on four-year degrees will not graduate at the same institution where they started within the first six years. At the same time, 30first year of college. As a result, colleges look for ways to serve students more efficiently and effectively. This is where data mining is introduced to provide some solutions to these problems. Educational data mining and learning analytics have been developed to provide tools for supporting the learning process, like monitor and measure student progress, but also, predict success or Most of the existing approaches focus on identifying students at risk who could benefit from further assistance in order to successfully complete a course or activity. A fundamental task in this process is to actually predict the student‘s performance in terms of grades. While reasonable prediction accuracy has been achieved [14, 10], there is a significant weakness of the models proposed to identify the poor-performing students [18]. Usually, these models tend to be over-optimistic for the performance of students, as the majority of the students do well, or have satisfactory enough performance.

1.2 Motivation

However, success and failure can be relative or not. For example, a B grade might be considered a bad grade for an excellent student, while being a good grade for a very weak student. We investigated different ways to define groups of students taking a course: failing students, students dropping the

class, students performing worse than expected and students performing worse than expected, while taking into consideration the difficulty of a course. In order to gain more insight into the learning process and its most important characteristics, we have created features that capture possible factors that influence the grades at the end of the semester. Using these features, we present a comprehensive study to answer the following questions: which features are good indicators of a student‘s performance? which features are the most important? The findings are interesting, as different features are the most important for different classification tasks.

1.3 Observation

One of our goals is to study which factors are important indicators of a student‘s performance, We categorize each extracted feature to one of the the 8 groups, according to Table 1. Afterwards, for each classification task, we built RF classifiers using only the features belonging to one of the above groups. We selected to use RF over GB, as they achieve similar performance in less training time. The accuracy achieved for a model using a single group of features is expected to be less than the accuracy when using all the features. The percentage of accuracy that a model using only the features belonging to one group manages to achieve, in terms of the F1 score, are presented on Fig. 2. In this bar chart, we can see the percentage of accuracy achieved from all the different feature groups for all the discussed classification tasks. The higher the percentage achieved by a single group of features, the more predictive ability these features have. The goal is to get a split that allow us to make a confident prediction. Consider the m-dimensional space that is defined by the feature vectors x, of length m. There, every training instance corresponds to a single point. A Linear SVM looks for a decision boundary between two classes, a hyperplane that bisects the data with the largest possible margin between the two different classes. The margin on each side of the hyperplane is the area with no data points in it

1.3.1 Proposed System

We are going to propose the system by using which the user can give a test on specific educational or subject categories. When student complete the test, system will calculate the performance of the user by using the algorithm decision tree. The system will suggest to the teacher that on which topics the user is weak or need to study again. To solve the problems faced with manual examination writing, there is need for a computerized system to handle all the works. We propose an application that will provide a working environment that will be flexible and will provide ease of work and will reduce the time for report generation and other paper works. Today many organizations are conducting online examinations worldwide successfully and issue results online but they are not measuring the performance of the student and teacher not know about the weak points of the students and we are focusing on this issue. The main advantage is that it can be evaluation of answers can be fully automated for all questions and other essay type questions can be evaluated manually or through automated system, depending on the nature of the question‘s and the requirements. This record is also seen by teacher for analyzing the performance of student and they can take a suitable action to improve the student performance before the student in the critical area. This technique will monitor and evaluate the 3 student academic performance at different year levels before the final test in order to forecast the weaknesses of the students. Teacher can play the role of admin who have authority to adding subjects, topics and questions for test purpose.

1.3.2 Advantages of proposed system

This paper focuses on analysis student academic performance by using advantage of data mining techniques model. The main advantage is that it can be evaluation of answers can be fully automated for all questions and other essay type questions can be evaluated manually or through automated system

CHAPTER 2

Literature Survey

2.1 Literature Review

In this paper [1], an approach for classification of eye disease using the Random Forest algorithm is proposed. They have developed the model for classification which is the first automatic segmentation model for the layers and extraction of pathologies. It can be employed for early detection of common eye diseases. The performance of the system is so effective that can easily detect the symptoms. In another paper [2], a classification control is proposed. Their paper is based on rotation forest classifier with feature extraction for classification with high accuracy rate. They have acquired an average accuracy of 93.44paper [3] authors have proven that for the classification the random forests can provide the better intersection of understandability and precision. Their model can easily discriminate and helps to decrease the amount of bias available in the classifications. These results can be used by other automatic machine learning classifiers to improve results. The research work presented in the paper [4] indicates that sentiment analysis of Twitter feeds. Their method is helpful to forecast stock prices. They have used the incremental active learning approach for a stream-based setting. It is used for training data from a data stream for hand-labeling. The sentiment analysis in on the stream based active learning will support for the financial domain. The online stock data available in the form of streams support for analysis and predict the future stock prices for a particular company. Their model is more reliable as well as trustworthiness of the prediction. The paper [5] presents a new approach for unsupervised training using a decision forest. The approach is based on a key insight that is crucial to stable learning. We demonstrate that decision forests can provide this necessary conditioning for training stability. They have developed a new method for quantitatively measuring the performance

and also placing importance on generalization in learning. The author of [6] has developed a model for a suite of explanatory 'predictor' variables and the target property that can establish explicit associations between both of them. Its specific learning process influences the model's validity and portability. They have proposed a hybrid approach for leaf area index (LAI) estimation. In their model, a complimentary training dataset of independent LAI was derived. Their approach is combining data mining operations with physically based constraints via a hybrid training approach. The research work [7] improves the performance of Google Earth Engine. The author has presented a paradigm-shift in deriving high-resolution. They have used random forest machine learning algorithms on a large volume of datasets from multiple sources. In the research work [8–12] author has used random forests in conjunction with aggregate crop statistics for crop type mapping. They have simulated and validated the methodology and extracted the features using machine learning models.

CHAPTER 3

Technology study

3.1 Python

3.1.1 What is Python ?

Below are some facts about Python. Python is currently the most widely used multi-purpose, high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber... etc. The biggest strength of Python is huge collection of standard library which can be used for the following – Machine Learning GUI Applications (like Kivy, Tkinter, PyQt etc.) Web frameworks like Django (used by YouTube, Instagram, Dropbox) Image processing (like OpenCV, Pillow) Web scraping (like Scrapy, BeautifulSoup, Selenium) Test frameworks Multimedia Advantages of Python :- Let‘s see how Python dominates over other languages.

1. **Extensive Libraries** Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don‘t have to write the complete code for that manually.

2. **Extensible** As we have seen earlier, Python can be extended to other languages. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. **Embeddable** Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add scripting capabilities to our code in the

other language.

4. **Improved Productivity** The language‘s simplicity and extensive libraries render programmers more productive than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

5. **IOT Opportunities** Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

6. **Simple and Easy** When working with Java, you may have to create a class to print Hello World. But in Python, just a print statement will do. It is also quite easy to learn, understand, and code. This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

7. **Readable** Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and indentation is mandatory. This further aids the readability of the code.

8. **Object-Oriented** This language supports both the procedural and object-oriented programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the encapsulation of data and functions into one.

9. **Free and Open-Source** Like we said earlier, Python is freely available. But not only can you download Python for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

10. **Portable** When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn‘t the same with Python. Here, you need to code. only once, and you can run it anywhere. This is called Write Once Run Anywhere (WORA). However, you need to be careful enough not to include any system-dependent features.

11. **Interpreted** Lastly, we will say that it is an interpreted language. Since statements are executed one by one, debugging is easier than in compiled languages. Any doubts till now in the advantages of Python? Mention in the

comment section.

Advantages of Python Over Other Languages

1. **Less Coding** Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don‘t have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.

2. **Affordable** Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support. The 2019 Github annual survey showed us that Python has overtaken Java in the most popular programming language category.

3. **Python is for Everyone** Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and machine learning, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

Disadvantages of Python So far, we‘ve seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let‘s now see the downsides of choosing Python over another language

1. **Speed Limitations** We have seen that Python code is executed line by line. But since Python is interpreted, it often results in slow execution. This, however, isn‘t a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

2. **Weak in Mobile Computing and Browsers** While it serves as an excellent server-side language, Python is much rarely seen on the client-side. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called Carbonnelle. The reason it is not so famous despite the existence of Brython is that it isn‘t that secure.

3. **Design Restrictions** As you know, Python is dynamically-typed. This

means that you don‘t need to declare the type of variable while writing the code. It uses duck-typing. But wait, what‘s that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can raise run-time errors.

4. **Underdeveloped Database Access Layers** Compared to more widely used technologies like JDBC (Java DataBase Connectivity) and ODBC (Open DataBase Connectivity), Python‘s database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

5. **Simple** No, we‘re not kidding. Python‘s simplicity can indeed be a problem. Take my example. I don‘t do Java, I‘m more of a Python person. To me, its syntax is so simple that the verbosity of Java code seems unnecessary. This was all about the Advantages and Disadvantages of Python Programming Language.

3.1.2 History of Python :-

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde 8 Informatica). The greatest achievement of ABC was to influence the design of Python. Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners1 , Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde en Informatica (CWI). I don't know how well people know ABC's influence on Python. I try to mention ABC's influence because I'm indebted to everything I learned during that project and to the people who worked on it." Later on in the same Interview, Guido van Rossum continued: "I remembered all my experience and some of my frustration with ABC. I decided to try to design a simple scripting language that possessed some of ABC's better properties,

but without its problems. So I started typing. I created a simple virtual machine, a simple parser, and a simple runtime. I made my own version of the various ABC parts that I liked. I created a basic syntax, used indentation for statement grouping instead of curly braces or begin-end blocks, and developed a small number of powerful data types: a hash table (or dictionary, as we call it), a list, strings, and numbers."

3.2 Machine Learning

3.2.1 what is Machine Learning ?

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data. Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain.Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

3.2.2 Categories Of Machine Learning :-

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.9 Supervised

learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section. Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

3.2.3 Need for Machine Learning :-

Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and haven‘t surpassed human intelligence in many aspects. Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, —to make decisions, based on data, with efficiency and scale‖. Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programing logic, in the problems that cannot be programmed inherently. The fact is that we can‘t do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

3.2.4 Challenges in Machines Learning :-

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are Quality of data Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.10 Time-Consuming task Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval. Lack of specialist persons As ML technology is still in its infancy stage, availability of expert resources is a tough job. No clear objective for formulating business problems Having no clear objective and welldefined goal for business problems is another key challenge for ML because this technology is not that mature yet. Issue of overfitting underfitting If the model is overfitting or underfitting, it cannot be represented well for the problem. Curse of dimensionality Another challenge ML model faces is too many features of data points. This can be a real hindrance.

3.2.5 Applications of Machines Learning

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML

1. Emotion analysis
2. Sentiment analysis
3. Error detection and prevention
4. Weather forecasting and prediction
5. Stock market analysis and forecasting
6. Speech synthesis
7. Speech recognition
8. Customer segmentation

9. Object recognition
10. Fraud detection
11. Fraud prevention

3.2.6 Types of Machine Learning :-

Supervised Learning – This involves learning from a training dataset with labeled data using classification and regression models. This learning process continues until the required level of performance is achieved.

Unsupervised Learning – This involves using unlabelled data and then finding the underlying structure in the data in order to learn more and more about the data itself using factor and cluster analysis models.

Semi-supervised Learning – This involves using unlabelled data like Unsupervised Learning with a small amount of labeled data. Using labeled data vastly increases the learning accuracy and is also more cost-effective than Supervised Learning.

Reinforcement Learning – This involves learning optimal actions through trial and error. So the next action is decided by learning behaviors that are based on the current state and that will maximize the reward in the future.

3.2.7 Advantages of Machine learning :-

1. **Easily identifies trends and patterns** - Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

2. **No human intervention needed (automation)** With ML, you don‘t need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. 14 ML is also good at recognizing spam.

3. **Continuous Improvement** As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. **Handling multi-dimensional and multi-variety data** Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

5. **Wide Applications** You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

3.2.8 Disadvantages of Machine Learning :-

1. **Data Acquisition** Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2. **Time and Resources** ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

3. **Interpretation of Results** Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4. **High error-susceptibility** Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

3.3 Modules Used in Project:-

3.3.1 Tensorflow :-

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015

3.3.2 Numpy :-

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

A powerful N-dimensional array object

Sophisticated (broadcasting) functions

Tools for integrating C/C++ and Fortran code

Useful linear algebra, Fourier transform, and random number capabilities17

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

3.3.3 Pandas :-

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with

Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

3.3.4 Scikit – learn :-

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. Python Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. 18 Python is Interpreted Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP. Python is Interactive you can actually sit at a Python prompt and interact with the interpreter directly to write your programs. Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that

avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

CHAPTER 4

System Design

4.1 System Architecture

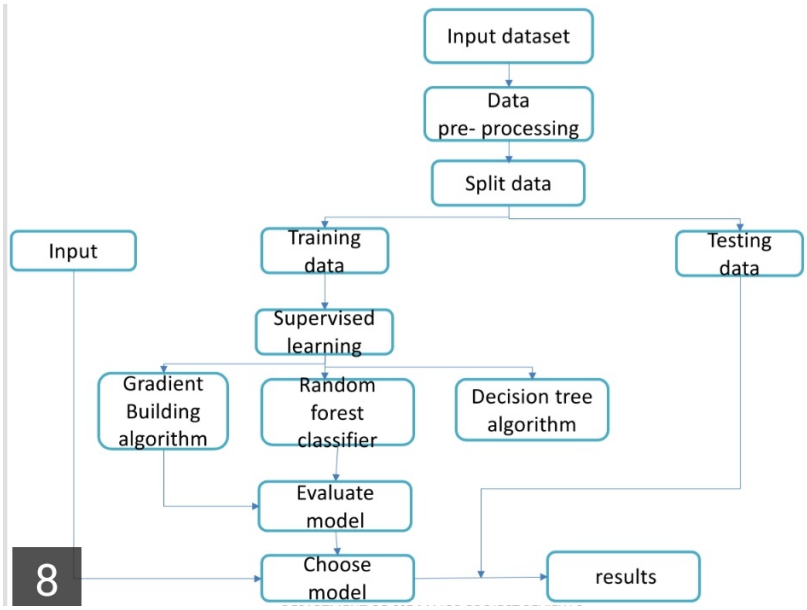


Figure 4.1: Architectural Diagram[1]

4.2 Module Description

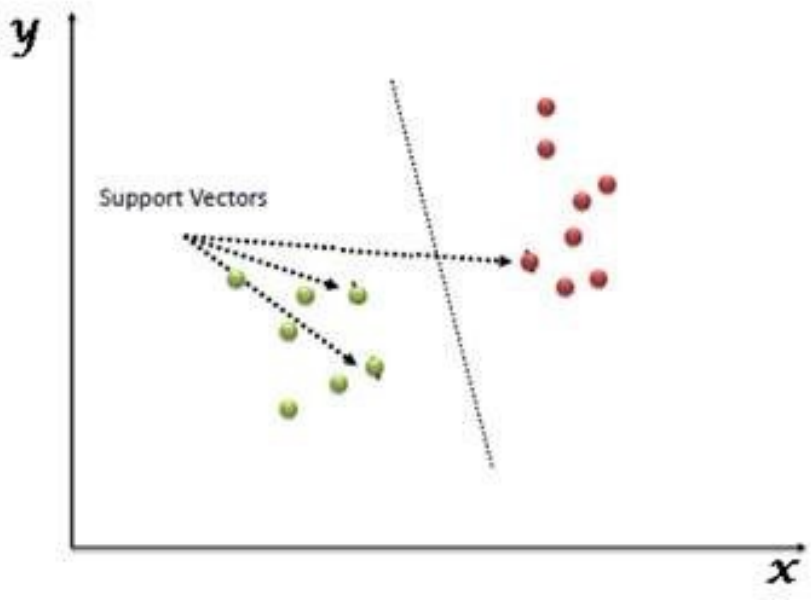
Feature selection is an important part in machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model.

In wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince classifiers from the features in each subset. The classifier consider the subset of feature with which the classification algorithm performs the best. To find the subset, the evaluator uses different search techniques like depth first search, random search, breadth first search or hybrid search. The filter 27 method uses an attribute evaluator along with a ranker to rank all the features in the dataset. Here one feature is omitted at a time that has lower ranks and then sees the predictive accuracy of the classification algorithm. Weights or rank put by the ranker algorithms are different than those by the classification algorithm. Wrapper method is useful for machine learning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features.

4.3 Algorithms used in this project

4.3.1 Support Vector Machine

—Support Vector Machine‖ (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot)

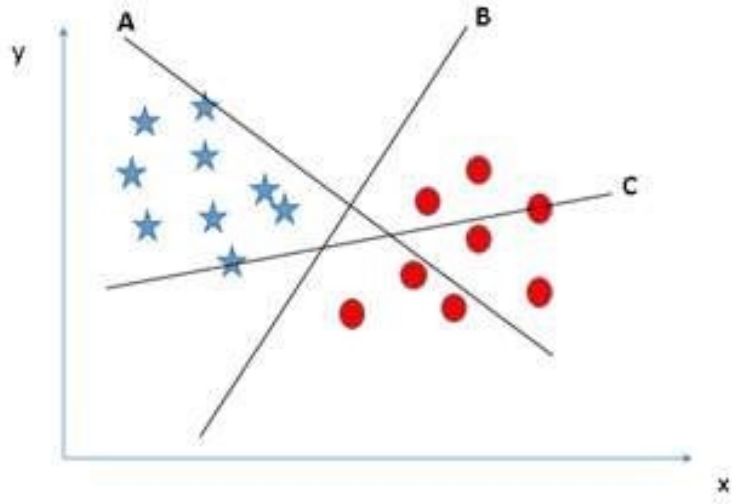


Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).You can look at support vector machines and a few examples of its working here.

4.3.2 How does it work?

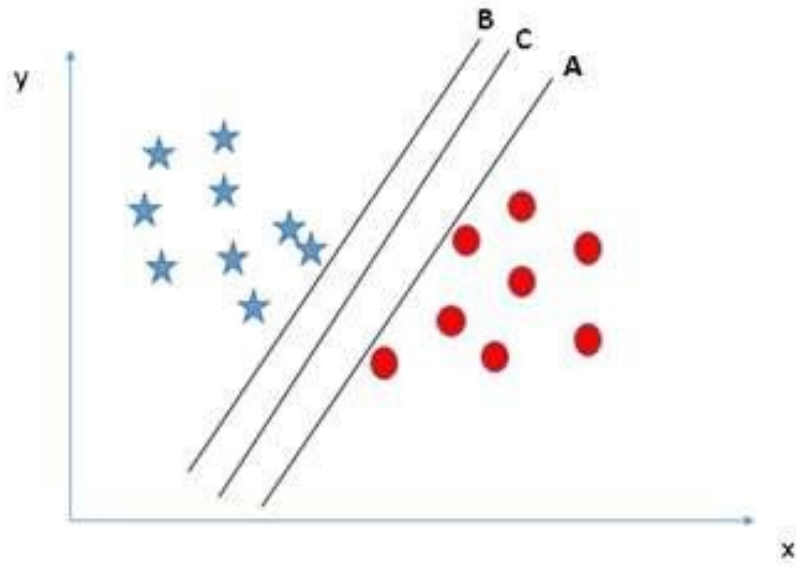
Above, we got accustomed to the process of segregating the two classes with a hyper-plane. Now the burning question is —How can we identify the right hyper-plane?‖. Don‘t worry, it‘s not as hard as you think!28 Let‘s understand: Identify the right hyper-plane (Scenario-1): Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.

$\boldsymbol{Identifytherighthyper-plane(Scenario-1)}$: Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.



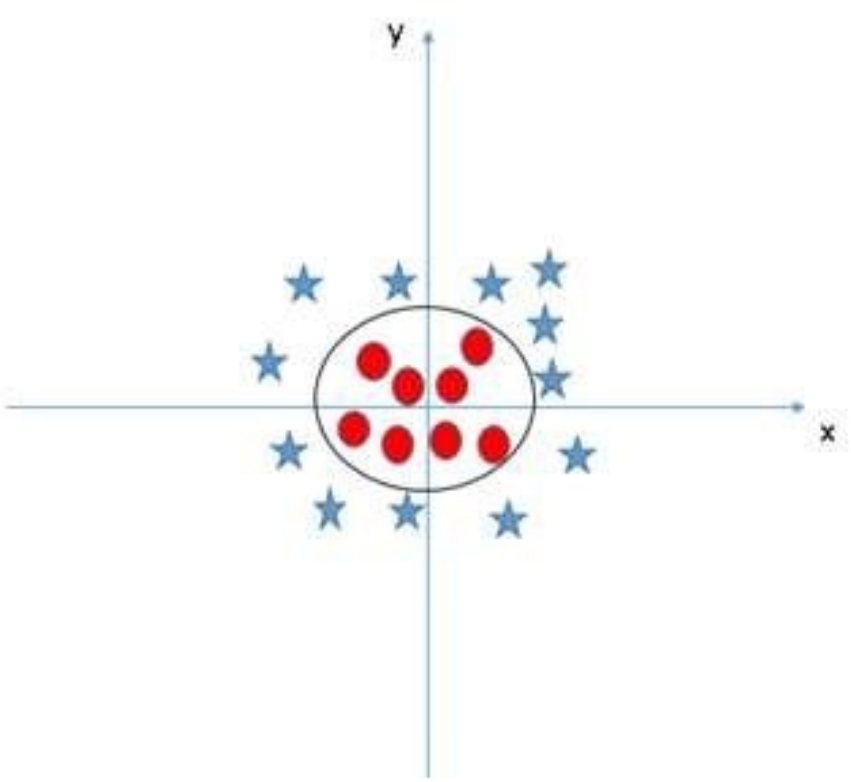
You need to remember a thumb rule to identify the right hyper-plane: —Select the hyper-plane which segregates the two classes better‖. In this scenario, hyper-plane —B‖ has excellently performed this job.

$\boldsymbol{Identifytherighthyper-plane(Scenario-2)}$: Here, we have three hyper-planes (A, B and C) and all are segregating the classes well. Now, How can we identify the right hyper-plane?



In the SVM classifier, it is easy to have a linear hyper-plane between these

two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyper-plane. No, the SVM algorithm has a technique called the kernel trick. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you‘ve defined.



Now, let‘s look at the methods to apply SVM classifier algorithm in a data science challenge.

4.3.3 Decision Tree Algorithm:

This algorithm will build training model by arranging all similar records in the same branch of tree and continue till all records arrange in entire tree. The complete tree will be referred as classification train model.

4.4 System Specification

4.4.1 Software Requirements

Functional requirements for a secure cloud storage service are straightforward:

1. The service should be able to store the user‘s data;

2. The data should be accessible through any devices connected to the Internet;

3. The service should be capable to synchronize the user‘s data between multiple devices (notebooks, smart phones, etc.);

4. The service should preserve all historical changes (versioning);

5. Data should be shareable with other users;

6. The service should support SSO; and

7. The service should be interoperable with other cloud storage services, enabling data migration from one CSP to another.

- Operating System: Windows
- Coding Language: Python 3.7

4.4.2 Hardware Requirements:

- Processor - Pentium –III
- Speed – 2.4 GHz
- RAM - 512 MB (min)
- Hard Disk - 20 GB
- Floppy Drive - 1.44 MB
- Key Board - Standard Keyboard
- Monitor – 15 VGA Colour

4.5 Detailed Design

UML is an acronym that stands for Unified Modeling Language. Simply put, UML is a modern approach to modeling and documenting software. In fact, it's one of the most popular business process modeling techniques. It is based on diagrammatic representations of software components. As the old proverb says: —a picture is worth a thousand words‖. By using visual representations, we are able to better understand possible flaws or errors in software or business processes. UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems. The need arose for a more unified way to visually represent those systems and as a result, in 1994-1996, the UML was developed by three software engineers working at Rational Software. It was later adopted as the standard in 1997 and has remained the standard ever since, receiving only a few updates.

***GOALS* :**

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.

2. Provide extendibility and specialization mechanisms to extend the core concepts.

3. Be independent of particular programming languages and development process.

4. Provide a formal basis for understanding the modeling language.

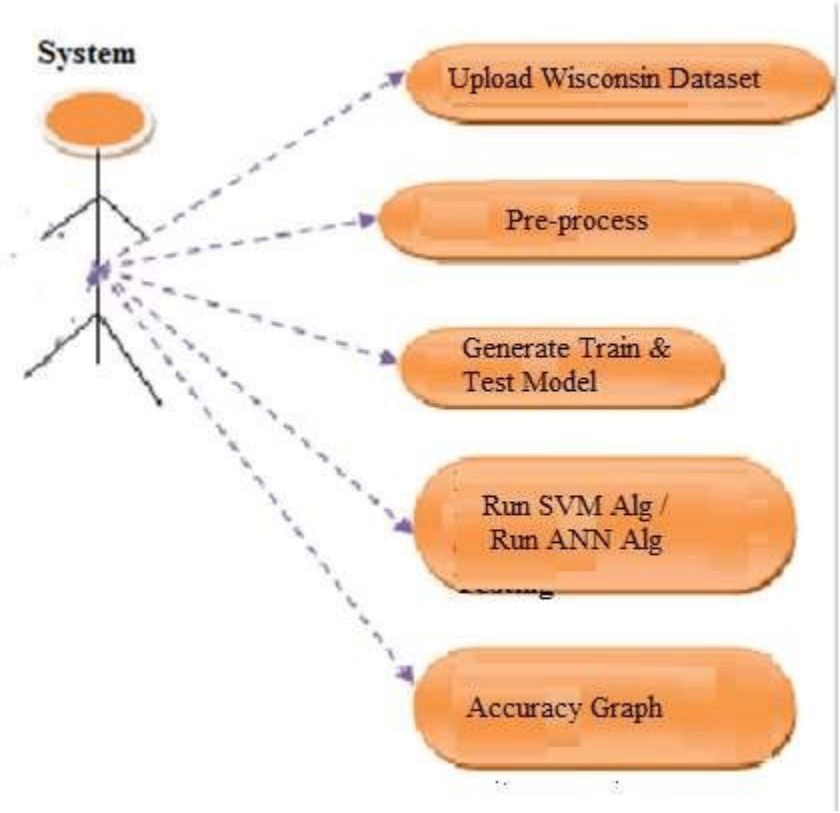
5. Encourage the growth of OO tools market.

6 Support higher level development concepts such as collaborations, frameworks, patterns and components.

4.5.1 Usecase Diagram :

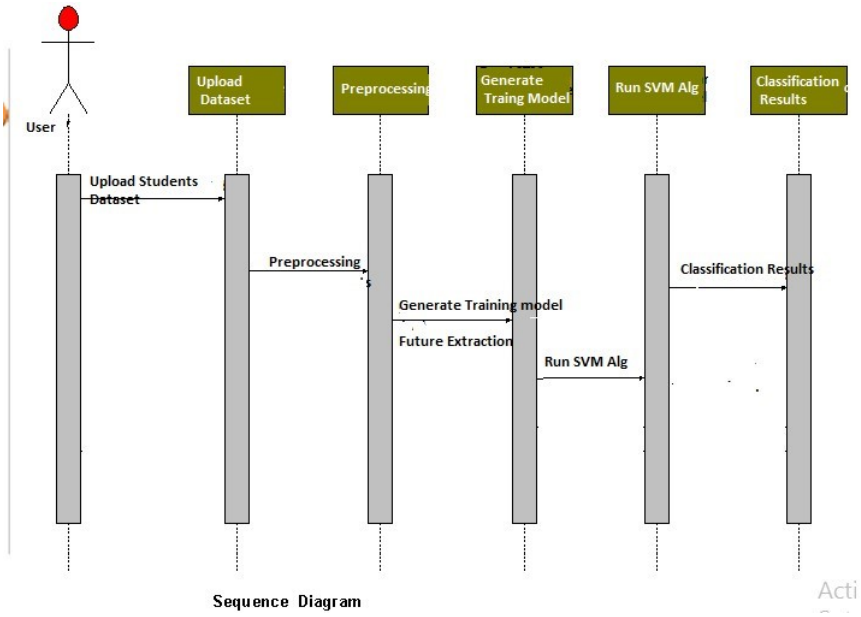
A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its

purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



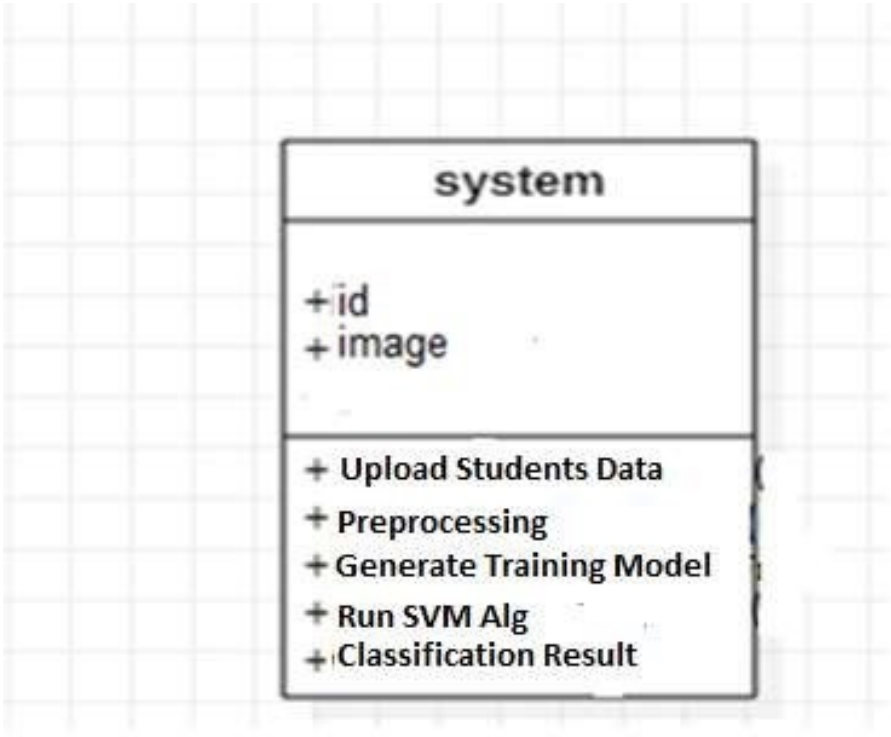
4.5.2 Sequence Diagram:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



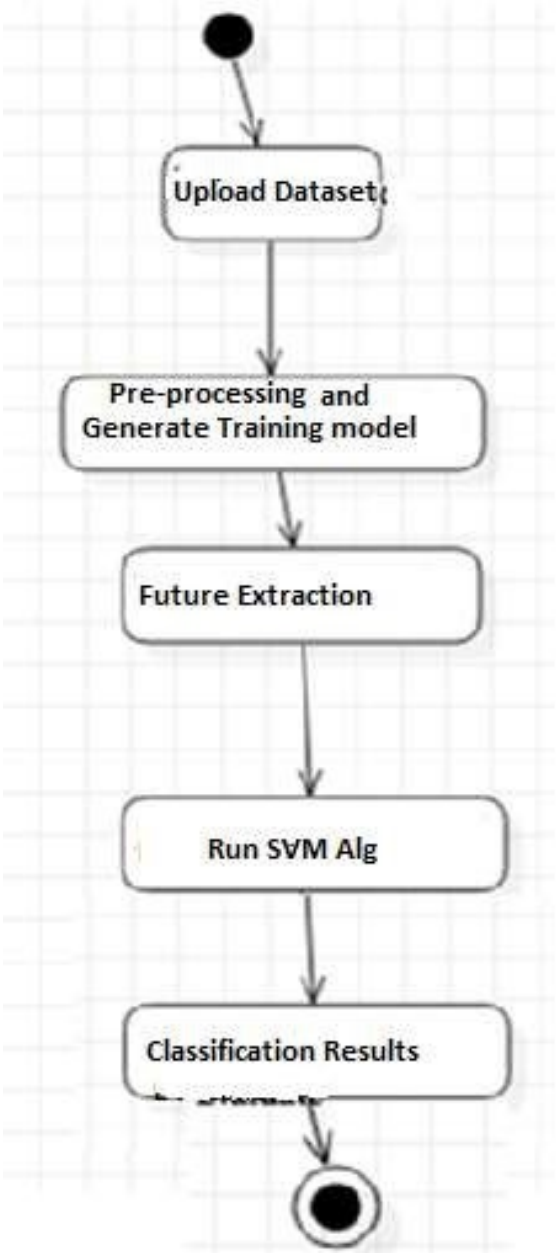
4.5.3 Class Diagram:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



4.5.4 Activity Diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



CHAPTER 5

Title of the Chapter

5.1 A Section

Lorem ipsum[2] dolor sit amet, consectetur adipiscing elit. Etiam consectetur libero dui, sit amet rutrum lectus mollis ac. Phasellus mattis augue quis auctor ullamcorper. Sed congue rutrum turpis, sit amet tincidunt erat laoreet ac. Morbi in feugiat erat, sit amet placerat lorem. Quisque cursus gravida nulla, nec pulvinar justo rutrum eget. Aenean et dolor vitae enim congue maximus. Praesent consequat finibus imperdiet. Sed ipsum erat, efficitur vitae urna a, convallis ornare ipsum. Quisque fringilla risus enim, ut elementum dui consectetur eu.

5.1.1 Part of a section

Vivamus sed enim quam. In facilisis consequat eros, id convallis diam congue et. Vivamus sed neque rutrum, gravida urna non, lobortis leo. Curabitur sagittis turpis sit amet dolor blandit lobortis. Fusce ac sem eget libero semper ultrices. Aliquam erat volutpat. Quisque tincidunt mi sapien, eu varius libero lacinia quis. Nullam dictum quam sed scelerisque aliquet. Mauris laoreet est nec dolor pellentesque dignissim. Praesent ac sapien erat. Mauris ut felis sit amet velit convallis cursus quis a tellus. Curabitur efficitur ultricies dui, eget vestibulum arcu venenatis ac. Nulla fermentum dolor a venenatis posuere. Vivamus eu luctus erat.

Here is a table attached :

n	F_n
1	1
2	1
3	2
4	3

Table 5.1: Fibonacci Table

Morbi pulvinar turpis at ligula sollicitudin, et finibus nulla vestibulum. Etiam molestie tincidunt molestie. Mauris augue ex, tincidunt non ante eget, ultrices ornare quam. Nulla pellentesque fringilla neque, eget facilisis arcu mollis vel. Nulla quam ipsum, vulputate a nisi ut, iaculis tempus ante. Nam efficitur, augue et cursus varius, nisi eros tincidunt ex, porta aliquet massa libero a mauris. Vestibulum finibus, dui sit amet dapibus tincidunt, libero quam venenatis elit, ut porta dui neque a nunc. Donec tincidunt eleifend mauris a luctus. Vivamus eget porttitor metus. Etiam eu porta orci. Suspendisse imperdiet orci sed lobortis vestibulum. Pellentesque sit amet euismod eros.

For an ideal gas,

$$P \cdot V = n \cdot R \cdot T \tag{5.1}$$

5.2 Another Section

5.2.1 Another subsection

Curabitur malesuada purus ac orci elementum, ac mollis lectus ornare. Nunc eget nunc non nunc viverra porttitor vel eget nibh. Phasellus nec finibus neque, vitae volutpat purus. Donec vitae sapien dictum, elementum ex eget, iaculis dui. Nulla elementum viverra purus.

Here is an image of Berlin Cathedral.

Random papers are referenced here[3] and here[4]. Go check the References page.

CHAPTER 6

Test Results

6.1 Types of tests

6.1.1 Unit testing

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.

6.1.2 System Test

System testing ensures 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 links and integration points.

6.1.3 White Box Testing

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

6.1.4 Black Box Testing

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, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot —see‖ into it. The test provides inputs and responds to outputs without considering how the software works.

6.1.5 Integration testing

Integration tests are designed to test integrated software 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.

6.1.6 Functional test

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 centered on the following items:54

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing

is complete, additional tests are identified and the effective value of current tests is determined.

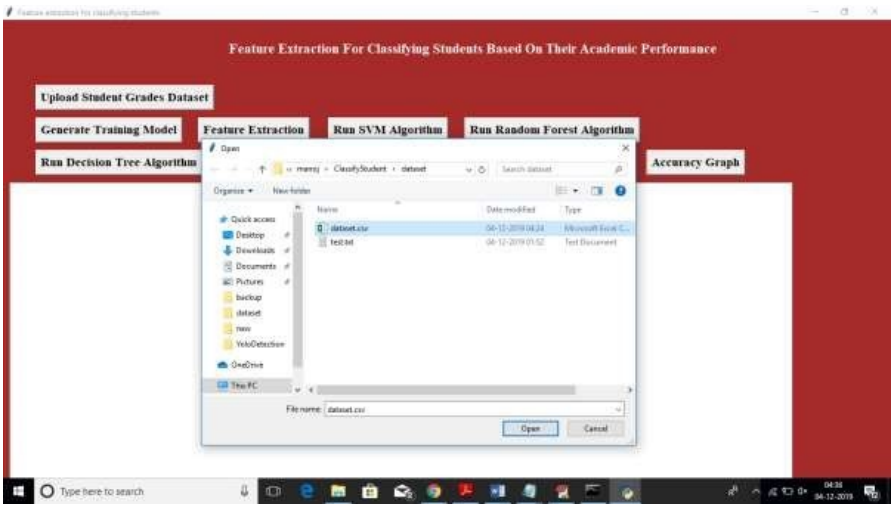
CHAPTER 7

Results and outputs

7.1 Outputs

To run this project double click on run file to get below screen

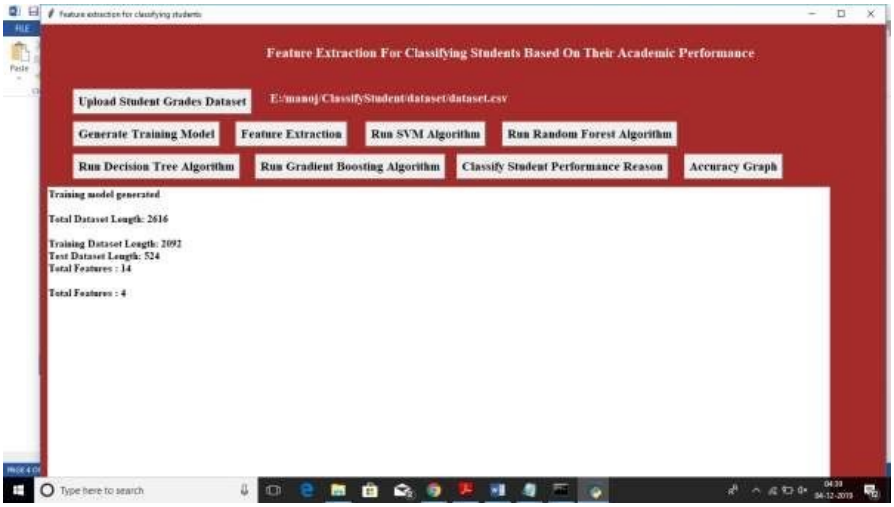
In above screen click on "Upload Student Grades Dataset" button to upload dataset



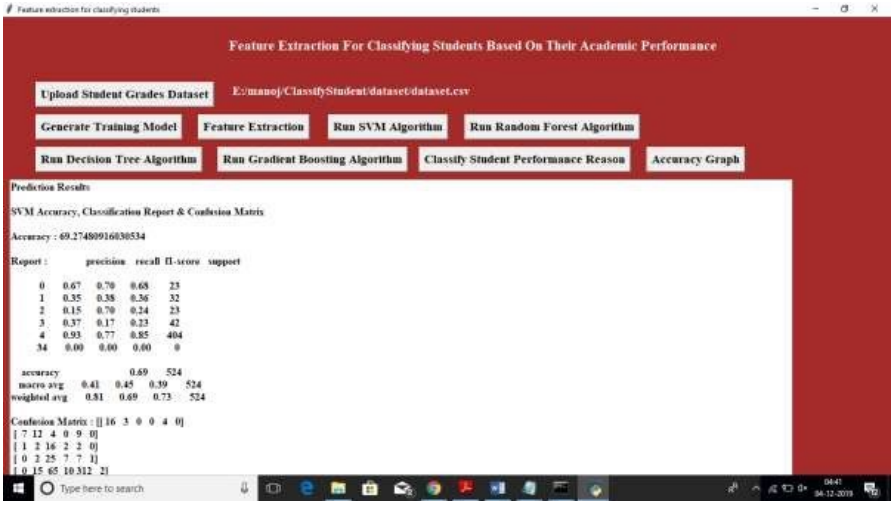
After uploading dataset will get below screen

Now click on Generate Training Model button to read dataset and build array for training purpose

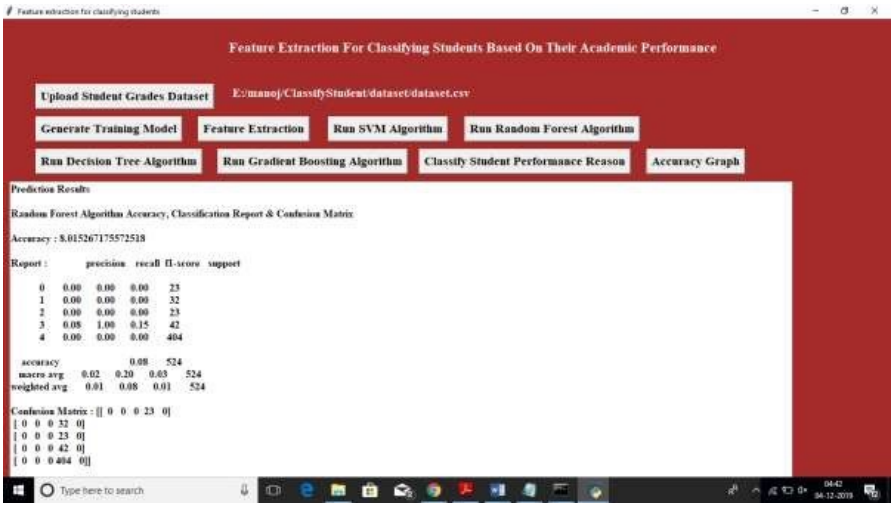
In above screen we can see total number of records in dataset and then displaying algorithm chooses how many records for training and testing purpose. Now click on Features Extraction button to extract features and assign as class label to the classifier algorithms.



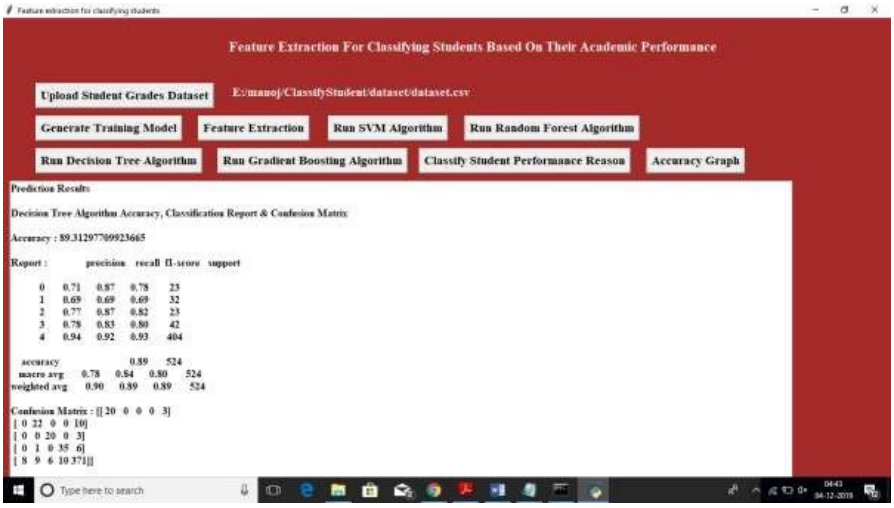
In above screen we can see dataset contains total 14 numeric features and extracted features are 4. After feature extraction click on Run SVM Algorithm to build SVM train model and to get classifier accuracy and FSCORE value



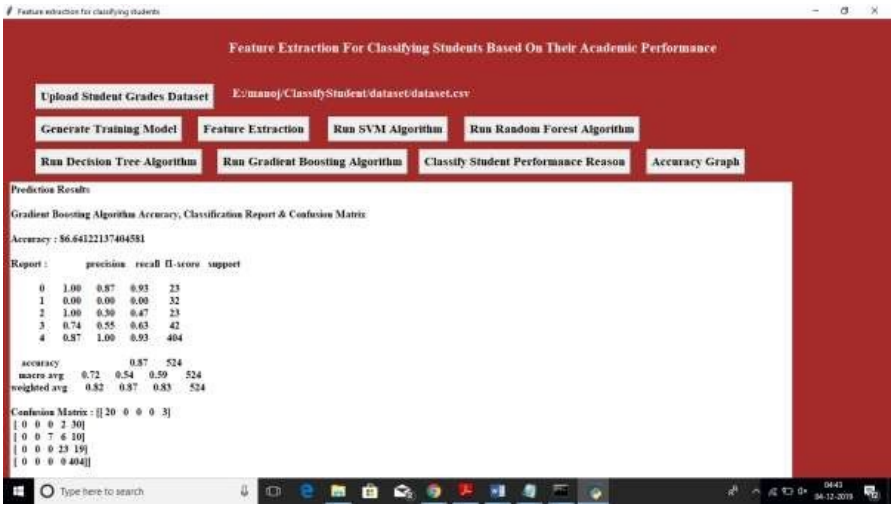
In above screen we can see SVM accuracy is 69on Run Random Forest Algorithm to build its model



In above screen random forest got only 8

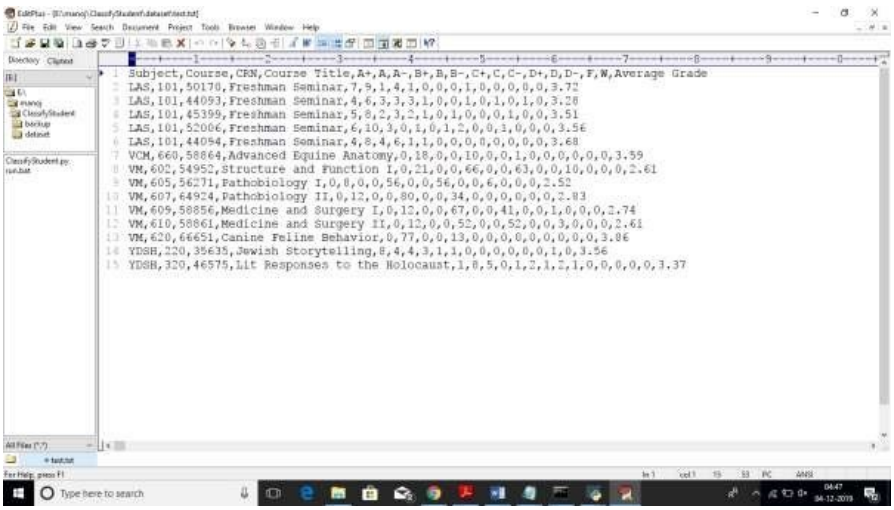


In above screen decision tree got 89

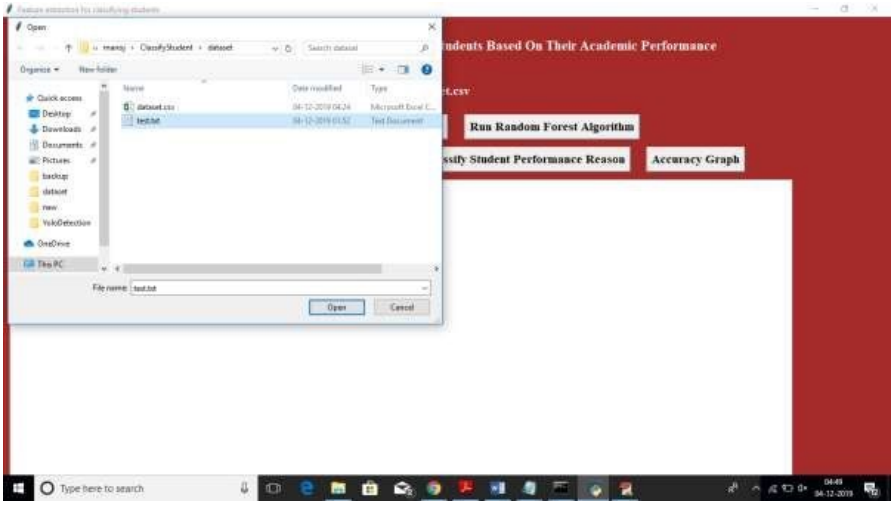


In above screen gradient boosting got 87

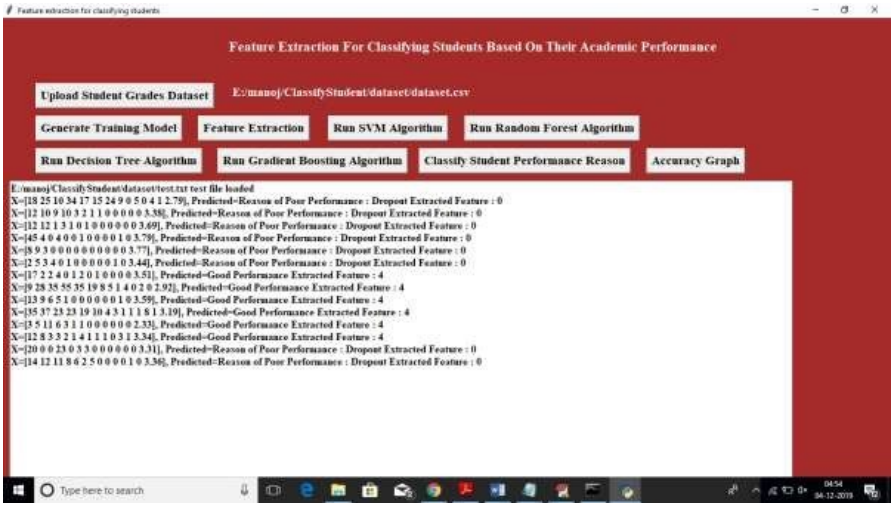
In above screen x-axis represents algorithm name and y-axis represents accuracy of that algorithm. Now we can test new student records on this train model to predict or classify new student performance. To check new student we need to upload text.txt test dataset from dataset folder and this dataset contains below data.



In above test data we don‘t have extracted feature values such as 0,1,2,3 or 4 and this value will be predicted or classify by this machine learning algorithms. Just we need to click on Classify Student Performance Reason button and upload test dataset then will get below result



After uploading test data will get below classification result



In above screen based on grades values application has given result as poor performance due to drop out or good performance

CHAPTER 8

Conclusions and Future Scope

8.1 Conclusion

The purpose of this paper is to accurate identify students that are at risk. These students might fail the class, drop it, or perform worst than they usually do. We extracted features from historical grading data, in order to test different simple and sophisticated classification methods based on big data approaches. The best performing methods are the Gradient Boosting and Random Forest classifiers, based on AUC and F1 score metrics. We also got interesting findings that can explain the student performance.

8.2 Future Scope

One of our goals is to study which factors are important indicators of a student‘s performance, so we performed the we can get many insights on the factors that affect student performance. For example, the features related to the students‘ grades (group 1) have a very good predictive capability in almost all the tasks, except the task of predicting the W grades. In this task, features related with the course‘s difficulty and popularity (group 4) as well as features that are course-specific (group 8), manage to achieve the same accuracy as when using all the features. This indicates that the reasons that a student drops a course are related more to the course, rather than to the students themselves. The next best indicator is the feature group about the student‘s course load during the semester The feature groups are behaving similarly for RelF and RelCF. However, we notice that for the RelCF task, the feature groups that are related with student-course specific features have slightly better performance, while the student-specific groups have slightly worst performance, compared to the task of RelF. This is happening because,

for RelCF, we take into consideration how other students usually perform on the target course. Every single group has enough information for the RF to utilize to achieve performance which is as good as 75feature

CHAPTER 9

References

[1] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001. [2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984. [3] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):273–297, 1995. [4] J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001. [5] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran. Machine learning based student grade prediction: A case study. arXiv preprint arXiv:1708.08744, 2017. [6] J. E. Knowles. Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. Journal of Educational Data Mining, 7(3):18–67, 2015.