A Project Report on

ANALYSIS AND PREDICTION OF STUDENTS PERFORMANCE USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of the Degree of

Bachelor of Technology
in
Computer Science and Engineering

By

VELAGALA SOWMYA REDDY	**19A91A0564**
BARRI SANDHYA	**19A91A0503**
K.YASHWANTH REDDY	**19A91A0520**

Under the Esteemed Supervision of

Dr.P.UDAYAKUMAR.,M.Tech,Ph.D

Professor(CSE)

Department of Computer Science and Engineering

ADITYA ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Affiliated to JNTUK Kakinada, Accredited by NAAC with 'A' Grade)

Aditya Nagar, ADB Road, Surampalem

2019 – 2023

ADITYA ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Affiliated to JNTUK Kakinada, Accredited by NAAC with 'A' Grade)

Aditya Nagar, ADB Road, Surampalem

2019 – 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the thesis entitled "**ANALYSIS AND PREDICTION OF STUDENTS PERFORMANCE USING MACHINE LEARNING"** is being submitted by

VELAGALA SOWMYA REDDY	**19A91A0564**
BARRI SANDHY	**19A91A0503**
K.YASHWANTH REDDY	**19A91A0520**

in partial fulfillment of the requirements for the award of degree of B**. Tech** in Computer Science and Engineering from **Jawaharlal Nehru Technological University Kakinada** is a record of bonafide work carried out by them at **Aditya Engineering College**

The results embodied in this Project report have not been submitted to any other University or Institute for the award of any degree or diploma.

PROJECT GUIDE	**HEAD OF THE DEPARTMENT**
Dr.P.Udayakumar.,M.Tech,Ph.D	**Dr.A.Vanathi.,M.E,Ph.D**

EXTERNAL EXAMINER

DECLARATION

We hereby declare that the project entitled **"ANALYSIS AND PREDICTION OF STUDENTS PERFORMANCE USING MACHINE LEARNING"** is a genuine project. This work has been submitted to the **ADITYA ENGINEERING COLLEGE,** Surampalem, permanently affiliated to **JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY, KAKINADA** in partial fulfillment of the **B.Tech** degree**.** We further declare that this project work has not been submitted in full or part of the award of any degree of this or any other educational institutions.

by

VELAGALA SOWMYA REDDY (19A91A0564)

BARRI SANDHYA (19A91A0503)

K.YASHWANTH REDDY (19A91A0520)

ACKNOWLEDGEMENT

It is with immense pleasure that we would like to express our indebted gratitude to our project guide **Dr. P. Udayakumar, Professor** who has guided us a lot and encouraged us in every step of the project work, **his** valuable moral support and guidance throughout the project helped us to a greater extent.

Our deepest thanks to our HOD **Dr. A. Vanathi, Associate Professor** for inspiring us all the way and for arranging all the facilities and resources needed for our project.

Our sincere thanks to project coordinator **Mrs. T. Sudha Rani, Associate Professor** for providing a great support and suggestions during our project work.

We wish to thank **Dr. S. Rama Sree, Professor** in CSE and Dean(Academics) for her support and suggestions during our project work.

We owe our sincere gratitude to **Dr. M.Sreenivasa Reddy, Principal** for providing a great support and for giving us the opportunity of doing the project.

We are thankful to our **College Management** for providing all the facilities in time to us for completion of our project.

Not to forget, **Faculty, Lab Technicians, non-teachin g staff and our friends** who have directly or indirectly helped and supported us in completing our project in time.

ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

VISION & MISSION OF THE INSTITUTE

Vision:

To emerge as a premier institute for quality technical education and innovation.

Mission:

M1: Provide learner centric technical education towards academic excellence

M2: Train on technology through collaborations

M3: Promote innovative research & development

M4: Involve industry institute interaction for societal needs

PRINCIPAL

PRINCIPAL
ADITYA ENGINEERING COLLEGE
SURAMPALEM - 533 437

VISION & MISSION OF THE DEPARTMENT

VISION:

To emerge as a competent Centre of excellence in the field of Computer Science and Engineering for industry and societal needs.

MISSION:

- Impart quality and value based education.
- Inculcate the inter personal skills and professional ethics.
- Enable research through state-of-the-art infrastructure.
- Collaborate with industries, government and professional societies.

Head of the Department
Head of the Department
Department of CSE
ADITYA ENGINEERING COLLEGE (A)

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

Graduates of the Program will

- **PEO 1:** Adopt new technologies and provide innovative solutions.
- **PEO 2:** Be employable, become an entrepreneur or researcher for a successful career.
- **PEO 3:** Demonstrate interpersonal, multi-disciplinary skills and professional ethics to serve society.

Head of the Department

Head of the Department
Department of CSE
ADITYA ENGINEERING COLLEGE (A)

PROGRAM OUTCOMES (POs)

After successful completion of the program, the graduates will be able to

PO 1 **Engineering Knowledge:** Apply knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.

PO 2 **Problem Analysis:** Identify, formulate, research literature and analyze complex engineering problems, reaching substantiated conclusions using first principles of mathematics, natural sciences and engineering sciences.

PO 3 **Design/Development of Solutions:** Design solutions for complex engineering problems and design systems, components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.

PO 4 **Conduct Investigations of Complex Problems:** Conduct investigations of complex problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of information to provide valideconclusions.

PO 5 **Modern Tool Usage:** Create, select and apply appropriate techniques, resources, and modern engineering and IT tools, including prediction and modelling, to complex engineering activities, with an understanding of the limitations.

PO 6 **The Engineer and Society:** Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.

PO 7 **Environment and Sustainability:** Understand the impact of professional engineering solutions in societal and environmental contexts and demonstrate knowledge of, and need for sustainable development.

PO 8 **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of engineering practice.

PO 9 **Individual and Teamwork:** Function effectively as an individual, and as a member or leader in diverse teams and in multidisciplinary settings.

PO 10 **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO 11 **Project Management and Finance:** Demonstrate knowledge and understanding of engineering management principles and apply these to one's own work, as a member and leader in a team and to manage projects in multidisciplinary environments.

PO 12 **Life-Long Learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Head of the Department
Head of the Department
Department of CSE
ADITYA ENGINEERING COLLEGE (A

PROGRAM SPECIFIC OUTCOMES (PSOs)

After successful completion of the program, the graduates will be able to

PSO 1: Develop efficient solutions to real world problems using the domains of Algorithms, Networks, database management and latest programming tools and techniques.

PSO 2: Provide data centric business solutions through emerging areas like IoT, AI , data analytics and Block Chain technologies.

Head of the Department
Head of the Department
Department of CSE
ADITYA ENGINEERING COLLEGE (A)

ABSTRACT

The main objective of any educational institution is providing the best education and knowledge to the students. However, student to student academic performance may vary. A single teacher can't monitor each and every student at once, due to this every year many students lag behind due to lack of proper advice and monitoring. If universities can analyze the low performance factors earlier and can predict the behavior of students, this information can help them take effective steps to increase student performance. So, Student performance prediction is very important to understand a student progress rate. So that it helps the teacher to know about the student like which student needs which kind of help. Not only students and teachers, parents also can get the knowledge on student performance in academics. Due to huge amount of data in educational databases, predicting the performance of students has became more difficult. A data mining research and prediction can be used in the education system for analyzing and predicting the student performance. Using educational machine learning algorithms- Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine, we can predict the student performance in academics. The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject and student background information we will predict the marks and depending on the predicted marks the student can know in which subject he/she is lagging and teachers can easily monitor the student performance and can give the student a proper advice to improve the student performance.

Keywords: Student Performance Prediction, Machine Learning, Decision Tree, Random Forest, K-Nearest Neighbour, Support Vector Machine.

S.No	INDEX	Page.No
1.	**INTRODUCTION**	01
	1.1 Introduction of Project area/Domain	01
	1.2 Existing System and Disadvantages	02
	1.3 Proposed System and Advantages	03
	1.4 Objectives of the Project	05
	1.5 Flowchart of the Project	05
2.	**REQUIREMENT ANALYSIS**	07
	2.1 Introduction of Requirement Analysis	07
	2.2 Hardware & Software Requirements	07
	2.3 Software Requirements Specification	14
3.	**LITERATURE SURVEY**	16
4.	**MODULES**	23
	4.1 Introduction to Modules	23
	4.2 Module Implementation	23
5.	**SYSTEM DESIGN**	25
	5.1 Introduction to System Design	25
	5.2 Block Diagram of Proposed System	25
	5.3 Input Design and Output Design	26
	5.4 UML Diagrams	27
6.	**SYTEM IMPLEMENTATION**	35
	6.1 Introduction	35
	6.2 Selected Software	40
	6.3 Sample Code	42
7.	**TESTING**	49
	7.1 Introduction	49
	7.2 Test cases	52
8.	**SCREENS & REPORTS**	55
	8.1 Introduction	55
	8.2 Screenshots	55
	8.3 Reports	59

9. CONCLUSION AND FUTURE SCOPE 62

9.1 Conclusion 62

9.2 Future Scope 62

10. BIBLIOGRAPHY 63

10.1 References 63

RESEARCH PAPER PUBLISHED

LIST OF FIGURES

Figure	Name of the Figure	Page. No
1.1	Flow Chart of the Project	06
2.1	Students Dataset	12
2.2	Students Dataset	13
2.3	Students Dataset	13
2.4	Characteristics of SRS	15
5.1	Block Diagram	25
5.2	Class Diagram	28
5.3	Use Case Diagram	30
5.4	Sequence Diagram	31
5.5	Collaboration Diagram	32
5.6	Activity Diagram	34
6.1	Decision Tree	36
6.2	Random Forest	37
6.3	Support Vector Machine	38
6.4	K-Nearest Neighbor	39
6.5	Google Colab	40
6.6	Google Colab Notebook	41
6.7	Google Colab Code Editor	41
7.1	Result of Test Case 1	52
7.2	Result of Test Case 2	53
7.3	Result of Test Case 3	54
7.4	Result of Test Case 4	54
8.1	Visualization Screenshots	56
8.2	Decision Tree Algorithm Execution Screenshot	57
8.3	Random Forest Algorithm Execution Screenshot	58
8.4	Support Vector Machine Algorithm Execution Screenshot	58
8.5	K-Nearest Neighbor Algorithm Execution Screenshot	59
8.6	Comparison of Algorithms	60
8.7	Output Report	61

LIST OF TABLES

Table	Name of the Table	Page. No
3.1	Comparison of Literature Review	20
7.1	Test case 1	52
7.2	Test case 2	53
7.3	Test case 3	53
7.4	Test case 4	54

1. INTRODUCTION

1.1 Introduction of Project Domain:

Machine Learning is a subset of artificial intelligence that is used in designing and developing algorithms through which computers can predict the behavior based on the given data. Machine learning algorithms mainly focus on pattern recognition and decision making. Machine learning is wherever data is being used and analyzed. It is a very useful tool for exposing secret patterns and important information that can otherwise be sought and comprehended with statistical methods. Nowadays in the area of education, there is a significant growth in using machine learning algorithms. Student performance prediction is very important to understand a student progress rate. So that it helps the teacher to know about the student like which student needs which kind of help.

In order to do this it is necessary to evaluate the output of students in any college, school or any other institution of education by performing diverse tests, evaluations and analysis. To comprehend and analyze academic performance and student behavior, both past and present strong tools are necessary to scientifically evaluate and forecast student results. Machine learning models will use the students past marks and if the model predict that students have higher probability of failing a course then alternative educational strategies would be employed to make his/her learning experience improved.

If universities can analyze the low performance factors earlier and can predict the behavior of students, this information can help them take effective steps to increase student performance. So, Student performance prediction is very important to understand a student progress rate. So that it helps the teacher to know about the student like which student needs which kind of help. Not only students and teachers, parents also can get the knowledge on student performance in academics. A data mining research and prediction can be used in the education system for analyzing and predicting the student performance. Using educational machine learning algorithms- Decision Tree, Random Forest, Support Vector Machine and KNN classifier we can predict the student performance in academics. The proposed system provides the efficient algorithm for analysis.

1.2 Existing System and Disadvantages:

The academic output of the student is generally stored in various formats in the student administration system like in files, documents, records, photographs and other formats. These data can be collected for valuable knowledge from the students. However, it is difficult to analyze the increasing amount of data of students through conventional statistical techniques and database management tools. For analysis Binomial logistic regression, Decision Tree, Support Vector Machine and KNN classifier are used So that students, teachers and parents can get the knowledge on student performance in academics. This method can help the teacher to decide about performance of the students and teachers can give advice to the student for improving their academics.

In this study, they discussed on implementation of machine learning in education. The outcome of this study is to predict student's academic performance. The data of students is used to develop a model that can predict student's performance in academics considering some background data of the student. The input data for this study should be student's dataset. This dataset should be taken as a tabular format which contains information related to students (that is age, gender, academics record, medical information). There are many algorithms that are used in predictive models. In this study we focus on how linear regression is implemented to the student's academic performance considering student's dataset.

The effectiveness of using machine learning in education field depends on the algorithm and the usage of the data. Choosing the algorithm to implement predicting the students' performance is important. The accuracy of the result depends on the Machine learning algorithm. The algorithm used to prove the thesis in this paper is Linear Regression. Present studies show that academic performances of the students are also dependent on student's background and other attributes. Many research works confirm that apart from the past academic performances student's background and other attributes indeed got a significant influence over students' performance. Machine learning has an emerging role in recent times in every sector, and it can also be used effectively in academia.

Disadvantages:

- It cannot be early predict the data in a less time.
- It takes more time for training.

1.3 Proposed System and Advantages:

Students are essential advantages of producing high quality graduates who excel in academics, and creative thought for any school, college and other educational institution. In order to do this it is necessary to evaluate the output of students in any college, school or any other institution of education. Due to huge amount of data in educational databases, predicting the performance of students has became more difficult. Using educational machine learning methods, we could potentially improve the performance and progress of students more efficiently in an efficient manner.

The proposed system uses machine learning algorithms Decision Tree, Random Forest, Support Vector Machine and KNN classifier are used So that students, teachers and parents can get the knowledge on student performance in academics. The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject we will predict the marks and depending on the predicted marks of the student in each subject the teachers can easily monitor the student performance.

Decision Tree:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. In either instance they are constructed the same way and are always used to visualize all possible outcomes and decision points that occur chronologically. Decision trees are most commonly used in the financial world for areas such as loan approval, portfolio management, and spending. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the

branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

Random Forest:

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

K-Nearest Neighbor:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to

the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Advantages of Proposed System:

- The proposed system produces the results with high Accuracy
- It takes less time for training
- This system predicts better results

1.4 Objectives of the Project:

The main objective is to evaluate the student progress rate. So that students, teachers and parents can get the knowledge on student performance in academics and teachers can give the student a proper advice to avoid the poor results and also can improve the student performance. This project provides a system that can analyze and predict the student performance accurately by using machine learning algorithms. So that students, teachers and parents can get the knowledge on student performance in academics.

1.5 Flowchart of the Project:

The flowchart of the proposed system describes about the flow of the project. First the dataset is collected from Kaggle and the set is preprocessed which includes data cleaning and data is visualized to understand the dataset and then the data set is divided into training and testing dataset after that the algorithms are implemented to predict the results. The flow of the project is clearly shown in the below figure.

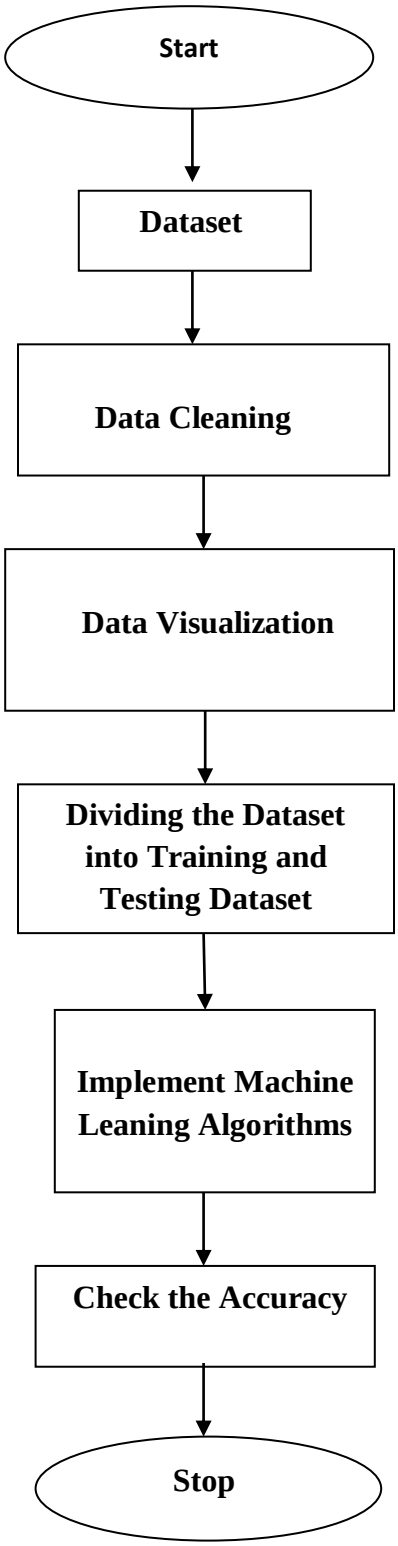


Fig 1.1 Flow Chart of the Project

2. REQUIREMENTS ANALYSIS

2.1 Introduction of Requirement Analysis:

Requirements analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product. These features, called requirements, must be quantifiable, relevant and detailed. Requirements analysis involves frequent communication with system users to determine specific feature expectations, resolution of ambiguity in requirements as demanded by the various users or groups of users, avoidance of feature creep and documentation of all aspects of the project development process from start to finish. Energy should be directed towards ensuring that the final system conforms to client needs rather than attempting to mold user expectations to fit the requirements.

This project is mainly based on machine learning where it uses the machine learning algorithms Decision Tree, Random Forest, Support Vector Machine, KNN classifier on historical data of student grades in one of the undergraduate courses and developed a model to predict the grades. To implement this project we need hardware and software requirements which are used to describe the working of the proposed methodology.

2.2 Hardware & Software Requirements:

Hardware Requirements:

- System : Pentium i3 Processor.
- Hard Disk : 500 GB.
- Ram : 2 GB

Software Requirements:

- Operating system : Windows 10.
- Coding Language : Python
- IDE : Google colab or Jupyter Notebook
- Streamlit
- Students Dataset

Processor:

A family of dual core, 64-bit, x86 CPUs from Intel intended for entry-level desktop and laptop computers. Introduced in 2010, the Core i3 is the third line in Intel's Core "i" branding (i3, i5, i7). The Core i3 chips do not include the Turbo Boost feature, which is found in i5 and i7 chips. The Core i5 chips are the midrange CPUs in the Core "i" line between the entry-level i3 and the high-performance i7 series. Certain models of the i5 family include the Turbo Boost feature. Core i5 processor is used to design the interface for this system. The purpose of this processor is An Intel Core i5 provides better performance against heavier and demanding applications and the Intel Core i5 comes in variations of two to four cores, all supporting four different threads simultaneously.

Hard Disk:

A hard drive is the tackle element that stores all of your digital content. Your documents, music, videos, programs, operation preferences, and operating system represent digital content stored on a hard drive. Hard drives can be external or internal. To store the different datasets from social media we require minimum 500 GB of hard disk.

RAM:

RAM stands for random access memory, and it's one of the most fundamental elements of computing. RAM is a temporary memory bank where your computer stores data it needs to retrieve quickly. RAM keeps data easily accessible so your processor can quickly find it without having to go into long-term storage to complete immediate processing tasks. To execute the proposed design we need minimum of 2 GB ram.

Operating System:

An operating system (OS) is the program that manages all the application programs in a computer. The application programs make use of the operating system by making requests for services through a defined application program interface (API). An operating system brings powerful benefits to computer software and software development. Without an operating system, every application would need to include its own UI, as well as the comprehensive code needed to handle all low-level functionality of the underlying

computer, such as disk storage, network interfaces and so on. Considering the vast array of underlying hardware available, this would vastly bloat the size of every application and make software development impractical. For our project we need the configuration of operating system as windows 10 or Linux that can only support the operations.

Python:

Python is a general-purpose, versatile, and powerful programming language. It's a great first language because Python code is concise and easy to read. Whatever you want to do, python can do it. From web development to machine learning to data science, Python is the language for you. Basically, a script is a text file containing the statements that comprise a Python program. Once you have created the script, you can execute it over and over without having to retype it each time. This is a very useful capability that allows you to type in a python program and to have it executed immediately in an interactive mode. To implement the project we have used the following libraries in python.

1. NumPy:

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.

Arrays are very frequently used in data science, where speed and resources are very important. It's the universal standard for working with numerical data in Python, and it's at the core of the scientific Python and PyData ecosystems. NumPy users include everyone from beginning coders to experienced researchers doing state-of-the-art scientific and industrial research and development. The NumPy API is used extensively in Pandas, SciPy, Matplotlib, scikit-learn, and most other data science and scientific Python packages.

2. Pandas:

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Pandas is built on top of the Numpy package, means Numpy is required for operating the Pandas. Before Pandas, Python was capable for data preparation, but it only provided limited support for data analysis. So, Pandas came into the picture and enhanced the capabilities of data analysis. It can perform five significant steps required for processing and analysis of data irrespective of the origin of the data, i.e., load, manipulate, prepare, model, and analyze**.**

3. Sklearn:

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. Scikit-learn offers a variety of supervised and unsupervised learning techniques. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Key concepts and features include:

- Algorithms for making decisions, such as Data are identified and categorised by classification as per the patterns, Regression is the process of forecasting or predicting data values using the historical and anticipated data average and Clustering is the automatic collection of datasets with related data.
- Predictive analysis is supported by various algorithms, including neural networks for pattern recognition and straightforward linear regression.
- Compatibility with the libraries of NumPy, pandas, and matplotlib. A predictive model can be built or trained on input data by computers using machine learning (ML), eliminating the need for explicit programming. A subset of AI is machine learning (AI).

Google Colab:

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Google Colab makes data science, deep learning, neural network, and machine learning accessible to individual researchers who cannot afford costly computational infrastructure. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

A Python program developer can use this notebook to write and execute random Python program codes just using a web browser. We can use Google Colabs like Jupyter notebooks. They are really convenient because Google Colab hosts them, so we don't use any of our own computer resources to run the notebook. We can also share these notebooks so other people can easily run our code, all with a standard environment since it's not dependent on our own local machines. However, we might need to install some libraries in our environment during initialization.

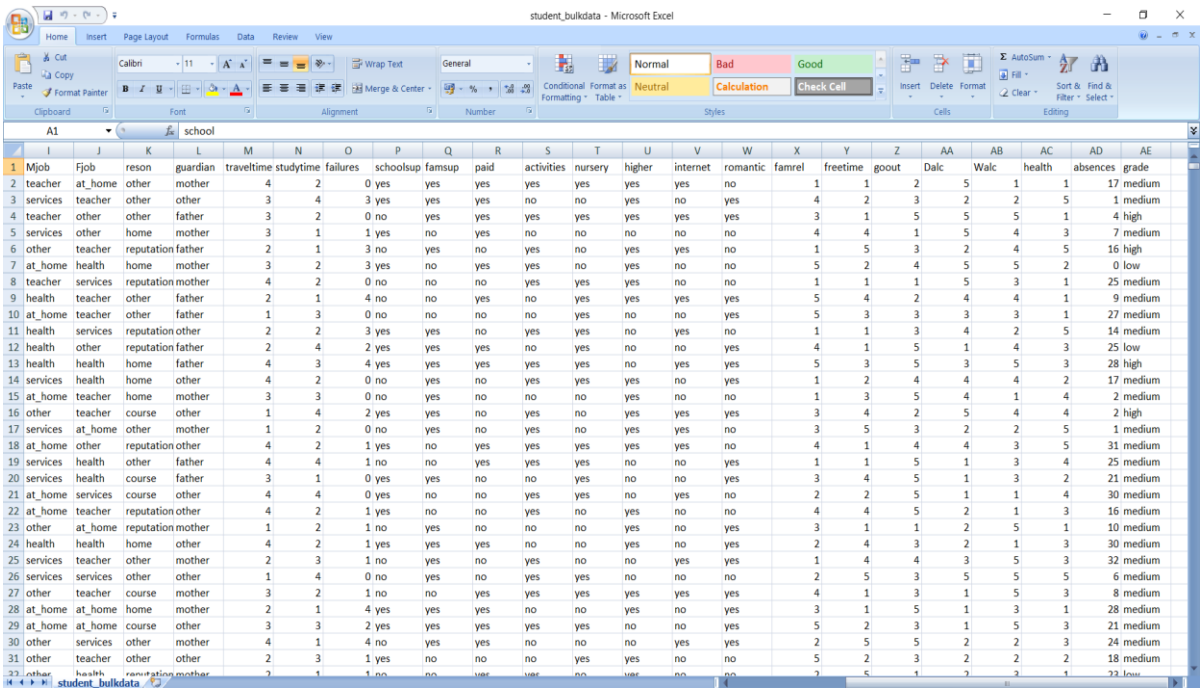
Streamlit:

Streamlit is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app. Streamlit allows you to write an app the same way you write a python code.

The streamlit has a distinctive data flow, any time something changes in your code or anything needs to be updated on the screen, streamlit reruns your python script entirely from the top to the bottom. This happens when the user interacts with the widgets like a select box or drop-down box or when the source code is changed. If you have some costly operations while rerunning your web app, like loading data from databases, you can use streamlit's st.cache method to cache those datasets, so that it loads faster.

Students Dataset:

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. Data sets can also consist of a collection of documents or files. We have taken the student dataset from kaggle on which we will apply data preprocessing techniques and we take the required attributes of the students to predict the marks in a subject by applying the machine learning algorithms Decision Tree, Random Forest, Support Vector Machine, KNN classifier.



	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Mjob	Fjob	reson	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	grade
2	teacher	at_home	other	mother	4	2	0	yes	yes	yes	yes	yes	yes	yes	no	1	1	2	5	1	1	17	medium
3	services	teacher	other	other	3	4	3	yes	yes	yes	no	no	yes	no	yes	4	2	3	2	2	5	1	medium
4	teacher	other	other	father	3	2	0	no	yes	yes	yes	yes	yes	yes	yes	3	1	5	5	5	1	4	high
5	services	other	home	mother	3	1	1	yes	no	yes	no	no	no	no	no	4	4	1	5	4	3	7	medium
6	other	teacher	reputation	father	2	1	3	no	yes	no	yes	no	yes	yes	no	1	5	3	2	4	5	16	high
7	at_home	health	home	mother	3	2	3	yes	no	yes	yes	no	yes	no	no	5	2	4	5	5	2	0	low
8	teacher	services	reputation	mother	4	2	0	no	no	no	yes	yes	yes	no	no	1	1	1	5	3	1	25	medium
9	health	teacher	other	father	2	1	4	no	no	yes	no	yes	yes	yes	yes	5	4	2	4	4	1	9	medium
10	at_home	teacher	other	father	1	3	0	no	no	no	no	yes	no	no	yes	5	3	3	3	3	1	27	medium
11	health	services	reputation	other	2	2	3	yes	yes	no	yes	yes	no	yes	no	1	1	3	4	2	5	14	medium
12	health	other	reputation	father	2	4	2	yes	yes	yes	no	yes	no	no	yes	4	1	5	1	4	3	25	low
13	health	health	home	father	4	3	4	yes	yes	yes	yes	yes	no	yes	yes	5	3	5	3	5	3	28	high
14	services	health	home	other	4	2	0	no	yes	no	yes	yes	yes	no	yes	1	2	4	4	4	2	17	medium
15	at_home	teacher	home	mother	3	3	0	no	yes	no	no	no	yes	no	no	1	3	5	4	1	4	2	medium
16	other	teacher	course	other	1	4	2	yes	yes	no	yes	no	yes	yes	yes	3	4	2	5	4	4	2	high
17	services	at_home	other	mother	1	2	0	no	yes	no	yes	no	yes	yes	no	3	5	3	2	2	5	1	medium
18	at_home	other	reputation	other	4	2	1	yes	no	yes	yes	yes	yes	yes	no	4	1	4	4	3	5	31	medium
19	services	health	other	father	4	4	1	no	no	yes	yes	yes	no	no	yes	1	1	5	1	3	4	25	medium
20	services	health	course	father	3	1	0	yes	yes	no	no	yes	no	no	yes	3	4	5	1	3	2	21	medium
21	at_home	services	course	other	4	4	0	yes	no	no	yes	yes	no	yes	no	2	2	5	1	1	4	30	medium
22	at_home	teacher	reputation	other	4	2	1	yes	no	no	yes	no	yes	no	no	4	4	5	2	1	3	16	medium
23	other	at_home	reputation	mother	1	2	1	no	yes	no	no	no	yes	no	yes	3	1	1	2	5	1	10	medium
24	health	health	home	other	4	2	1	yes	yes	yes	no	no	yes	no	yes	2	4	3	2	1	3	30	medium
25	services	teacher	other	mother	2	3	1	no	yes	no	yes	no	no	yes	yes	1	4	4	3	5	3	32	medium
26	services	services	other	other	1	4	0	no	yes	no	yes	yes	no	no	no	2	5	3	5	5	5	6	medium
27	other	teacher	course	mother	3	2	1	no	no	yes	yes	yes	yes	yes	yes	4	1	3	1	5	3	8	medium
28	at_home	at_home	home	mother	2	1	4	yes	yes	yes	no	no	yes	no	yes	3	1	5	1	3	1	28	medium
29	at_home	at_home	course	other	3	3	2	yes	yes	yes	yes	yes	no	no	yes	5	2	3	1	5	3	21	medium
30	other	services	other	mother	4	1	4	no	yes	yes	no	no	no	yes	yes	2	5	5	2	2	3	24	medium
31	other	teacher	other	other	2	3	1	yes	no	no	no	yes	yes	no	no	5	2	3	2	2	2	18	medium

Fig 2.1 Students Dataset

	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1518	at_home	services	reputation	other	4	1	2	yes	no	no	yes	yes	no	no	no	1	3	1	2	3	2	10	medium
1519	other	at_home	course	other	4	2	2	yes	no	yes	no	yes	yes	no	yes	1	5	3	2	2	5	8	medium
1520	services	teacher	home	father	3	4	0	no	no	yes	yes	no	no	yes	no	2	4	4	1	5	3	30	medium
1521	teacher	health	other	mother	3	4	2	no	yes	no	yes	no	no	yes	no	5	1	1	3	5	3	24	medium
1522	teacher	teacher	course	father	4	2	2	yes	yes	yes	no	yes	yes	yes	no	5	4	5	1	4	5	22	medium
1523	teacher	health	course	mother	2	3	0	yes	yes	yes	yes	no	yes	no	no	4	1	3	1	5	3	22	high
1524	at_home	other	course	father	3	2	0	yes	no	yes	no	yes	no	no	no	5	1	3	1	5	5	16	high
1525	at_home	teacher	home	mother	4	3	1	yes	no	yes	yes	no	yes	yes	yes	5	1	4	1	2	3	5	high
1526	other	teacher	course	mother	4	4	2	no	yes	no	no	no	no	no	no	5	4	2	5	5	1	18	medium
1527	health	services	reputation	father	2	3	0	yes	yes	no	yes	no	yes	yes	no	4	5	1	5	2	5	18	medium
1528	teacher	health	reputation	other	1	4	4	no	no	no	yes	no	yes	yes	no	2	3	4	3	4	1	3	high
1529	other	other	home	father	1	1	3	no	yes	no	no	no	yes	yes	yes	2	5	1	1	3	3	24	medium
1530	teacher	other	reputation	other	1	1	2	no	yes	no	no	no	yes	yes	no	1	3	4	4	4	3	5	high
1531	at_home	teacher	reputation	mother	1	3	3	yes	yes	no	no	no	no	no	yes	2	5	4	2	2	4	18	medium
1532	other	services	reputation	mother	1	3	1	no	no	yes	yes	no	no	no	yes	5	5	1	1	5	2	22	low
1533	other	at_home	course	father	3	4	4	no	yes	no	no	yes	yes	yes	yes	5	1	2	5	5	3	8	medium
1534	at_home	health	reputation	mother	1	1	0	no	no	no	yes	no	yes	yes	yes	1	3	5	3	3	5	5	medium
1535	at_home	health	home	mother	1	2	0	yes	no	yes	yes	yes	yes	no	yes	2	2	3	5	5	1	7	medium
1536	other	at_home	other	mother	2	4	3	no	yes	no	no	no	yes	yes	yes	1	5	2	4	3	5	1	medium
1537	teacher	services	course	other	4	3	1	no	no	yes	no	no	no	yes	yes	4	3	5	4	1	1	2	high
1538	other	other	reputation	father	4	3	2	no	yes	yes	yes	no	no	no	no	1	5	5	2	3	1	10	medium
1539	health	other	other	father	4	4	2	no	no	yes	yes	no	no	yes	yes	2	5	1	5	1	3	30	medium
1540	health	teacher	course	other	4	4	2	no	no	yes	no	no	no	yes	yes	5	4	1	5	2	5	32	medium
1541	teacher	health	course	other	4	1	0	yes	yes	yes	no	yes	yes	yes	no	1	3	1	5	4	1	23	low
1542	health	at_home	other	other	4	4	0	yes	no	no	no	no	no	no	no	3	2	5	4	2	4	6	medium
1543	health	teacher	reputation	other	3	2	1	yes	no	no	no	yes	yes	no	no	5	4	5	1	2	4	8	medium
1544	services	teacher	course	other	4	3	4	yes	no	yes	no	no	yes	yes	yes	1	2	1	4	4	2	23	medium
1545	health	at_home	other	father	4	3	1	no	yes	no	yes	no	no	yes	no	3	3	5	2	1	4	20	medium
1546	at_home	services	home	mother	3	1	0	no	yes	yes	no	yes	yes	no	no	4	1	4	1	4	5	22	medium
1547	other	health	course	father	1	4	4	no	no	no	yes	no	yes	no	no	4	5	4	1	5	1	12	medium
1548	health	teacher	course	mother	2	2	1	yes	yes	no	no	no	no	yes	no	1	3	4	2	5	2	10	low
1549	teacher	teacher	course	other	4	1	3	no	no	no	no	yes	no	no	yes	4	5	5	5	5	1	23	medium

Fig 2.2 Students Dataset

	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
2789	health	at_home	other	mother	3	4	0	no	yes	yes	no	no	no	yes	no	3	5	2	5	4	4	6	medium
2790	at_home	services	home	father	3	4	4	no	yes	no	yes	no	yes	yes	no	3	3	5	2	3	2	15	medium
2791	health	teacher	reputation	other	3	2	4	no	no	no	yes	yes	yes	yes	yes	5	1	4	2	5	4	0	low
2792	at_home	health	course	father	4	4	1	yes	no	no	yes	no	no	no	yes	4	1	3	3	1	1	18	high
2793	teacher	services	home	other	2	3	4	no	yes	yes	yes	yes	no	yes	no	5	4	2	3	4	5	0	high
2794	health	at_home	other	mother	2	2	3	no	yes	no	yes	no	yes	yes	yes	3	3	1	3	4	3	12	high
2795	services	health	other	father	4	1	2	no	no	yes	yes	yes	no	yes	yes	5	3	5	1	5	4	26	medium
2796	at_home	health	reputation	mother	3	1	3	no	no	no	no	no	yes	yes	yes	1	1	2	3	3	4	6	high
2797	health	at_home	course	father	1	2	3	no	yes	no	no	yes	yes	yes	yes	3	4	3	3	5	3	29	high
2798	other	health	reputation	other	4	4	1	no	no	no	yes	yes	yes	no	no	2	4	4	1	2	2	4	medium
2799	at_home	teacher	other	mother	4	1	3	yes	no	yes	no	no	no	yes	no	5	2	5	5	2	5	32	medium
2800	services	health	other	mother	2	2	3	yes	no	no	yes	no	no	yes	no	1	4	5	2	4	3	2	medium
2801	at_home	services	course	mother	1	2	0	yes	yes	yes	no	yes	yes	yes	yes	3	5	3	4	1	2	16	medium
2802	health	teacher	other	mother	2	4	2	no	no	yes	yes	yes	yes	yes	no	2	1	5	2	3	1	21	medium
2803	health	teacher	other	mother	2	4	4	no	yes	no	no	yes	no	yes	no	3	5	3	3	1	4	4	low
2804	health	other	course	mother	4	4	3	no	yes	yes	yes	yes	yes	yes	yes	2	5	4	2	5	2	15	medium
2805	services	at_home	other	mother	4	4	2	yes	no	yes	no	yes	no	no	yes	2	1	3	5	4	3	30	medium
2806	health	services	course	father	1	4	2	no	yes	yes	no	yes	yes	no	yes	2	4	3	4	3	1	1	medium
2807	health	teacher	other	other	2	4	0	yes	no	yes	yes	no	yes	no	yes	1	4	5	1	5	2	17	high
2808	teacher	services	other	other	1	4	1	no	no	no	yes	no	no	yes	yes	5	2	2	3	5	1	7	high
2809	at_home	services	home	father	4	3	4	no	yes	no	yes	no	yes	yes	yes	4	3	2	5	3	5	14	medium
2810	at_home	other	course	father	4	1	4	no	yes	no	no	no	yes	no	no	5	4	4	4	5	1	20	low
2811	services	at_home	course	mother	2	3	3	no	no	no	no	no	yes	yes	yes	3	3	2	1	3	5	5	high
2812	teacher	other	course	father	3	3	1	yes	yes	no	yes	yes	no	no	no	5	2	5	3	3	2	7	medium
2813	other	services	other	father	3	3	1	yes	no	no	no	no	yes	no	no	4	2	4	2	3	3	18	medium
2814	health	health	home	mother	3	4	1	no	no	no	no	no	yes	yes	yes	5	1	3	2	2	1	8	medium
2815	health	health	course	father	3	4	1	no	yes	yes	no	yes	yes	yes	no	1	4	1	1	1	5	13	medium
2816	at_home	other	home	mother	2	1	4	yes	yes	yes	yes	no	no	yes	yes	1	2	4	3	2	4	9	medium
2817	health	health	course	mother	3	2	3	no	no	no	yes	no	yes	yes	no	4	3	1	1	1	1	22	high
2818	health	at_home	home	father	4	3	3	yes	yes	yes	yes	no	no	yes	yes	3	3	1	1	4	4	11	high
2819	health	health	home	father	1	4	2	yes	yes	no	no	yes	yes	no	no	3	4	4	1	2	2	2	medium
2820	health	other	other	father	3	4	2	no	no	yes	no	yes	yes	yes	no	4	3	3	4	3	3	8	high

Fig 2.3 Students Dataset

2.3 Software Requirements Specification:

The production of the requirements stage of the software development process is Software Requirements Specifications (SRS) (also called a requirements document). This report lays a foundation for software engineering activities and is constructing when entire requirements are elicited and analyzed. SRS is a formal report, which acts as a representation of software that enables the customers to review whether it (SRS) is according to their requirements. Also, it comprises user requirements for a system as well as detailed specifications of the system requirements.

The SRS is a specification for a specific software product, program, or set of applications that perform particular functions in a specific environment. It serves several goals depending on who is writing it. First, the SRS could be written by the client of a system. Second, the SRS could be written by a developer of the system. The two methods create entirely various situations and establish different purposes for the document altogether. The first case, SRS, is used to define the needs and expectation of the users. The second case, SRS, is written for various purposes and serves as a contract document between customer and developer.

A software requirements specification is a document that describes requirements for a software product, program or set of programs. Requirements in the software requirements specification are expressed in normal language and are not concerned with technical implementation. That's what the design documents are for. In software development, the software requirements specification represents the results of the requirements analysis and describes the requirements of the software under development. Though it is traditionally created as a document, it can also be created in different forms, for example – a very simple one – in spoken form.

The SRS is a specification for a specific software product, program, or set of applications that perform particular functions in a specific environment. It serves several goals depending on who is writing it. First, the SRS could be written by the client of a system. Second, the SRS could be written by a developer of the system. The two methods

create entirely various situations and establish different purposes for the document altogether. The first case, SRS, is used to define the needs and expectation of the users.

The second case, SRS, is written for various purposes and serves as a contract document between customer and developer. The software requirements specification document lists sufficient and necessary requirements for the project development.To derive the requirements, the developer needs to have clear and thorough understanding of the products under development. This is achieved through detailed and continuous communications with the project team and customer throughout the software development process.

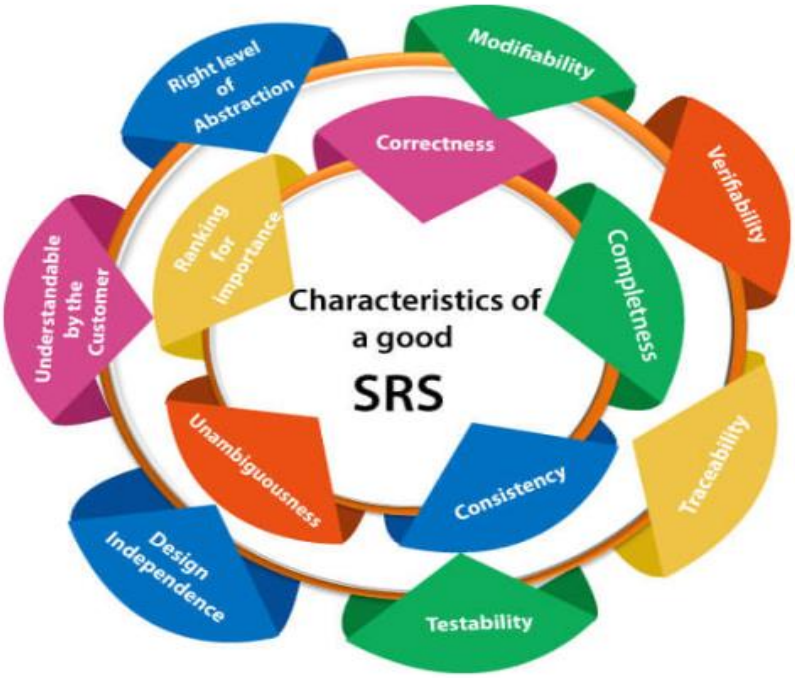


Fig 2.4 Characteristics of SRS

3. LITERATURE SURVEY

[1] Boddeti Sravani and Myneni Madhu Bala, "Prediction of Student Performance Using Linear Regression". 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, pp. 1-5, 2020

In this paper they implemented linear regression, which is machine learning algorithm to predict the student's performance in academics. This study focuses on supervised learning, more specifically on predictive analysis. Predicting student's academic performance is very important because it can indicate the teachers about the students who are possibly to drop out from the course and prediction can provide additional assistance to the students who need to improve their academic performance. The dataset used in this paper comprised of 100 students, which is a small dataset.

This study is on implementation of machine learning in education. The outcome of this study is to predict student's academic performance. The data of students is used to develop a model that can predict student's performance in academics considering some background data of the student. The input data for this study should be student's dataset. This dataset should be taken as a tabular format which contains information related to students (that is age, gender, academics record, medical information). In this study they focused on how linear regression is implemented to the student's academic performance considering student's dataset [1].

[2] D K Arun, V Namratha, B V Ramyashree, Yashita P Jain, Antara Roy Choudhary, ''Student Academic Performance Prediction Using Educational Data Mining'' Int. J. Pure Appl. Math, vol. 118, no. 18, pp. 2703–2707, 2021

This paper discusses about the field of Educational Data Mining concentrates on prediction more frequently as compared to generating exact results for future prediction. In order to keep track of the changes being in class patterns, a regular analysis must be there in the educational field. This paper presents the original results from an Educational data mining exploration design enforced by collecting huge data sets from BMS College of Engineering.

It aimed at revealing the high eventuality of data mining operations for university operation and preventing the students from the risk of getting a low- Grade Point Average(GPA). WEKA platform was used to perform analysis with different machine learning models for both the classification and regression problems. Various classification and regression algorithms were analyzed and tested for accuracy for a student's performance in subject wise analysis and GPA prediction. For the former, ensemble system using voting was performed on the classification algorithms. They were also distributed in different groups and compared based on certain factors, the results of which are discussed in this paper [2].

[3] E. Alyahyan and D. Düştegör, ‘‘Predicting academic success in higher education: Literature review and best practices,’’ Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, pp. 93-97, Dec. 2020

This paper mainly focuses on student success, that plays a vital part in educational institutions, as it's frequently used as a metric for the institution’s performance. Beforehand discovery of scholars at threat, along with preventative measures, can drastically improve their success. Recently, machine learning algorithms have been considerably used for prediction purpose. While there is a plethora of success stories in the literature, these techniques are mainly accessible to “computer science”, or more precisely, “artificial intelligence” literate educators. Indeed, the effective and efficient application of data mining methods entail many decisions, like how to define student’s success, through which student attributes to focus on, up to which machine learning method is more appropriate to the given problem.

This study aims to provide a step-by-step set of guidelines for educators willing to apply data mining techniques to predict student success. For this, the literature has been reviewed, and the state-of-the-art has been compiled into a systematic process, where possible decisions and parameters are comprehensively covered and explained along with arguments. This study will provide to educators an easier access to data mining techniques, enabling all the potential of their application to the field of education [3].

[4] Manjari Chitti1, Padmini Chitti 2, Manoj Jayabalan: "Need for Interpretable Student Performance Prediction." 2020 IEEE 13th International Conference on Developments in eSystems Engineering (DeSE), pp. 269-272, 2020.

The education domain is growing at an exponential rate, maturing with the introduction of innovative and improved offerings to the learning community by using various Education Data Mining techniques (EDM). Significant amount of research is being carried out around EDM, using various dimensions like student's performance, dropout rates, individual cognitive capabilities, teacher's, administrators' performance, course delivery, and content authoring. Several data analytics techniques are used to gain knowledge from educational data. However, the prediction models generated are complex and not interpretable in understanding why and how a prediction has arrived at from the produced results.

The interpretability of the model becomes important when dealing with large datasets and complex models. The derived analytics and the models will be acceptable and trusted if they are interpretable. It helps the educational organizations to advance their learning processes in incfreasing students' performance and reduction in the student dropout rate. This study reviews EDM focusing on the factors influencing student's predictions, various algorithms used, and identified the gaps. The study also gives an insight into how the "black-box" decisions of the prediction model are made, the role of various Artificial intelligence techniques in making the model results interpretable, and their contribution in producing efficient results [4].

[5] E. Tatar and D. Düştegör, ''Prediction of academic performance at undergraduate graduation: Course grades or grade point average'' Appl. Sci., vol. 10, no. 14, pp. 1–15, 2020

A private university must consider numerous effects in accepting prospective scholars. scholars enrolled are anticipated to stay until their studies are completed, have good academic performance, and be suitable to graduate on time. Private universities, from the morning of the admission of new scholars, it's necessary to choose which students are accepted to achieve the quality of education in the study program. This paper aims to study

the class and class order of variable significance to scholars' length of stay and academic performance labeled scale. The system is used to predict the performance.

This study uses rank methods information gain and gain ratio to help other methods $\chi 2$ and random forest. A dataset of 7676 students, from 2010- 2021, from a operation program of a private university in Indonesia, is used. This study collects data from the faculty-specific department from the university's academic admissions as inputs. The result of the study shows that is the student reaching the specific requirements. Origin of High School, Selection Test Score, and Gender get resolve votes. This study is unique because it sheds light on the case particularity to Indonesia [5].

[6] Hartatik Kusrini Agung Budi Prasetio "Prediction of Student Graduation with Naive Bayes Algorithm" 2021 at 13:04:34 UTC from IEEE Xplore, pp. 103-111, 2021.

The research carried out in this study is the development and analysis of student performance in the academic field using the Naive Bayes algorithm so that it can help teachers and students see early predictions, and help teachers to see the progress and predictions of student. The purpose of this research is to study student achievement prediction models that have model values. The prediction model optimization step by selecting the variable used in the prediction model development is IPS1,2,3,4. The data used in this study are the results of observations from universities.

The result of this research is the prediction of Student Achievement Development with Naive Bayes Algorithm based on the marks of the student in semester 1.2,3,4 variable and added value are UN rate, Gender, and status stay. From the results of research conducted it was found that the application of the naïve Bayes algorithm for model 1 is a model prediction using variable IP students result in an accuracy of 75% and R2 = 68,2%. Model 2 for prediction are used 8 variables namely Nim, Gender, Residence, IPS 1, IPS2, IPS3, IPS4, study period and student status result accuracy of prediction 89% and with a prediction level of R2 = 71,4%. This certainly improves the performance of the training data efficiency prediction model [6].

Table 3.1 Comparison of Literature Review

S.No	Title	Proposed Methodology	Demerits/Future Direction
1	**Prediction of Student Performance Using Linear Regression**	In this paper they implemented linear regression, which is machine learning algorithm to predict the student's performance in academics.	The dataset used in this paper comprised of 100 students, which is a small dataset.
2	**Student Academic Performance Prediction Using Educational Data Mining**	This paper presents the original results from an Educational data mining exploration design enforced by collecting sets from BMS College of Engineering.	This require more time for training dataset
3	**Predicting academic success in higher education: Literature review and best practices**	This paper mainly focuses on student success, that plays a vital part in educational institutions, as it's frequently used as a metric for the institution's performance.	This paper could not provide the accurate results

4	**Need for Interpretable Student Performance Prediction.**	In this paper significant amount of research is being carried out around EDM, using various dimensions like student's performance, dropout rates, individual cognitive capabilities, teacher's, administrators' performance, course delivery, and content authoring.	This cannot be applied on large dataset
5	**Prediction of academic performance at undergraduate graduation: Course grades or grade point average**	This paper aims to study the class and class order of variable significance to scholars length of stay and academic performance labeled scale.	This paper uses High School, Selection Test Score which not much accurate as less attributes are used

6	**Prediction of Student Graduation with Naive Bayes Algorithm**	The data used in this study are the results of observations from universities. The result of this research is the prediction of Student Achievement Development with Naive Bayes Algorithm based on the marks of the student in semester 1.2,3,4 variable and added value are UN rate, Gender, and status stay.	In this there is comparision of 2 models and small dataset is used

4. MODULES

4.1 Introduction to Modules:

A module is a collection of source files and build settings that let you divide your project into discrete units of functionality. We have divided the project into 6 modules which are the steps involved in analysis and prediction of students performance using machine learning. The 6 modules include

- Data Collection
- Data Preprocessing
- Splitting the data into training and testing
- Algorithm implementation on training dataset
- Accuracy score on test dataset
- Making the predictions with the model

4.2 Module Implementation:

Data Collection:

Data collection is the process of gathering the information which is the first step in implementation of a machine learning model. There are many techniques to collect the data like web scraping, manual interventions etc. We have collected the student's dataset from "kaggle" website.

Data Preprocessing:

Data Preprocessing involves data cleaning and data transformation. The dataset is first checked for duplicates and null values and they are removed for better performance of the machine learning algorithms. and we take the required attributes of the students from the dataset.

Splitting the data into training and testing:

The preprocessed dataset will be splitted into training and testing dataset. where the algorithms will be trained upon the training dataset and testing data set will be tested for predicting the results.

Algorithm implementation on training dataset:

The proposed system uses machine learning algorithms Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and KNN classifier are used So that students, teachers and parents can get the knowledge on student performance in academics. The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject we will predict the marks and depending on the predicted marks of the student in each subject the teachers can easily monitor the student performance.

Decision Tree is a tree-structured Classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Random Forest algorithm is a machine learning algorithm that grows and combines multiple decision trees to create a forest. KNN algorithm classifies the dataset into different categories and it finds the distance between the test data for which output is needed. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. In this the dataset will be trained with the machine learning algorithms, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor.

Accuracy on testing dataset:

The testing dataset will generate the accuracy which will vary based on the machine learning algorithms. Accuracy in machine learning is an evaluation metric that measures the number of correct predictions made by model in relation to the number of predictions made. We calculate it by dividing number of correct predictions to total number of predictions.

Making the predictions:

Depending on the accuracy of the machine learning algorithms, the algorithm with most accuracy will be used to predict the student performance.

5. SYSTEM DESIGN

5.1 Introduction to System Design:

The system design activity starts after the system requirements analysis has been completed. System design is the art of procuring and deploying appropriate resources to achieve system goals. The system design activity starts after the system requirements analysis has been completed. System development is the process of creating or altering systems, along with the processes, practices, models, and methodologies used to develop them. The purpose of the system design process is to provide sufficient detailed data and information about the system. System Design focuses on needs and required functionality.

5.2 Block diagram of proposed system:

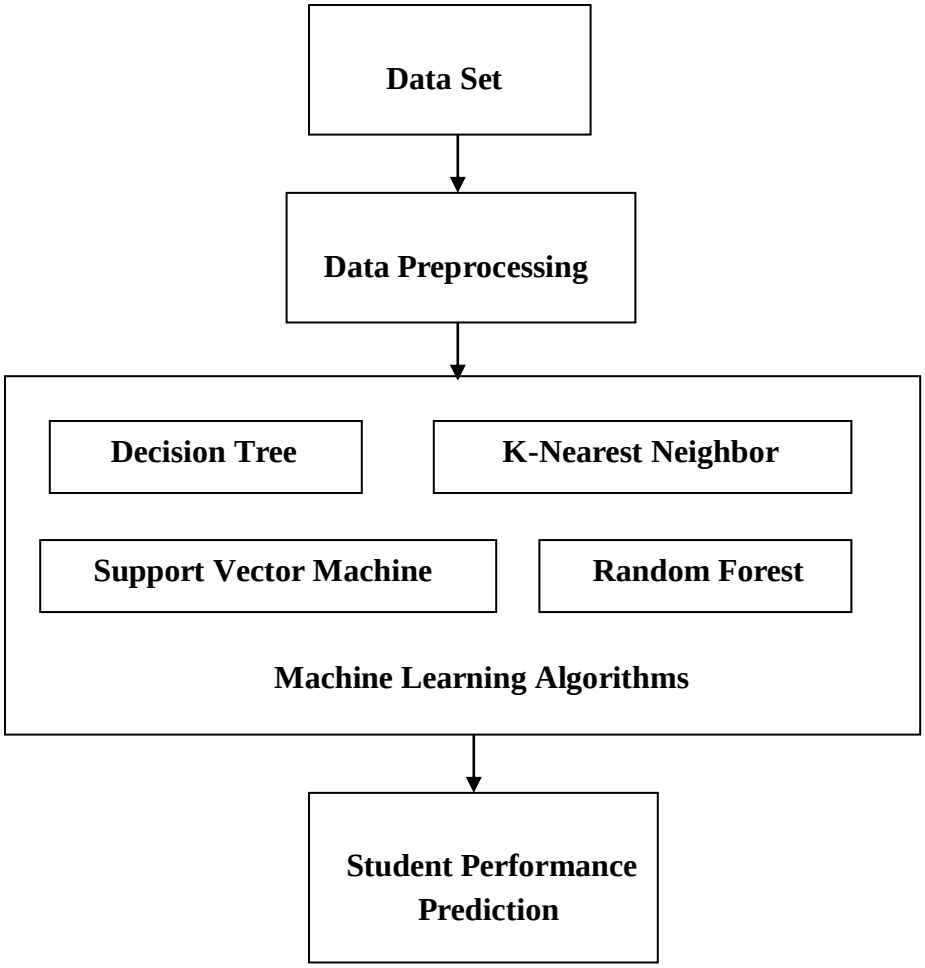


Fig 5.1 Block Diagram

5.3 Input Design and Output Design:

Input Design:

Input is the raw data that is processed to produce output. The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

The objectives of input design are

- To design data entry and input procedures
- To reduce input volume
- To design source documents for data capture or devise other data capture methods
- To design input data records, data entry screens, user interface screens, etc.
- To use validation checks and develop effective input controls.

Output Design:

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts. A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs.

The objectives of output design are

- To develop output design that serves the intended purpose and eliminates the production of unwanted output.
- To develop the output design that meets the end users requirements.
- To deliver the appropriate quantity of output.
- To form the output in appropriate format and direct it to the right person.
- To make the output available on time for making good decisions.

5.4 UML Diagrams:

UML, which stands for Unified Modeling Language, is a way to visually represent the architecture, design, and implementation of complex software systems. When you're writing code, there are thousands of lines in an application, and it's difficult to keep track of the relationships and hierarchies within a software system. UML diagrams divide that software system into components and subcomponents.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects-oriented software and the software development process. UML diagrams divide that software system into components and subcomponents. The UML uses mostly graphical notations to express the design of software projects. The Object Management Group is a consortium that created the UML standard. This topic provides a brief overview of UML. The basic difference between the entity-relationship model and the UML model is that, instead of designing entities as this information illustrates, you model objects.

UML is a standardized modeling language that can be used across different programming languages and development processes, so the majority of software developers will understand it and be able to apply it to their work. Though many engineers dread diagrams, they're useful in an Agile development environment: they keep development productive and focused. Instead of thinking them as just a "nice to have," treat your UML diagrams as core aspects of documentation. UML diagrams can help engineering teams:

- ✓ Bring new team members or developers switching teams up to speed quickly.
- ✓ Navigate source code.
- ✓ Plan out new features before any programming takes place.
- ✓ Communicate with technical and non-technical audiences more easily.

5.4.1 Class Diagram:

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. Class

diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object-oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram. The purpose of class diagram is to model the static view of an application. Class diagrams are the only diagrams which can be directly mapped with object-oriented languages and thus widely used at the time of construction. Here, we have implemented two classes with their attributes. Classes are System, User and their attributes can be represented with the help of class diagram.

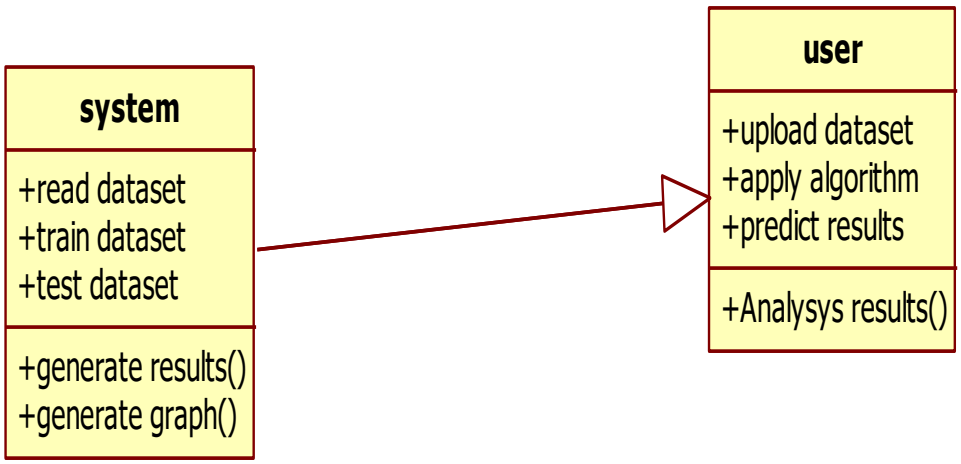


Fig 5.2 Class Diagram

5.4.2 Use case 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. These diagrams also identify the interactions between the system and its actors.

The use cases and actors in use-case diagrams describe what the system does and how the actors use it, but not how the system operates internally. Use-case diagrams illustrate and define the context and requirements of either an entire system or the important parts of the system. Use-case diagrams would typically develop in the early phases of a project and refer to them throughout the development process. 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.

UML use case diagrams are ideal for:

- Representing the goals of system-user interactions.
- Defining and organizing functional requirements in a system.
- Specifying the context and requirements of a system.
- Modeling the basic flow of events in a use case.

The use cases and actors in use-case diagrams describe what the system does and how the actors use it, but not how the system operates internally. Use-case diagrams illustrate and define the context and requirements of either an entire system or the important parts of the system. We can model a complex system with a single use-case diagram, or create many use-case diagrams to model the components of the system. Use-case diagrams would typically develop in the early phases of a project and refer to them throughout the development process.

While gathering requirements, we can create use-case diagrams to capture the requirements and to present to others what the system should do during the analysis and design phases, we can use the use cases and actors from your system. Use-case diagrams to identify the classes that the system requires. During the testing phase, we can use use-case diagrams to identify tests for the system.

The components of Use Case diagram are:

Actors: An actor represents a role of a user that interacts with the system that you are modelling. The user can be a human user, an organization, a machine, or another external system, usually drawn like skeleton.

Use Cases: A use case describes a function that a system performs to achieve the user's goal A use case must yield an observed result that is of value to the user of the system. They are represented by oval shape and named with the function.

Relationships in use case diagrams: Relationships between actors and use cases- represented using straight arrows. There are five types of relationships in use case diagrams they are: using straight arrows. They are Association between an actor and a use case, Generalization of an actor, Extend relationship between two use cases, Include relationship between two use cases, Generalization of a use case.

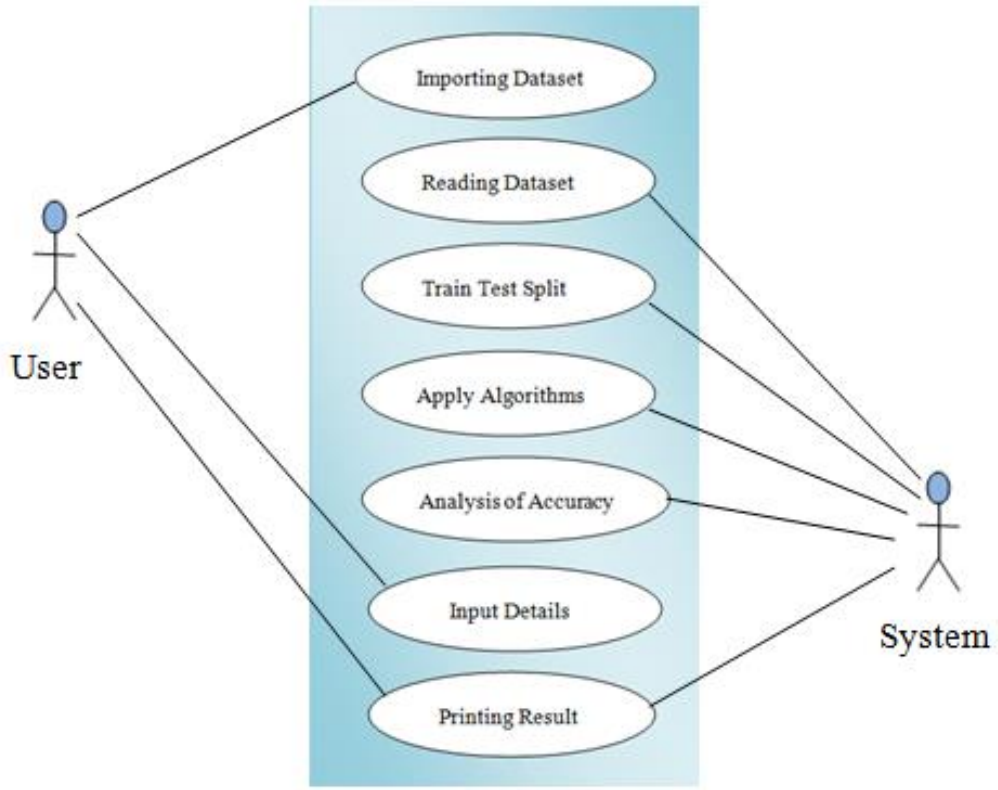


Fig 5.3 Use case Diagram

5.4.3 Sequence Diagram:

A sequence diagram is a type of interaction diagram because it describes how and in what order a group of objects works together. These diagrams are used by software developers and business professionals to understand requirements for a new system or to document an existing process. Sequence diagrams are sometimes known as event diagrams or event scenarios. Sequence diagrams can be useful references for businesses and other organizations.

Try drawing a sequence diagram to:

- Represent the details of a UML use case.
- Model the logic of a sophisticated procedure, function, or operation.
- See how objects and components interact with each other to complete a process.
- Plan and understand the detailed functionality of an existing or future scenario.

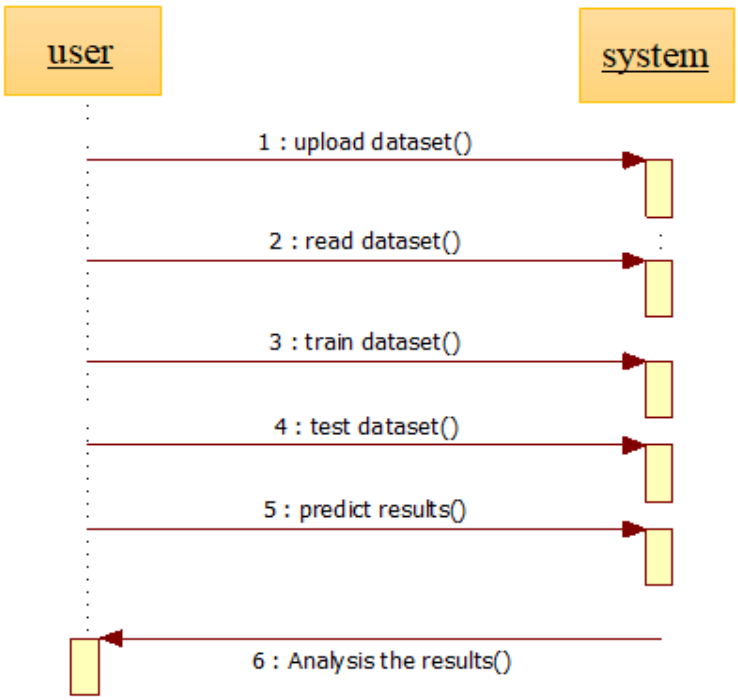


Fig 5.4 Sequence Diagram

5.4.4 Collaboration Diagram:

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization where as the collaboration diagram shows the object organization.

The collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system.

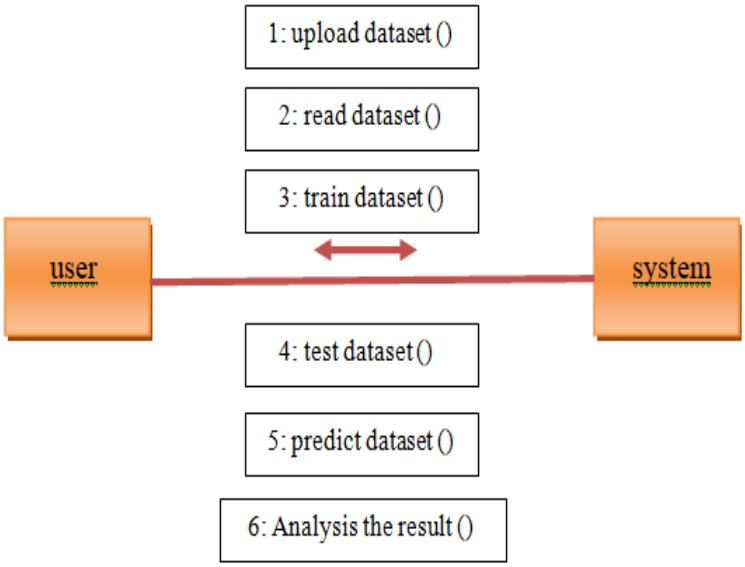


Fig 5.5 Collaboration Diagram

5.4.5 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. An activity diagram visually presents a series of actions or flow of control in a system similar to a flowchart or a data flow diagram.

Typically, an event needs to be achieved by some operations, particularly where the operation is intended to achieve a number of different things that require coordination, or how the events in a single use case relate to one another, in particular, use cases where activities may overlap and require coordination. Activity diagrams are often used in business process modeling. They can also describe the steps in a use case diagram. Activities modeled can be sequential and concurrent. In both cases an activity diagram will have a beginning (an initial state) and an end (a final state).

Basic Activity Diagram Notations and Symbols:

Initial State or Start Point: A small filled circle followed by an arrow represents the initial action state or the start point for any activity diagram. For activity diagram using swimlanes, make sure the start point is placed in the top left corner of the first column.

Activity or Action State: An action state represents the non-interruptible action of objects. You can draw an action state in SmartDraw using a rectangle with rounded corners.

Action Flow: Action flows, also called edges and paths, illustrate the transitions from one action state to another. They are usually drawn with an arrowed line.

Object Flow: Object flow refers to the creation and modification of objects by activities. An object flow arrow from an action to an object means that the action creates or influences the object and it indicates that the action state uses the object.

Decisions and Branching: A diamond represents a decision with alternate paths. When an activity requires a decision prior to moving on to the next activity, add a diamond between the two activities. The outgoing alternates should be labeled with a condition or guard expression. You can also label one of the paths "else".

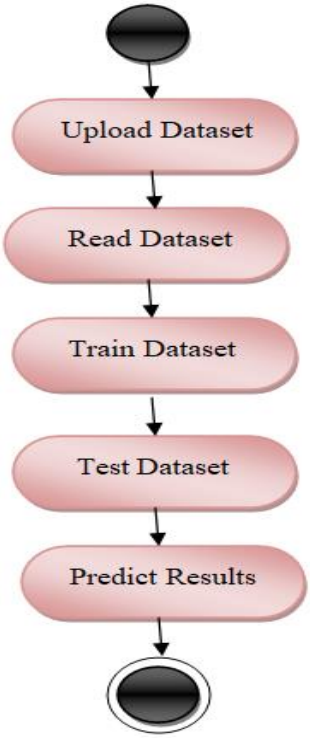


Fig 5.6 Activity Diagram

6. SYSTEM IMPLEMENTATION

6.1 Introduction:

Systems implementation is a set of procedures performed to complete the design (as necessary) contained in the approved systems design document and to test, install, and begin to use the new or revised Information System. Implementation is the process that actually yields the lowest-level system elements in the system hierarchy (system breakdown structure). System elements are made, bought, or reused. Production involves the hardware fabrication processes of forming, removing, joining, and finishing, the software realization processes of coding and testing, or the operational procedures development processes for operators' roles.

If implementation involves a production process, a manufacturing system which uses the established technical and management processes may be required. In the prediction step, the model is used to predict the response for given data. We used four algorithms to implement the model. The purpose of the implementation process is to design and create (or fabricate) a system element conforming to that element's design properties and/or requirements. The element is constructed employing appropriate technologies and industry practices.

Decision Tree:

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. In either instance they are constructed the same way and are always used to visualize all possible outcomes and decision points that occur chronologically. Decision trees are most commonly used in the financial world for areas such as loan approval, portfolio management, and spending. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas

Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, andbased on the answer (Yes/No), it further split the tree into sub trees.

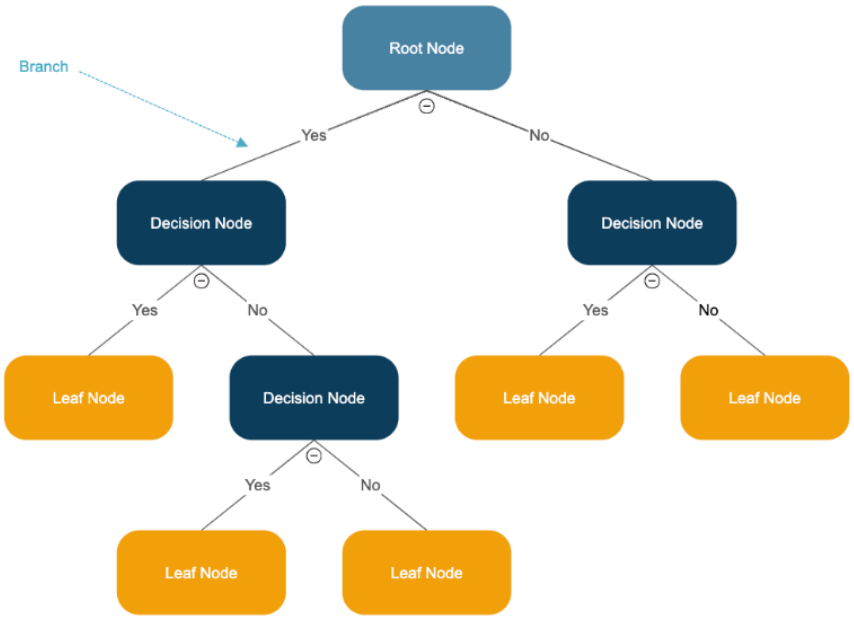


Fig 6.1 Decision Tree

Random Forest:

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains

several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

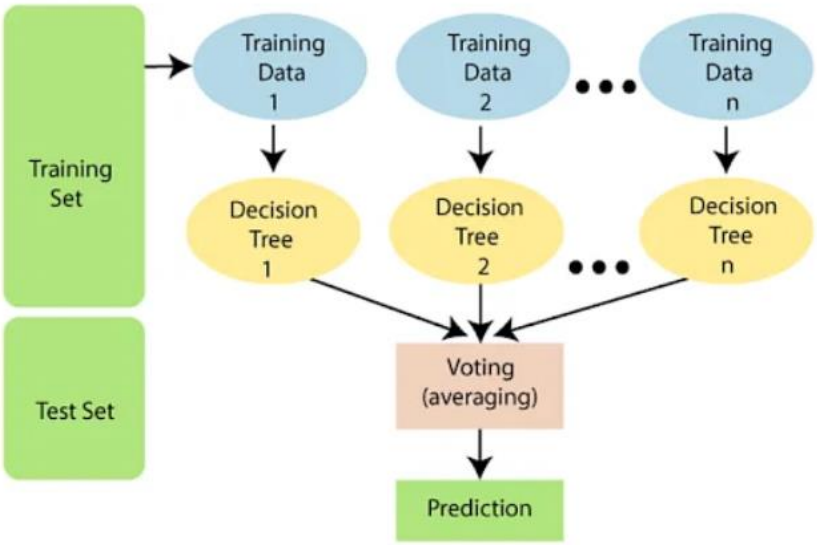


Fig 6.2 Random Forest

Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional

space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane. There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane. We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

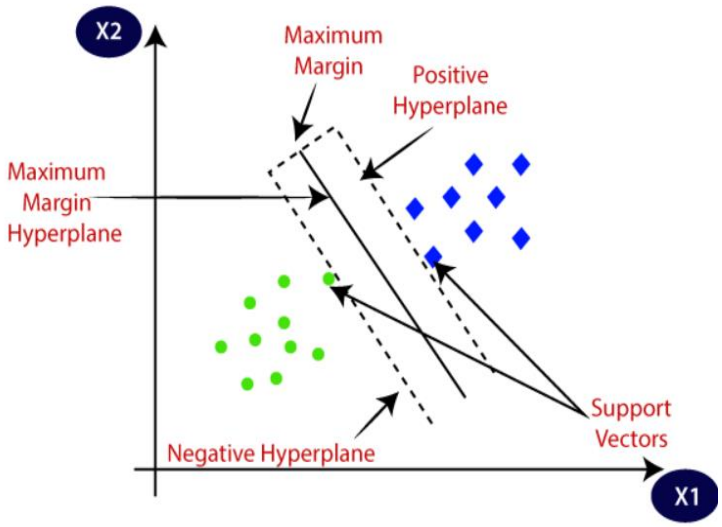


Fig 6.3 Support Vector Machine

K-Nearest Neighbor:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram.

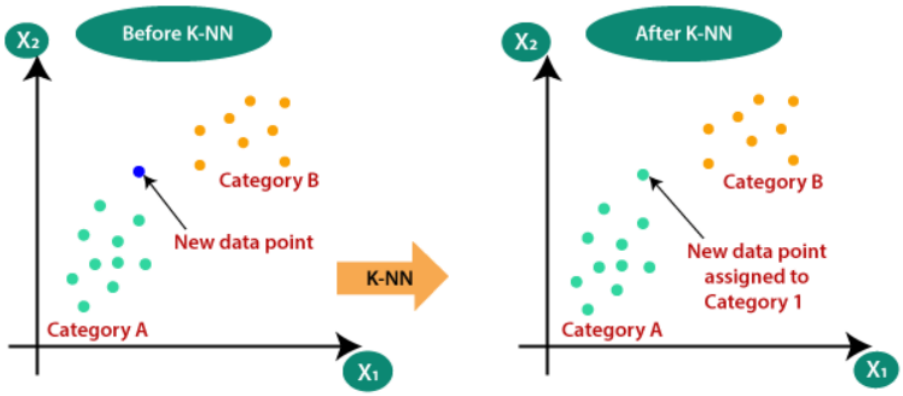


Fig 6.4 K-Nearest Neighbor

6.2 Selected Software:

Google Colab:

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members just the way you edit documents in Google Docs. Google Colab makes data science, deep learning, neural network, and machine learning accessible to individual researchers who can not afford costly computational infrastructure. Colab supports many popular machine learning libraries which can be easily loaded in your notebook. A Python program developer can use this notebook to write and execute random Python program codes just using a web browser.

We can use Google Colabs like Jupyter notebooks. They are really convenient because Google Colab hosts them, so we don't use any of our own computer resources to run the notebook. We can also share these notebooks so other people can easily run our code, all with a standard environment since it's not dependent on our own local machines. However, we might need to install some libraries in our environment during initialization.

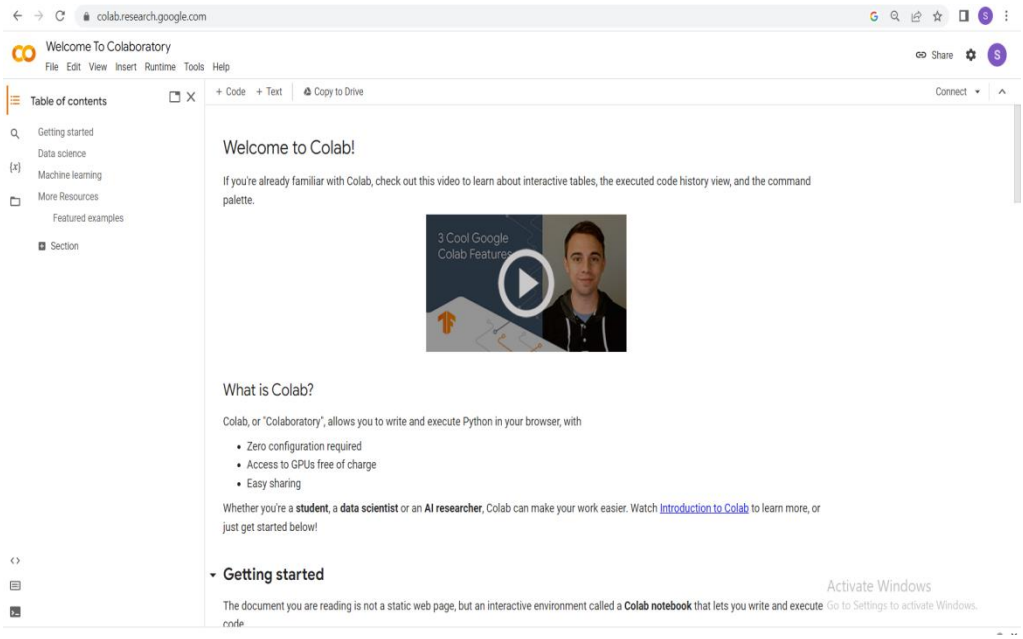


Fig 6.5 Google Colab

Creating a notebook in Colab:

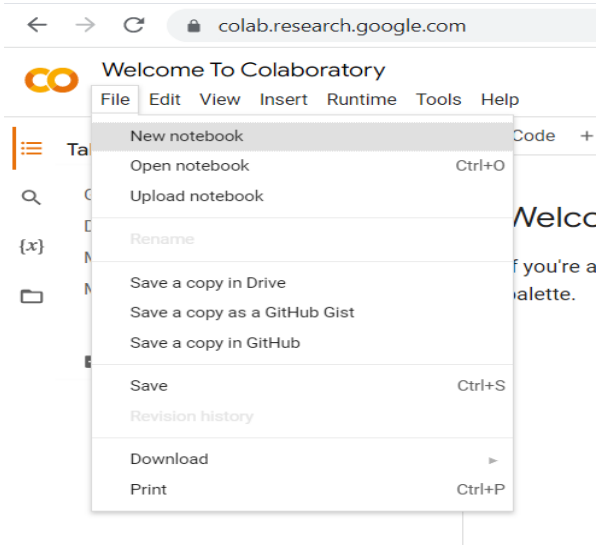


Fig 6.6 Google Colab Notebook

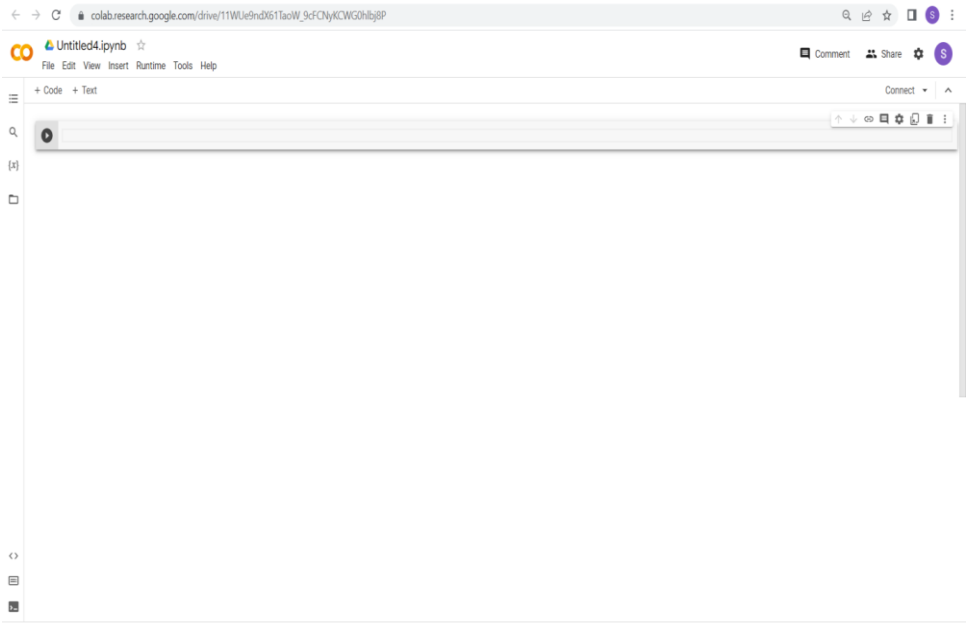


Fig 6.7 Google Colab Code Editor

Streamlit:

Streamlit is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app. If the source code of the streamlit's python script changes the app shows whether to rerun the application or not in the top-right corner. You can also select the 'Always rerun' option to rerun always when the source script changes.

Streamlit allows you to write an app the same way you write a python code. The streamlit has a distinctive data flow, any time something changes in your code or anything needs to be updated on the screen, streamlit reruns your python script entirely from the top to the bottom. This happens when the user interacts with the widgets like a select box or drop-down box or when the source code is changed. If you have some costly operations while rerunning your web app, like loading data from databases, you can use streamlit's st.cache method to cache those datasets, so that it loads faster.

6.3 Sample Code:

Streamlit code:

```
import streamlit as st
import pickle

dic1={'Gabriel Pereira':0, 'Mousinho da Silveria':1}
dic3={'Female':0, 'Male':1}
dic4={'Urban':1,'Rural':0}
dic5={'<=3':1,'>3':0}
dic6={'Living Together':1, 'Apart':0}
dic7={'None':0, 'Primary(upto 4th)':1,'High School(5th to 9th)':2,"10+2":3,'Higher
Education':4}
dic8={'Teacher':0, 'Health care realated':1,'Civil Services':2,'Other':3,'At Home':4}
```

```
dic9={'School Reputation':3,'Course preference':0,'Close to Home':4,'Other':2}
dic10={'Mother':1,'Father':0,'Other':2}
dic11={'<15min':1,'15-30min':2,'30-60min':3,'>60min':4}
dic12={'<2hrs':1,'2-5hrs':2,'5-10hrs':3,'>10hrs':4}
dic13={'0':0,'1':1,'2':2,'3':3,'more':4}
dic14={'True':1,'False':0}

st.title('Analysis and Prediction of Student Performace')

School = st.selectbox(
    'School',
    ('Gabriel Pereira', 'Mousinho da Silveria'))
st.write('You selected:', School)

Sex = st.selectbox(
    'Sex',
    ('Female', 'Male'))

#-------------------------- age (15-22)----------------------------

age=st.number_input("Age", min_value=15, max_value=22)

Address = st.selectbox(
    'Address',
    ('Urban', 'Rural'))

Family_Size = st.selectbox(
    'Family Size',
    ('<=3', '>3'))
```

```
Pstatus = st.selectbox(
    "Parent's Cohabitation Status",
    ('Living Together', 'Apart'))

st.write('You selected:', Pstatus)

Medu = st.selectbox(
    "Mother's Education",
    ('None', 'Primary(upto 4th)','Highschool(5th to 9th)','10+2','Higher Education'))

Fedu = st.selectbox(
    "Father's Education",
    ('None', 'Primary(upto 4th)','Highschool(5th to 9th)','10+2','Higher Education'))

Mjob = st.selectbox(
    'Mother Job',
    ('Teacher', 'Health care realated','Civil Services','Other','At Home'))


Fjob = st.selectbox(
    'Father Job',
    ('Teacher', 'Health care realated','Civil Services','At Home','Other'))

reson = st.selectbox(
    'Reson to Choose this School',
    ( 'School Reputation','Course preference','Close to Home','Other'))

guardian  = st.selectbox(
    "Student's Guardian",
    ('Mother','Father','Other'))
```

```
traveltime  = st.selectbox(
    "Travel time(from home to school)",
    ('<15min','15-30min','30-60min','>60min'))

studytime  = st.selectbox(
    "Weekly Study Time",
    ('<2hrs','2-5hrs','5-10hrs','>10hrs'))

failures  = st.selectbox(
    "No.of past class failures",
    ('0','1','2','3','more'))

schoolsup  = st.selectbox(
    "Extra Educational Support",
    ('True','False'))

famsup = st.selectbox(
    "Family Educational Support",
    ('True','False'))

paid  = st.selectbox(
    "Extra Paid Classes Within The Course Subject ",
    ('True','False'))

activities = st.selectbox(
    "Extra-Curricular Activities ",
    ('True','False'))

nursery = st.selectbox(
    "Attended Nursery School",
    ('True','False'))
```

```
higher= st.selectbox(
    "Wants To Take Higher Education",
    ('True','False'))

internet = st.selectbox(
    "Internet Access At Home ",
    ('True','False'))

romantic = st.selectbox(
    "With A Romantic Relationship ",
    ('True','False'))

#--------famrel-----------------
famrel=st.number_input("Quality Of Family Relationships", min_value=1, max_value=5)

#-------freetime----------------
freetime=st.number_input("Free Time After School", min_value=1, max_value=5)


#-------goout------------------
goout=st.number_input("Go Out With Friends", min_value=1, max_value=5)

#-------dalc-------------------
dalc=st.number_input("Workday Alcohol Consumption", min_value=1, max_value=5)

#-------walc-----------------
walc=st.number_input("Weekend Alcohol Consumption", min_value=1, max_value=5)
```

```
#------health----------
health=st.number_input("Current Health Status", min_value=1, max_value=5)

#------absent---------
absences=st.number_input("No.of School Absences",min_value=0, max_value=32)

Algo = st.selectbox(
    "Select Algorithms ",
    ('Decision Trees','Support Vector Machie','Randomforest','KNN'))

dic_algo={'Decision Trees':0,'Support Vector Machie':1,'Randomforest':2,'KNN':3}

counter=dic_algo[Algo]

dec=pickle.load(open("decisiontree.sav",'rb'))
svm=pickle.load(open("svm.sav",'rb'))
random=pickle.load(open("randomforest.sav",'rb'))
knn=pickle.load(open("knn.sav",'rb'))


if st.button('Predict'):
values=[dic1[School],dic3[Sex],age,dic4[Address],dic5[Family_Size],dic6[Pstatus],dic7[Me
du],dic7[Fedu],dic8[Mjob],dic8[Fjob],dic9[reson],dic10[guardian],dic11[traveltime],
dic12[studytime],dic13[failures],dic14[schoolsup],dic14[famsup],dic14[paid],dic14[activitie
s],dic14[nursery],dic14[higher],dic14[internet],dic14[romantic],famrel,freetime,goout,dalc,
walc,health,absences]

    pred1=dec.predict([values])
    pred2=svm.predict([values])
    pred3=random.predict([values])
    pred4=knn.predict([values])
```

```
    if counter==0:
        st.write("Decisointree:"+str(pred1[0]))

    if counter==1:
        st.write("Support Vector Machine:"+str(pred2[0]))

    if counter==2:
        st.write("Random Forest:"+str(pred3[0]))

    if counter==3:
        st.write("KNN:"+str(pred4[0]))

else:
    st.write('click above for prediction')
```

7. TESTING

7.1 Introduction

Software Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements. Software Testing is important because if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

Software testing is a process of identifying the correctness of software by considering its all attributes (Reliability, Scalability, Portability, Re-usability, Usability) and evaluating the execution of software components to find the software bugs or errors or defects. Software testing provides an independent view and objective of the software and gives surety of fitness of the software. It involves testing of all components under the required services to confirm that whether it is satisfying the specified requirements or not. The process is also providing the client with information about the quality of the software. Software Testing is Important because if there are any bugs or errors in the software, it can be identified early and can be solved before delivery of the software product. Properly tested software product ensures reliability, security and high performance which further results in time saving, cost effectiveness and customer satisfaction.

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, 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 tests. Each test type addresses a specific testing requirement. Testing is important because software bugs could be expensive or even dangerous. Software bugs can potentially cause monetary and human loss

Types of Testing:

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.

2. 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.

3. Regression Testing

Regression testing is a method of testing that is used to ensure that changes made to the software do not introduce new bugs or cause existing functionality to break. It is typically done after changes have been made to the code, such as bug fixes or new features, and is used to verify that the software still works as intended.

4. Performance Testing

It is designed to test the run-time performance of software within the context of an integrated system. It is used to test the speed and effectiveness of the program. It is also called load testing. In it we check, what is the performance of the system in the given load.

5. Object-Oriented Testing

This testing is a combination of various testing techniques that help to verify and validate object-oriented software. This testing is done in the following manner:

- Testing of Requirements
- Design and Analysis of Testing
- Testing of Code
- Integration testing
- System testing
- User Testing

6. Acceptance Testing

Acceptance testing is done by the customers to check whether the delivered products perform the desired tasks or not, as stated in requirements. We use this OOT, for discussing test plans and for executing the projects.

7. System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

8. 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.

9. 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.

7.2 Test Cases:

The test case is defined as a group of conditions under which a tester determines whether a software application is working as per the customer's requirements or not. Test case designing includes preconditions, case name, input conditions, and expected result. A test case is a first level action and derived from test scenarios. Test case gives detailed information about testing strategy, testing process, preconditions, and expected output. These are executed during the testing process to check whether the software application is performing the task for that it was developed or not. Detailed test case documentation works as a full proof guard for the testing team.

Test Case 1: Here, we are going to look test case-1 where we give the input by the user to check the performance of the student.

Table 7.1 Test Case 1

Test Case ID	TC#001
Description	When the user gives the data for prediction.
Test Step	Give the valid details of a student.
Expected Output	Student Performance will be High.
Actual Output	Student Performance will be High.
Result	Pass

Fig 7.1 Result of Test Case 1

Test Case 2: Here, we are going to look test case-2 where we give the input by the user to check the performance of the student.

Table 7.2 Test case 2

Test Case ID	TC#002
Description	When the user gives the data for prediction.
Test Step	Give the valid details of a student.
Expected Output	Student Performance will be Low.
Actual Output	Student Performance will be Low.
Result	Pass

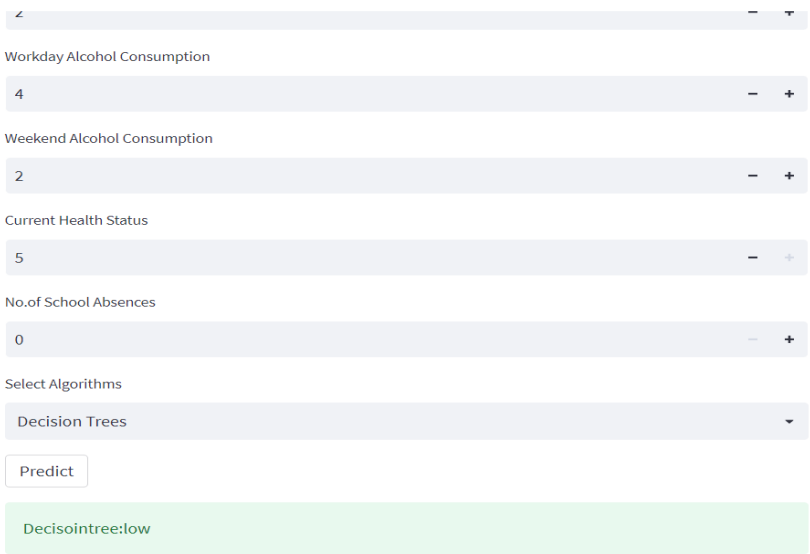


Fig 7.2 Result of Test Case 2

Test Case 3: Here, we are going to look test case-3 where we give the input by the user to check the performance of the student.

Table 7.3 Test case 3

Test Case ID	TC#003
Description	When the user gives the data for prediction.
Test Step	Give the valid details of a student.
Expected Output	Student Performance will be Medium.
Actual Output	Student Performance will be Medium.
Result	Pass

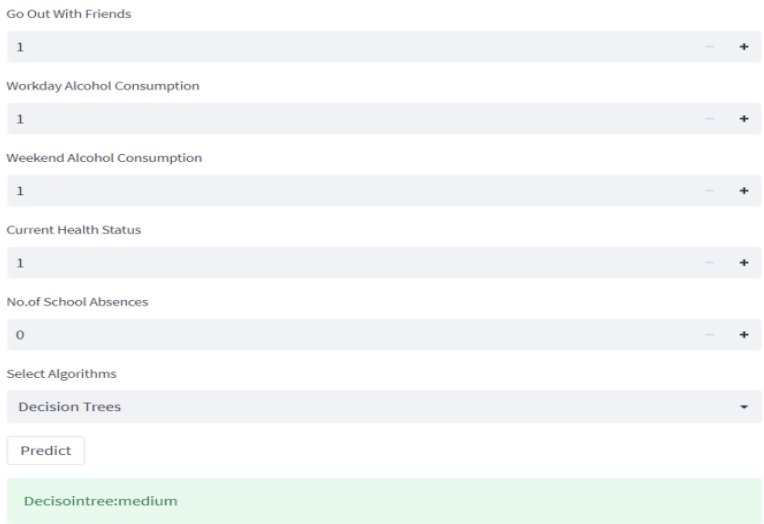


Fig 7.3 Result of Test Case 3

Test Case 4: Here, we are going to look test case-4 where we give the input by the user to check the performance of the student.

Table 7.4 Test case 4

Test Case ID	TC#004
Description	When the user gives the data for prediction.
Test Step	Give the valid details of a student.
Expected Output	Student Performance will be Low.
Actual Output	Student Performance will be Low.
Result	Pass

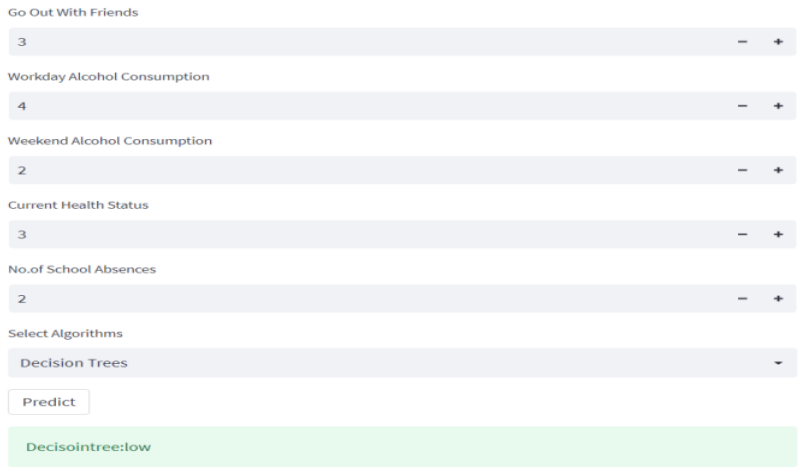


Fig 7.4 Result of Test Case 4

8. SCREENS & REPORTS

8.1 Introduction:

The screen displays the type of input the user has to be given to the system. When the user adds the data to the input screen called the webpage it will classify the attack based on the model. It will detect four types of attacks based on the machine learning algorithms. Designing the screen is most important part of our project. This widget is very customizable with a great design. A flat surface on which a picture or series of pictures is projected or reflected the surface on which the image appears in an electronic display (as in a television set, radar receiver, or computer terminal). The term screen design usually refers to either the process of determining the visual appearance and content of a single visual frame or the end result of that process. A term used to describe the viewing portion of a display.

8.2 Screenshots:

A screenshot, sometimes referred to as a screen cap or screengrab, is an image that shows the contents of a computer display. Screenshots let you capture exactly what you're seeing on your screen to share with others or reference later. Taking, saving, and sharing screenshots can be extremely helpful.

```
[38] import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     %matplotlib inline
     import seaborn as sns
```

```
[39] df=pd.read_csv('student_bulkdata.csv')
```

```
[40] df.head()
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	grade
0	MS	M	17	R	LE3	A	1	1	teacher	at_home	...	yes	no	1	1	2	5	1	1	17	medium
1	MS	F	21	U	LE3	T	2	1	services	teacher	...	no	yes	4	2	3	2	2	5	1	medium
2	GP	M	15	R	LE3	A	0	2	teacher	other	...	yes	yes	3	1	5	5	5	1	4	high
3	MS	F	15	R	LE3	T	1	4	services	other	...	no	no	4	4	1	5	4	3	7	medium
4	GP	F	18	R	LE3	A	4	0	other	teacher	...	yes	no	1	5	3	2	4	5	16	high

5 rows × 31 columns

```
[41] df.isnull().sum()

school        0
sex           0
age           0
```

```
df.isnull().sum()
```

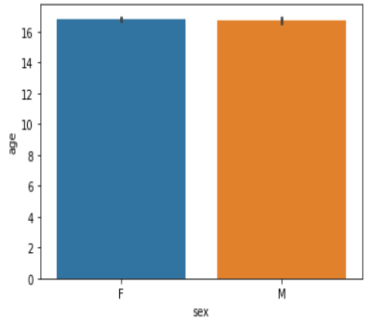
```
school          0
sex             0
age             0
address         0
famsize         0
Pstatus         0
Medu            0
Fedu            0
Mjob            0
Fjob            0
reason          0
guardian        0
traveltime      0
studytime       0
failures        0
schoolsup       0
famsup          0
paid            0
activities      0
nursery         0
higher          0
internet        0
romantic        0
famrel          0
freetime        0
goout           0
Dalc            0
Walc            0
health          0
absences        0
G1              0
G2              0
G3              0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 31 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      5000 non-null   int64
 1   sex         5000 non-null   int64
 2   age         5000 non-null   int64
 3   address     5000 non-null   int64
 4   famsize     5000 non-null   int64
 5   Pstatus     5000 non-null   int64
 6   Medu        5000 non-null   int64
 7   Fedu        5000 non-null   int64
 8   Mjob        5000 non-null   int64
 9   Fjob        5000 non-null   int64
 10  reson       5000 non-null   int64
 11  guardian    5000 non-null   int64
 12  traveltime  5000 non-null   int64
 13  studytime   5000 non-null   int64
 14  failures    5000 non-null   int64
 15  schoolsup   5000 non-null   int64
 16  famsup      5000 non-null   int64
 17  paid        5000 non-null   int64
 18  activities  5000 non-null   int64
 19  nursery     5000 non-null   int64
 20  higher      5000 non-null   int64
 21  internet    5000 non-null   int64
 22  romantic    5000 non-null   int64
 23  famrel      5000 non-null   int64
 24  freetime    5000 non-null   int64
 25  goout       5000 non-null   int64
 26  Dalc        5000 non-null   int64
 27  Walc        5000 non-null   int64
 28  health      5000 non-null   int64
 29  absences    5000 non-null   int64
```

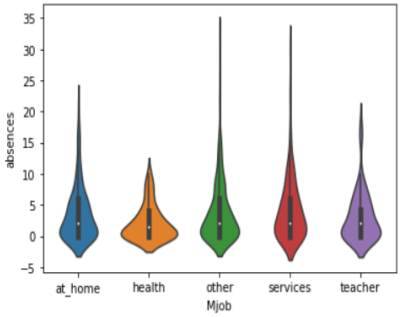
```
[7] sns.barplot(x='sex',y='age',data=df)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f8971fcaaf0>



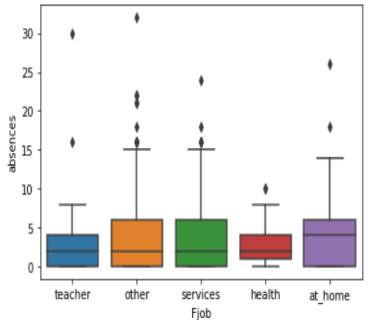

```
sns.violinplot(x='Mjob',y='absences',data=df)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f897198a850>




```
[8] sns.boxplot(x='Fjob',y='absences',data=df)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f8971f0a5b0>




```
[10] sns.barplot(x='Pstatus',y='age',data=df)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f896f916b20>

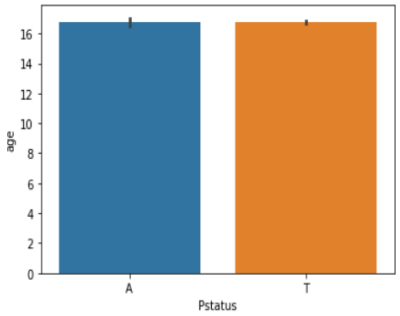


Fig 8.1 Visualization screenshots

```
[18] from sklearn.preprocessing import LabelEncoder

[19] df['school'].value_counts()

GP    423
MS    226
Name: school, dtype: int64

[20] df['school'] = df['school'].replace(('GP','MS'),(0,1))

[21] df['sex'].value_counts()

F    383
M    266
Name: sex, dtype: int64

[22] df['sex'] = df['sex'].replace(('F','M'),(0,1))

[23] df.keys()

Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'grades'],
      dtype='object')

[24] from sklearn.preprocessing import LabelEncoder
     le = LabelEncoder()
     for i in df.keys():
         if df[i].dtype == 'object' and i!='grades':
             df[i]=le.fit_transform(df[i])
```

```
[59] x = df.drop(['grade'],axis = 1)
     y = df['grade']

[60] from sklearn.model_selection import train_test_split
     x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=33, shuffle =True)

     from sklearn.tree import DecisionTreeClassifier

     import time
     start=time.time()


     model = DecisionTreeClassifier(criterion='gini',max_depth=10,random_state=33) #criterion can be entropy
     model.fit(x_train, y_train)
     train_dt_acc = model.score(x_train, y_train)

     print('model Train Score is : ' , model.score(x_train, y_train))

     y_test = model.predict(x_test)
     test_dt_acc = model.score(x_test, y_test)
     print('model test Score is : ' , model.score(x_test, y_test))

     end=time.time()
     time_dt=end-start

model Train Score is :  0.971
model test Score is :  1.0

     x.keys()

Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reson', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences'],
      dtype='object')
```

Fig 8.2 Decision Tree Algorithm Execution Screenshot

```
[87] model.predict([x.iloc[5]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names
  warnings.warn(
array(['low'], dtype=object)

[64] from sklearn.ensemble import RandomForestClassifier

     import time
     start=time.time()

     rf = RandomForestClassifier(criterion='gini',max_depth=10,random_state=33) #criterion can be entropy
     rf.fit(x_train, y_train)
     train_rf_acc = rf.score(x_train, y_train)
     print('model Train Score is : ' , rf.score(x_train, y_train))

     y_test = rf.predict(x_test)
     test_rf_acc = rf.score(x_test, y_test)
     print('model test Score is : ' , rf.score(x_test, y_test))

     end=time.time()
     time_rf=end-start

model Train Score is :  0.95525
model test Score is :  1.0

rf.predict([x.iloc[5]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature names
  warnings.warn(
array(['low'], dtype=object)
```

Fig 8.3 Random Forest Algorithm Execution Screenshot

```
[89] rf.predict([x.iloc[5]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature names
  warnings.warn(
array(['low'], dtype=object)

[66] from sklearn.svm import SVC

     import time
     start=time.time()

     sv = SVC()
     sv.fit(x_train, y_train)
     train_sv_acc = sv.score(x_train, y_train)
     print('model Train Score is : ' , sv.score(x_train, y_train))

     y_test = sv.predict(x_test)
     test_sv_acc = sv.score(x_test, y_test)
     print('model test Score is : ' , sv.score(x_test, y_test))

     end=time.time()
     time_sv=end-start

model Train Score is :  0.71325
model test Score is :  1.0

sv.predict([x.iloc[60]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but SVC was fitted with feature names
  warnings.warn(
array(['medium'], dtype=object)
```

Fig 8.4 Support Vector Machine Algorithm Execution Screenshot

```
[90] sv.predict([x.iloc[60]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but SVC was fitted with feature names
  warnings.warn(
array(['medium'], dtype=object)
```

```
[68] from sklearn.neighbors import KNeighborsClassifier

     import time
     start=time.time()

     knn = KNeighborsClassifier() #criterion can be entropy
     knn.fit(x_train, y_train)
     train_knn_acc = knn.score(x_train, y_train)
     print('model Train Score is : ' , knn.score(x_train, y_train))

     y_test = knn.predict(x_test)
     test_knn_acc = sv.score(x_test, y_test)
     print('model test Score is : ' , knn.score(x_test, y_test))

     end=time.time()
     time_knn=end-start

model Train Score is :  0.7755
model test Score is :  1.0
```

```
knn.predict([x.iloc[60]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but KNeighborsClassifier was fitted with feature names
  warnings.warn(
array(['medium'], dtype=object)
```

Fig 8.5 K-Nearest Neighbor Algorithm Execution Screenshot

8.3 Reports:

```
[91] knn.predict([x.iloc[60]])

/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but KNeighborsClassifier was fitted with feature names
  warnings.warn(
array(['medium'], dtype=object)
```

```
[69] data1 = pd.DataFrame({'Algorithm':['Decision Tree','Random Forest','SVM','KNN'],
                           'Accuracy':[train_dt_acc,train_rf_acc,train_sv_acc,train_knn_acc]})
```

```
[70] data2=pd.DataFrame({'Algorithm':['Decision Tree','Random Forest','SVM','KNN'],'ExecutionTime':[time_dt,time_rf,time_sv,time_knn]})
```

```
data1
```

	Algorithm	Accuracy
0	Decision Tree	0.97100
1	Random Forest	0.95525
2	SVM	0.71325
3	KNN	0.77550

```
data2
```

	Algorithm	ExecutionTime
0	Decision Tree	0.059526
1	Random Forest	0.728328
2	SVM	4.392275
3	KNN	1.529283

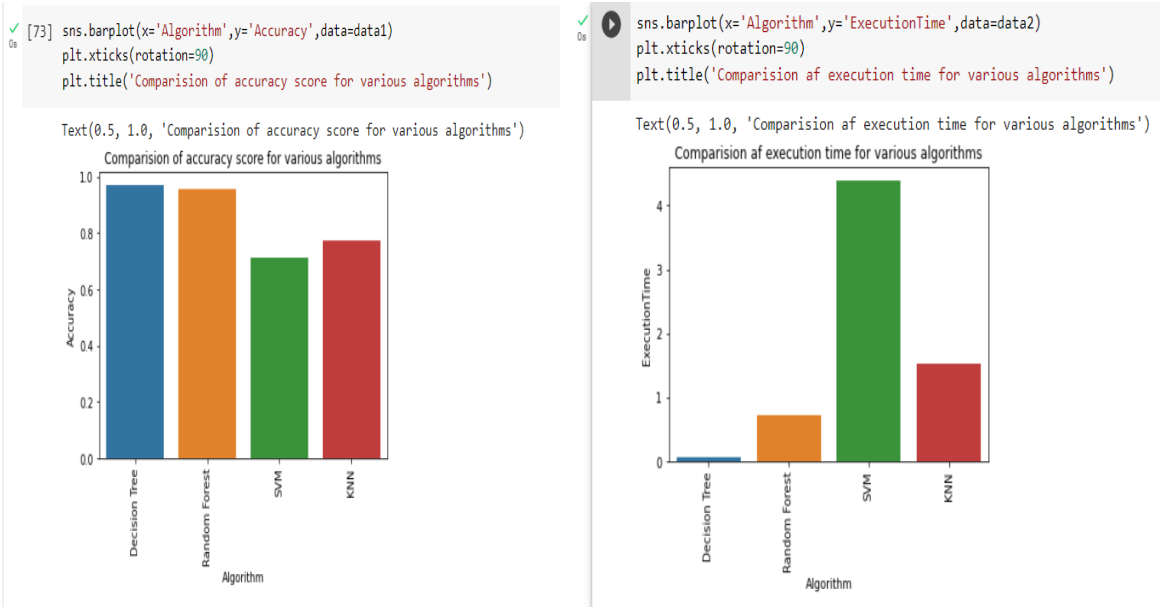
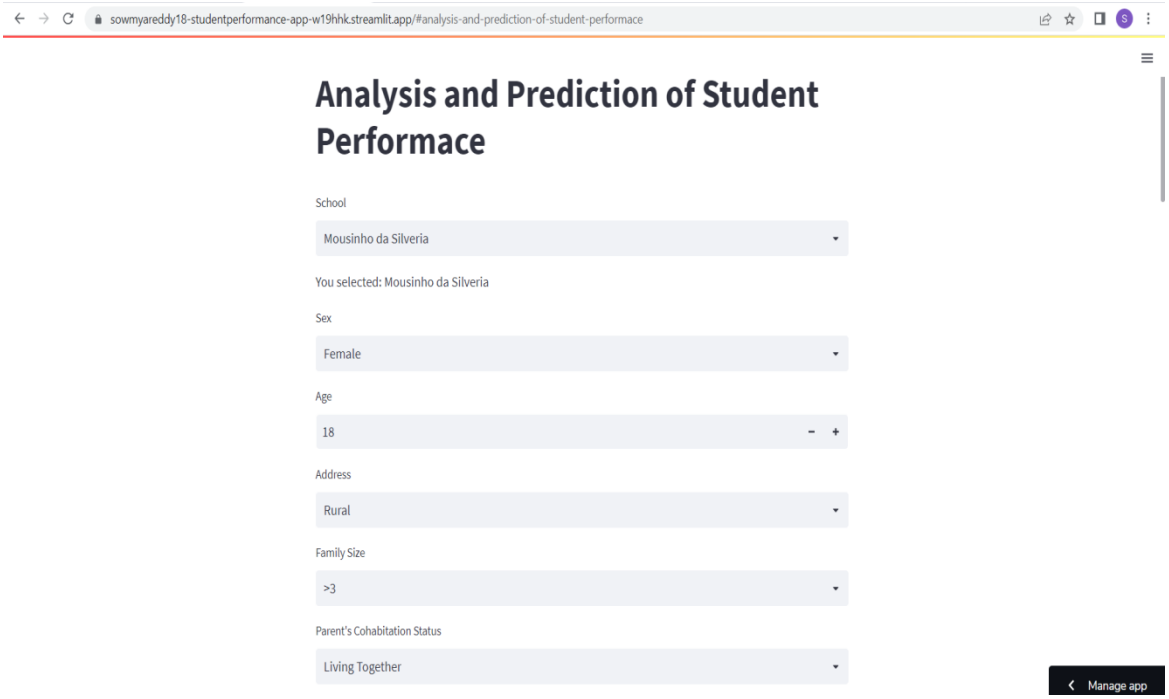


Fig 8.6 Comparison of Algorithms



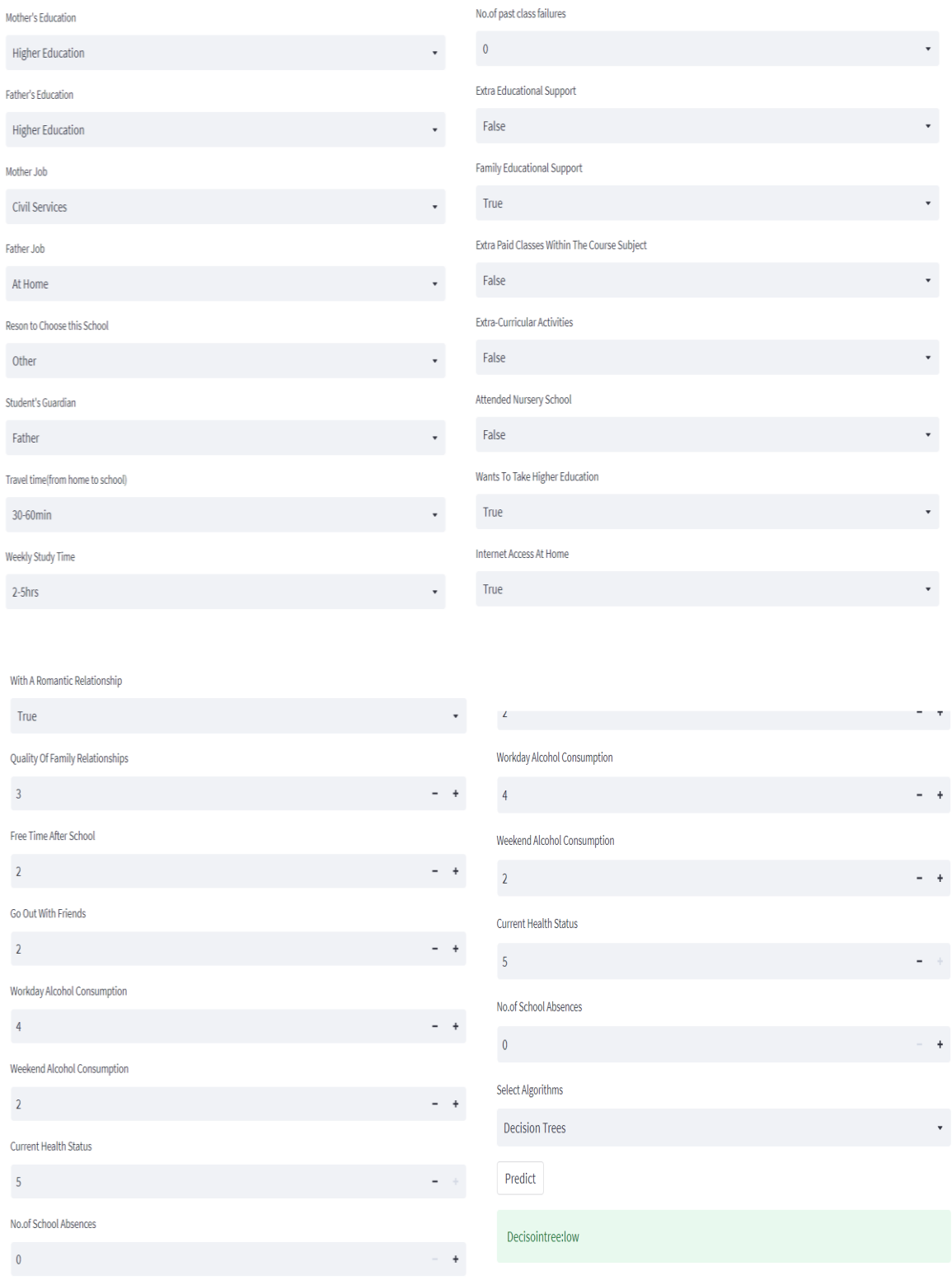


Fig 8.7 Output Report

9. CONCLUSION AND FUTURE SCOPE

9.1 Conclusion

The effectiveness of using machine learning in education field depends on the algorithm and the usage of the dataset. Choosing the algorithm to implement predicting the student performance is important. The accuracy of the result depends on the Machine learning algorithm that is being used. Present studies show that academic performances of the students are also dependent on student's background and other attributes. Many research works confirm that apart from the past academic performances student's background and other attributes indeed got a significant influence over students' performance. Machine learning has an emerging role in recent times in every sector, and it can also be used effectively in academics. In our project we have used machine learning algorithms- Decision Tree, Random Forest, Support Vector Machine and KNN classifier we can predict the student performance in academics. The proposed system provides the efficient algorithm for analysis.

9.2 Future Scope

The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject and student background information we will predict the marks and depending on the predicted marks the student can know in which subject he/she is lagging and teachers can easily monitor the student performance and can give the student a proper advice to improve the student performance. In the future the efficient and accurate algorithm should be estimated to analyze and predict the student performance and can be used in any particular college website to predict the performance of the students in that college.

10 BIBILOGRAPHY

References

[1] Boddeti Sravani and Myneni Madhu Bala, "**Prediction of Student Performance Using Linear Regression**". 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, pp.1-5, 2020.

[2] D K Arun, V Namratha, B V Ramyashree, Yashita P Jain, Antara Roy Choudhary, ''**Student Academic Performance Prediction Using Educational Data Mining**'' Int. J. Pure Appl. Math, vol. 118, no. 18, pp. 2703–2707, 2021.

[3] E. Alyahyan and D. Düştegör, ''**Predicting academic success in higher education: Literature review and best practices**'' Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, pp. 93-97, Dec. 2020.

[4] Manjari Chitti1 , Padmini Chitti 2 , Manoj Jayabalan: "**Need for Interpretable Student Performance Prediction.**" 2020 IEEE 13th International Conference on Developments in eSystems Engineering (DeSE), pp.269-272, 2020.

[5] E. Tatar and D. Düştegör, ''**Prediction of academic performance at undergraduate graduation: Course grades or grade point average**'' Appl. Sci., vol. 10, no. 14, pp. 1–15, 2020.

[6] Hartatik Kusrini Agung Budi Prasetio "**Prediction of Student Graduation with Naive Bayes Algorithm**" 2021 at 13:04:34 UTC from IEEE Xplore, pp. 103-111, 2021.

[7] S. Saeed, H. Gull, and S. Z. Iqbal, "**Web 2.0 Usage by Saudi Female Students for Information Sharing in Public Sector University a Pilot Study**," Int. J. Public Adm. 2017 Volume 4, Issue 3, July 2017, pp. 79–90.

[8] Dhanashree Mane, Pranali Namdas, Pooja Gargade, Dnyaneshwari Jagtap, S.S. Rathi" **Predicting student Performance Using Machine Learning Approach**". VJER Vishwakarma Journal of Engineering Research, pp. 1208-1212, December 2018.

[9] Wei Wang, Han Yu, Chunyan Miao" **Deep model for dropout prediction in moocs**". In ICCSE 2017, Proceedings of the 2nd International Conference on Crowd Science and Engineering, pp, 26–32, July 2017.

[10] B. Jeon, E. Shafran, L. Breitfeller, J. Levin, and C. P. Rose, "**Time-series Insights into the Process of Passing or Failing Online University Courses using Neural-Induced Interpretable Student States**", pp. 321-324, 2019.

[11] A. Barredo Arrieta et al., "**Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI**," Inf. Fusion, vol. 58, no. October 2019, pp. 82–115, 2019.

[12] Aprillia, Ristu S, Sari S, "**Determination of the Best Model in the Naive Bayes Classifier Method in Determining the Nutritional Status of Toddlers with Consider Parameter Independence**", JURNAL ITSMART Vol 4. No 1. Juni 2015 ISSN : 2301–7201, 2015.

ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Computer Science and Engineering

Academic Year: 2022-23

Project Title: **Analysis and Prediction of Students Performance using Machine Learning**

Type of Project: **Application Oriented/Design Oriented/Societal based/ Research Oriented/Industry Oriented**

Project Guide: Dr.P.UdayaKumar

Project Team: V.Sowmya Reddy (19A91A0564)

B.Sandhya (19A91A0503)

K.Yashwanth Reddy (19A91A0520)

ABSTRACT

The main objective of any educational institution is providing the best knowledge and education to the students. However, student to student academic performance may vary. A single teacher can't guide each and every student at once, due to this every year many students may lag behind due to lack of proper monitoring and advice. If universities can analyze the behavior of students and can predict the low performance factors earlier, then that information can help them to take effective steps in order to increase student performance. So, Student performance analysis and prediction is very important to understand about a student progress rate. So that this helps the teacher to know about the details of the student like which student needs which kind of help. Not only students and teachers, parents also can get the knowledge on student performance in academics. Predicting the performance of students became more difficult due to large amount of data in educational databases. A data mining research and prediction can be used in the education system for analyzing and predicting the student performance. Using educational machine learning algorithms Decision Tree, Random Forest, K-Nearest Neighbour, Support Vector Machine, we can predict the student performance in academics. The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject and student background information we will predict the final marks and depending on the predicted marks the student can know in which subject he/she is lagging and teachers can easily monitor the student performance and can give the student a proper advice to the student.

Signature of the Team members

COURSE OUTCOMES

PROJECT PART 1

Regulation: AR19	L	T	P	C
Course Code:191CS7P04	0	0	4	2

CO1	Identify a real life / engineering problem
CO2	Perform extensive investigation with prior knowledge
CO3	Interpret problem formulation and solution through critical thinking
CO4	Develop the work plan, schedule and estimate the cost
CO5	Identify the resources required to initiate project work

PROJECT PART 2

Regulation :AR19	L	T	P	C
Course Code:191CS8P05	0	0	14	7

CO6	Apply the domain knowledge to arrive at a framework to solve the problem
CO7	Design solution using research-based knowledge and modern tools and interpret the results
CO8	Assess the obtained solution in the context of engineering framework addressing the societal and environmental concerns adhering to professional ethics
CO9	Demonstrate communication skills effectively to work as a team, for guide interaction and presentations.
CO10	Prepare technical documentation/reports with effective written communication skills

CO-PO MAPPING

PROJECT PART-I 191CS7P04														
CO/PO	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PO12**	**PSO 1**	**PSO 2**
CO1	3	1								1		1	2	2
CO2	1	1	1								3	1	1	2
CO3		2				2	3					1	2	2
CO4	1	1									3	1	2	2
CO5	1	1									3	1	2	2

PROJECT PART-II191CS8P05														
CO/PO	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PO12**	**PSO 1**	**PSO 2**
CO6	1	3	2	2	1				2	2	2	3	2	1
CO7	1	3	2		1	1			2	2	2	2	2	1
CO8	1	3	3	2			1	1	2	2	2	2	2	1
CO9	1	3	2	1					2	2	2	2	2	1
CO10	1	3	2	1	1	1		1	1	2	1	1	2	1

CONCLUSION

The effectiveness of using machine learning in education field depends on the algorithm and the usage of the dataset. Choosing the algorithm to implement predicting the student performance is important. The accuracy of the result depends on the Machine learning algorithm that is being used. Present studies show that academic performances of the students are also dependent on student's background and other attributes. Many research works confirm that apart from the past academic performances student's background and other attributes indeed got a significant influence over students' performance. Machine learning has an emerging role in recent times in every sector, and it can also be used effectively in academics. In our project we have used machine learning algorithms- Decision Tree, Random Forest, Support Vector Machine and KNN classifier we can predict the student performance in academics. The proposed system provides the efficient algorithm for analysis.

FUTURE SCOPE

The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject and student's background information we will predict the marks and depending on the predicted marks the student can know in which subject he/she is lagging and teachers can easily monitor the student performance and can give the student a proper advice to improve the student performance. In the future the efficient and accurate algorithm should be estimated to analyze and predict the student performance and can be used in any particular college website to predict the performance of the students in that college.

ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Computer Science and Engineering

Academic Year: 2022-23

Project Title: **Analysis and Prediction of Students Performance using Machine Learning**

Type of Project: **Application Oriented/Design Oriented/Societal based/ Research Oriented/Industry Oriented**

Project Guide: Dr.P.UdayaKumar

PROGRAM OUTCOMES			
PO1	Engineering knowledge	PO7	Environment and sustainability
PO2	Problem analysis	PO8	Ethics
PO3	Design/development of solutions	PO9	Individual and team work
PO4	Conduct investigations of complex problems	PO10	Communication
PO5	Modern tool usage	PO11	Project management and finance
PO6	The engineer and society	PO12	Lifelong learning

PROGRAM SPECIFIC OUTCOMES	
PSO1	Develop efficient solutions to the real world problems using the domains of algorithms, networks, database management and latest programming tools and techniques.
PSO2	Provide the data centric business solutions through emerging areas like IoT ,AI ,data analytics and Block chain technologies.

CONCLUSION STATEMENTS

S.No	Description	Attained COs	Attained POs/PSOs
1	Explained about the importance of analysis and prediction of student performance in their academics.	CO1	PO1
2	Different Machine Learning algorithms that are used to predict the student's performance were discussed.	CO1,CO2	PO1,PO3 PO5
3	Researched about ways to predict the student's performance prediction	CO3	PO2, PO7, PSO1
4	Literature was reviewed from 2014 to 2021 by various authors on student's performance prediction.	CO2	PO2, PO7 PO12
5	Experimental procedure to use Machine Learning algorithms such as Random Forest, Decision tree, Logistic Regression, SVM, KNN are used to predict the student performance.	CO2,CO3	PO2,PO3, PO5 PO9, PSO1
6	Compared this method with existing methods.	CO2	PO4,PO7
7	Recorded accuracy results while working with different machine learning algorithms with student dataset.	CO5	PO2, PO4, PO9
8	Concluded the findings and analyzed the results to present it.	CO2,CO4 CO5	PO9,PO10 PO11, PSO2

Signature of the Team members

V.Sowmya Reddy (19A91A0564)

B. Sandhya (19A91A0503)

K.Yashwanth Reddy (19A91A0520)

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 436-442

e-ISSN : 2583-1062

Impact Factor : 5.725

ANALYSIS AND PREDICTION OF STUDENTS PERFORMANCE USING MACHINE LEARNING

Velagala Sowmya Reddy[1], Barri Sandhya[2], K. Yashwanth Reddy[3], Dr. P. Udaya Kumar[4]

[1,2,3]Cse, B. Tech Aditya Engineering College, Surampalem

[4]Guide, M. Tech, Ph.D, Aditya Engineering College, Surampalem

ABSTRACT

Any educational institution's primary goal is providing the best knowledge and education to the students. However, student to student academic performance may vary. A single teacher can't guide each and every student at once, due to this every year many students may lag behind due to lack of proper monitoring and advice. If universities can analyze the behavior of students and can predict the low performance factors earlier, then that information can help them to take effective steps in order to increase student performance. So, Student performance analysis and prediction is very important to understand about a student progress rate. So that this helps the teacher to know about the details of the student like which student needs which kind of help. Not only students and teachers, parents also can get the knowledge on student performance in academics. Due to the abundance of data in educational databases, it is now increasingly challenging to predict students' success.. A data mining research and prediction can be used in the education system for analyzing and predicting the student performance. Using educational machine learning techniques Support Vector Machine, K-Nearest Neighbour, Decision Tree, and Random Forest The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject and student background details we will predict the marks and depending on the predicted marks the student can know in which subject he/she is lagging and teachers can easily monitor the student performance and can give the student a proper advice to the student.

Keywords–Student Performance Prediction, Machine Learning, Decision Tree, Random Forest, K-Nearest Neighbour, Support Vector Machine.

1. INTRODUCTION

A subset of artificial intelligence called "machine learning" is used to build and create algorithms that let computers predict behaviour based on incoming data. The majority of machine learning algorithms emphasise decision-making and pattern identification. Machine learning is wherever data is being used and analyzed. It is a very useful tool for exposing secret patterns and important information that can otherwise be sought and comprehended with statistical methods. In the field of education, the use of machine learning algorithms has increased significantly in recent years.. Student performance analysis and prediction is very important to understand about a student progress rate. So that it helps the teacher to know about the student like which student needs which kind of help.If universities can analyze the the behavior of students and can predict low performance factors earlier, this information can help them to take effective steps to increase student performance. So, Student performance analysis and prediction is very important to understand a student progress rate. So that it helps the teacher to know about the student like which student needs which kind of help. Not only students and teachers, parents also can get the knowledge on student performance in academics. A data mining research and prediction can be used in the education system for analyzing and predicting the student performance. We can forecast student academic achievement using the educational machine learning algorithms Decision Tree, Random Forest, Support Vector Machine, and KNN classifier. The proposed system provides the efficient algorithm for analysis.

2. LITERATURE REVIEW

[1] PREDICTION OF STUDENT PERFORMANCE USING LINEAR REGRESSION-They used the linear regression approach in this paper. In this case study, supervised learning and more specifically, predictive analysis are the main topics. The ability to predict student academic success is crucial since it informs both students and teachers of those who may struggle in a course and those who need to boost their grade point average. The dataset used in this paper comprised of 100 students, which is a small dataset. Student datasets are the study's primary source of data. The dataset used for this study is tabular in nature and includes data on students' age, gender, academic performance, and health. In this case study, the implementation of linear regression to forecast the student's academic performance while taking into account the student's dataset was the main focus. [1].

[2] STUDENT ACADEMIC PERFORMANCE PREDICTION USING EDUCATIONAL DATA MINING-This study examines educational data mining, which focuses more on making predictions than on producing precise results

for future predictions. A regular analysis must be carried out in the educational field in order to keep track of the changes that are being in class patterns. In this case study exploratory design, which was made possible by gathering data sets from BMS College of Engineering, the initial findings of an educational data mining forecast are provided. Numerous classification and regression methods were examined and tried for improved accuracy in subject-by-subject analysis and GPA prediction of a student's performance. Voting on the classification algorithms was carried out for the prior ensemble system. These were also divided into other groups and contrasted according to specific criteria; the outcomes are discussed in this paper. [2].

[3] PREDICTING ACADEMIC SUCCESS IN HIGHER EDUCATION: LITERATURE REVIEW AND BEST PRACTICES-This paper primarily focuses on student achievement, which is crucial for educational institutions and usually used as a performance indicator. Prior to the identification of students who may be in danger, preventative measures can increase student achievement. Machine learning techniques have recently been applied to prediction. Although there are many success stories in the literature, "computer science" or more typically "artificial intelligence" educated educators tend to use these strategies the most. Many judgements are made possible by the use of data mining that is successful and efficient techniques, including how to define student achievement, which student characteristics to emphasise, and even which machine learning technique is best for the particular issue at hand. This study made data mining techniques more accessible to educators, allowing the full potential of their application to the given challenge in the field of education. [3].

3. PROPOSED METHODOLOGY

Students are essential for producing high quality graduates who excel in academics, and creative thought for any school, college and other educational institution. In order to get the quality students it is necessary to evaluate the output of students in order to forecast student achievement. Data abundance in educational databases has made it increasingly difficult to predict students' success. We can more effectively and efficiently increase student performance and growth by using educational machine learning algorithms.The proposed system uses machine learning methods Decision Tree, Random Forest, Support Vector Machine and KNN classifier are used to predict the student performance that is used by students, teachers and parents to get the knowledge on student performance in academics. The proposed system provides the efficient algorithm for analysis. Based on the previous marks of the student in each subject we will predict the marks and depending on the predicted marks of the student in each subject the teachers can easily monitor the student performance. This project provides a system that can analyze and predict the student performance accurately by using machine learning algorithmsThe flowchart of the proposed system describes about the flow of the project. First the dataset is collected from Kaggle and the set is preprocessed which includes data cleaning and data is visualized to understand the dataset and then, the data set is split into a training and a testing dataset, and finally, the algorithms are put into practise to make predictions. The flow of the project is clearly shown in the below figure.

Fig 3.1 Flow Chart of the Project

Hardware Requirements:

- System : Pentium i3 Processor
- Hard Disk : 500 GB
- Ram : 2 GB

Software Requirements:

- Operating system : Windows 10
- Coding Language : Python
- IDE : Google colab or Jupyter Notebook
- Streamlit
- Students Dataset

4. ALGORITHMS USED

Decision Tree:

A supervised learning approach called a decision tree can be used to solve classification and regression problems, however it is typically chosen for classification problems. Decision trees are always used to visualise all potential outcomes and decision points that take place chronologically and are produced in both situations in the same manner. In the world of finance, decision trees are most frequently utilised in areas like budgeting, portfolio management, and loan approval

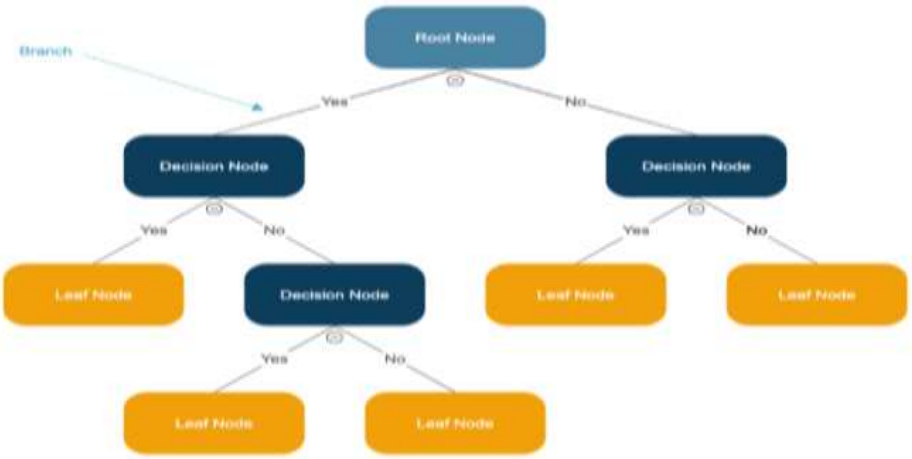


Fig 4.1 Decision Tree

Random Forest:

A very well known supervised machine learning approach called the Random Forest approach is utilised to solve classification and regression issues. We are aware that a forest has several trees, and the more trees there are, the more effective it will be. Similar to this, the accuracy and problem-solving capacity of a Random Forest Algorithm increase with the number of trees in the algorithm.

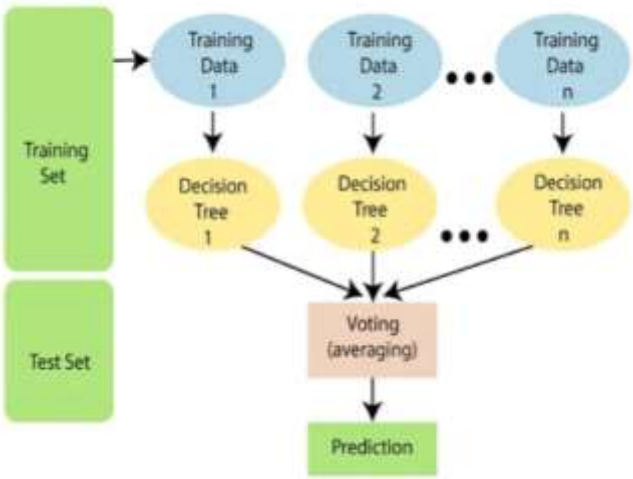


Fig 4.2 Random Forest

Support Vector Machine:

The SVM algorithm's objective is to determine the optimum line or decision boundary that can divide n-dimensional space into classes so that subsequent data points can be quickly assigned to the appropriate category. A hyperplane is the name given to this optimal decision boundary. The dataset's features determine the hyperplane's dimensions, therefore if there are only two characteristics, the hyperplane will be a straight line. Additionally, if there are three features, the hyperplane will only have two dimensions

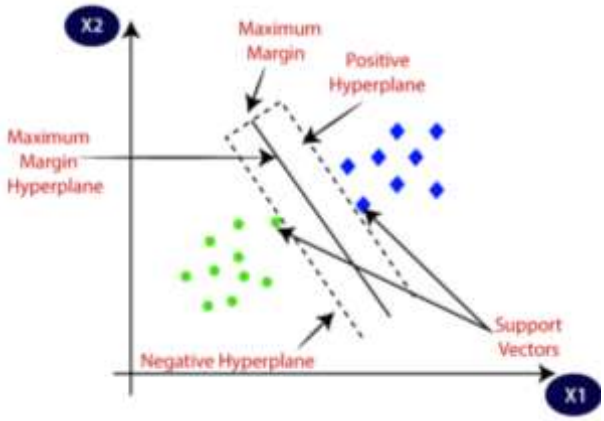


Fig 4.3 Support Vector Machine

K-Nearest Neighbor:

One of the simplest machine learning algorithms used in supervised learning techniques is K-Nearest Neighbour. The new case is placed in the category that is most similar to the existing categories after the K-NN algorithm analyses the similarity between the new data and existing cases. To put it simply, the K-NN algorithm collects all the data that is available and categorises new data based on similarities

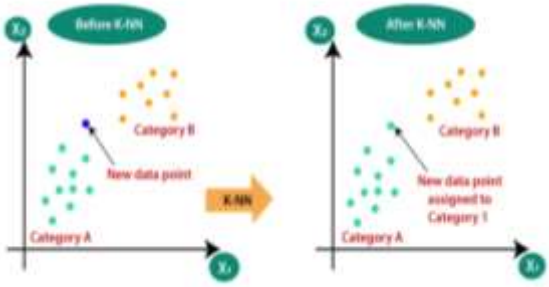


Fig 4.4 K-Nearest Neighbor

5. EXPERIMENTAL RESULTS

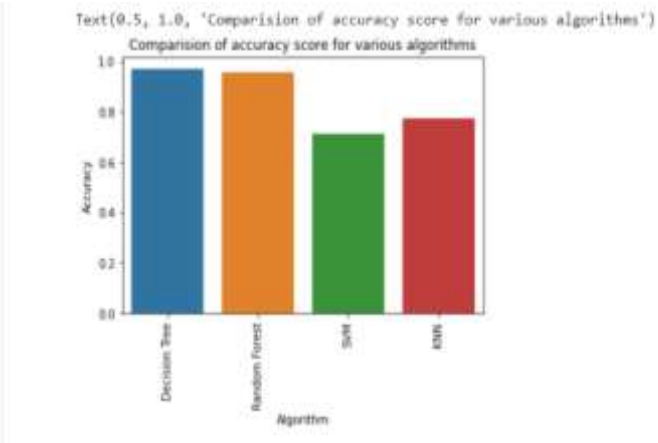


Fig 5.1 Comparison of Accuracies

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 436-442

e-ISSN : 2583-1062

Impact Factor : 5.725

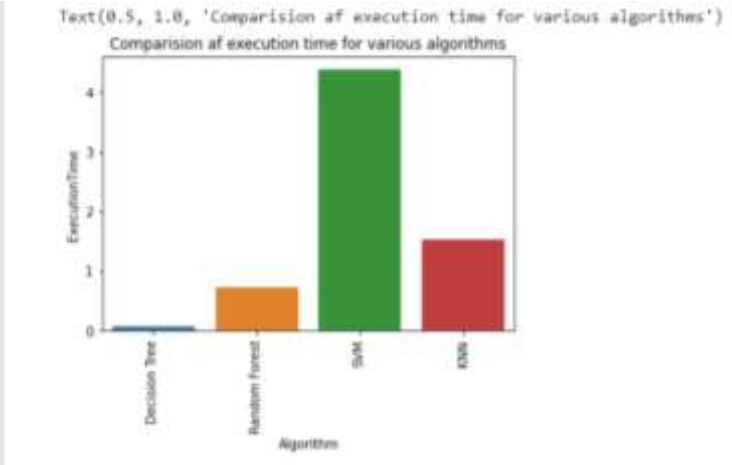


Fig 5.2 Comparison of Execution Times

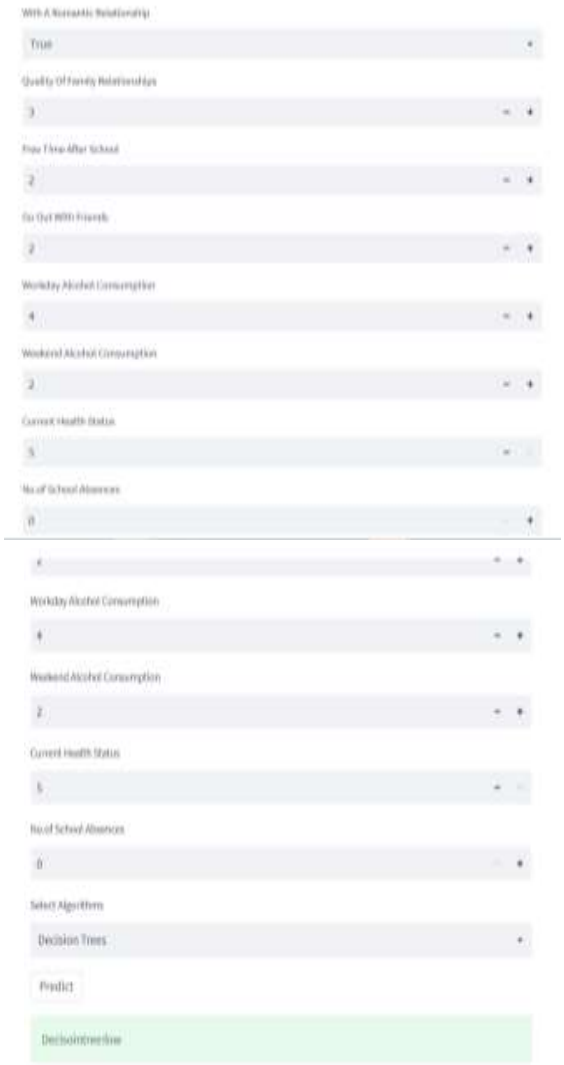
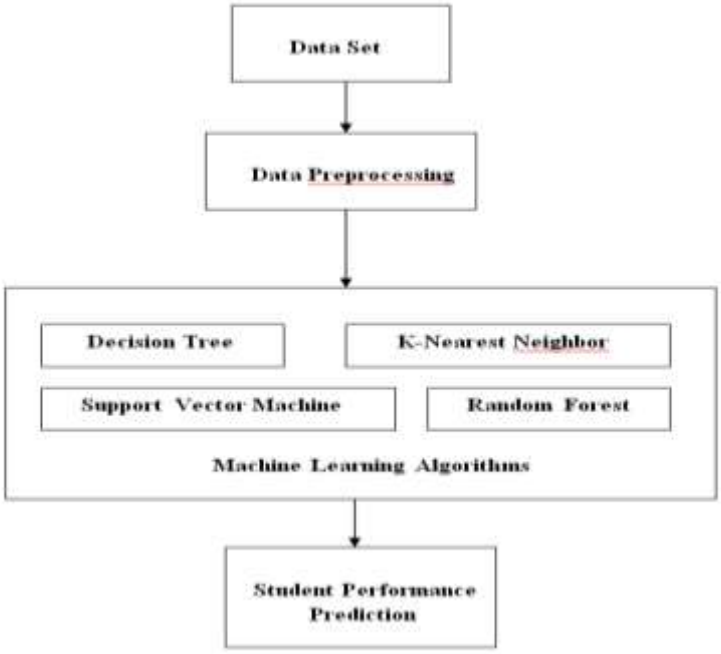


Fig 5.3 Output Screenshots

6. BLOCK DIAGRAM



7. ADVANTAGES

- The proposed system produces the results with high Accuracy
- It takes less time for training
- This system predicts better results

8. CONCLUSION AND FUTURE SCOPE

Any machine learning algorithm's efficacy in the field of education depends on the algorithm's utilisation of the dataset and the dataset itself. It's crucial to carefully select the algorithm to use when predicting student success. The machine learning method being utilised affects how accurate the forecast is. Current case studies demonstrate that a student's history and other characteristics have an impact on their academic achievement. Numerous studies support the idea that, in addition to prior academic achievements, a student's history and other characteristics may also have a substantial impact on how well they do. Data mining in education can benefit from the application of machine learning, which has been increasingly important in all industries lately. To predict student academic success in our project, we used the machine learning algorithms Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbour. The efficient algorithm for analysis and prediction is provided by the proposed system.The suggested system offers an effective analytical algorithm. Based on the previous marks of the student in each subject and some background information we will predict the marks and depending on the predicted marks the student can know in which subject he/she is lagging and teachers can easily monitor the student performance and can provide the learner with appropriate recommendations to enhance performance. Future analyses and predictions of student success should be made using an efficient and accurate algorithm, which can be used to any college website to forecast student performance there.

9. REFERENCES

[1] Boddeti Sravani And Myneni Madhu Bala, “PREDICTION OF STUDENT PERFORMANCE USING LINEAR REGRESSION”. 2020 International Conference For Emerging Technology (INCET) Belgaum, India. Jun 5-7, Pp.1-5, 2020.

[2] D K Arun, V Namratha, B V Ramyashree, Yashita P Jain And Antara Roy Choudhary, ‘‘STUDENT ACADEMIC PERFORMANCE PREDICTION USING EDUCATIONAL DATA MINING’’ Int. J. Pure Appl. Math, Vol. 118, No. 18, Pp. 2703–2707, 2021.

[3] E. Alyahyan And D. Düştegör, ‘‘PREDICTING ACADEMIC SUCCESS IN HIGHER EDUCATION: LITERATURE REVIEW AND BEST PRACTICES’’ Int. J. Educ. Technol. Higher Educ., Vol. 17, No. 1, Pp. 93-97, Dec. 2020.

[4] Manjari Chitti1 , Padmini Chitti 2 And Manoj Jayabalan: “NEED FOR INTERPRETABLE STUDENT PERFORMANCE PREDICTION.” 2020 IEEE 13th International Conference On Developments In Esystems Engineering (Dese), Pp.269-272, 2020.

[5] S. Saeed, H. Gull, And S. Z. Iqbal, “WEB 2.0 USAGE BY SAUDI FEMALE STUDENTS FOR INFORMATION SHARING IN PUBLIC SECTOR UNIVERSITY A PILOT STUDY” Int. J. Public Adm. 2017 Volume 4, Issue 3, July 2017, Pp. 79–90.

[6] Dhanashree Mane, Pranali Namdas, Pooja Gargade, Dnyaneshwari Jagtap And S.S. Rathi ”PREDICTING STUDENT PERFORMANCE USING MACHINE LEARNING APPROACH”. VJER Vishwakarma Journal Of Engineering Research, Pp. 1208-1212, December 2018.

[7] Wei Wang, Han Yu And Chunyan Miao ”DEEP MODEL FOR DROPOUT PREDICTION IN MOOCS”. In ICCSE 2017, Proceedings Of The 2nd International Conference On Crowd Science And Engineering, Pp, 26–32, July 2017.

[8] A. Barredo Arrieta Et Al., “Explainable Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities And Challenges Toward Responsible AI,” Inf. Fusion, Vol. 58, Pp. 82–115, 2019.