

MISSION MARKS

MACHINE LEARNING PROJECT

A

Major Project Report

Submitted by:

Agnim Seth(0887CS161001)

Apoorv Mishra(0887CS161006)

Sourabh Baghel(0887CS161030)

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Under the Guidance

of

Mr.Anurag Kumar

(Asst.Professor, CSE)



DR. APJ ABDUL KALAM UNIVERSITY INSTITUTE OF TECHNOLOGY

JHABUA, M.P. (INDIA) - 457661

(AFFILIATED TO RGPV, BHOPAL, M.P. (INDIA))

JUNE 2020

DECLARATION

We Agnim Seth(0887CS161001),Apoorv Mishra(0887CS161006),Sourabh Baghel(0887CS161030), hereby declare that this project work entitled “Mission Marks: An Machine Learning Project” was carried out by us under the supervision of Prof.Anurag Kumar, AP, Department of CSE, Dr. A.P.J Abdul Kalam University Institute Of Technology, Jhabua (M.P.). This project work is submitted to the Department of Computer Science and Engineering during the academic year 2019-2020.

Place:

Date:

Name

Signature

Agnim Seth(0887CS161001)

Apoorv Mishra(0887CS161006)

Sourabh Baghel(0887CS161030)

DR. APJ ABDUL KALAM
UNIVERSITY INSTITUTE OF TECHNOLOGY,
JHABUA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



SESSION 2019-2020

CERTIFICATE OF APPROVAL

This is to certify that the work embodies in this report entitled “***Mission Marks: An Machine Learning Project*** ” being submitted by *Agnim Seth* (0887CS161001), *Apoorv Mishra* (0887CS161006), *Sourabh Baghel* (0887CS161030) carried out the project work under our supervision and guidance in the “**Department of Computer Science & Engineering**”, **Dr. APJ Abdul Kalam University Institute Of Technology, Jhabua (M.P.)**.

(Internal Examiner)

(External Examiner)

Name:

Name:

Designation: AP

Designation:

Institute: Dr. APJ Abdul Kalam UIT, Jhabua

Institute

Date:

Date:

DR. APJ ABDUL KALAM
UNIVERSITY INSTITUTE OF TECHNOLOGY,
JHABUA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



SESSION 2019-2020

CERTIFICATE

This is to certify that the work embodies in this report entitled “***Mission Marks: An Machine Learning Project*** ” being submitted by *Agnim seth* (0887CS161001), *Apoorv Mishra* (0887CS161006) *Sourabh Baghel* (0887CS161030) carried out the project work under my supervision and guidance in the “**Department of Computer Science & Engineering**”, Dr. APJ Abdul Kalam UIT, Jhabua (M.P.).

Guided & Approved By:

Prof. Anurag Kumar

Asst. Prof.

Department of CSE

Dr. APJ Abdul Kalam UIT, Jhabua

Forwarded By:

Prof. Vaishali Ahirwar

HOD

Department of CSE

Dr. APJ Abdul Kalam UIT, Jhabua

ACKNOWLEDGEMENT

Knowledge is an expression of experience gained in life. It is the choicest possession that should be happily shared with others.

In this regard We feel great pleasure in submitting this major project report on “*Mission Marks: An Machine Learning Project*”. It gives me immense pleasure to express my deepest sense of gratitude and sincere thanks to my highly respected and esteemed guide **Prof. Anurag Kumar** (Dept. of CSE), Dr. APJ Abdul Kalam UIT, Jhabua for their valuable guidance, encouragement and help for completing this work. Their useful suggestions for this whole work and cooperative behavior are sincerely acknowledged. During this project, we received a lot of help, advice and co-operation from our esteemed faculty and other distinguished persons. We would also like to thank **Prof. Vaishali Ahirwar** (H.O.D. Dept. of CSE) for their valuable guidance through the course of the project without whose encouragement the project wouldn't have been a success.

We are grateful to our Principal Dr. **K.S. Chandel** and college authorities for their support and all those who have directly or indirectly helped us during the project work.

At the end I would like to express my sincere thanks to all my friends and others who helped me directly or indirectly during this project work.

Abstract

The ultimate goal of any educational institution is offering the best educational experience and knowledge to the students. Identifying the students who need extra support and taking the appropriate actions to enhance their performance plays an important role in achieving that goal.

Machine Learning is a cornerstone when it comes to artificial intelligence and big data analysis. It provides powerful algorithms that are capable of recognizing patterns, classifying data, and, basically, learn by themselves to perform a specific task. This field has incredibly grown in popularity these days, however, it still remains unknown for the majority of people, and even for most professionals.

In this project, a model is proposed to predict the performance of students in an academic organization. The algorithm employed is a machine learning technique called Multiple Linear Regression. Further, the importance of several different attributes, or "features" is considered, in order to determine which of these are correlated with student performance. Finally, the results of an experiment follow, showcasing the power of machine learning in such an application.

The main aim of this project is also to prove the possibility of training and modeling a small dataset size and the feasibility of creating a prediction model with credible accuracy rate. This project explores as well the possibility of identifying the key indicators in the small dataset, which will be utilized in creating the prediction model, using analysis, visualization and algorithms. Best indicators were fed into multiple linear regression machine learning algorithms to evaluate them for the most accurate model.

Keywords: Machine learning, Multiple linear regression, visualization, Prediction

Table of contents

| | |
|--|-----|
| Title page | i |
| Declaration | ii |
| Certificate of approval | iii |
| Certificate | iv |
| Acknowledgements | v |
| Abstract | vi |
| Table of contents | vii |
| List of Figures | ix |
| List of Tables | x |
| 1. Introduction | 1 |
| 2. Machine Learning (ML) | 2 |
| 2.1. A little bit of history | 2 |
| 2.2. What is Machine Learning? | 4 |
| 2.3. Types of ML | 7 |
| 2.4. State of the art | 7 |
| 3. Literature Survey | 11 |
| 4. Mission Marks | 12 |
| 4.1. Description of the project | 12 |
| 4.1.1. How is the plan going to be solved? | 12 |
| 4.1.2. workflow | 13 |
| 4.2. Dataset Analysis | 14 |
| 4.2.1. Dataset: by head() Function | 14 |
| 4.2.2. Info of Dataset | 15 |
| 4.2.3. Description of Dataset | 15 |
| 4.3. Data Visualization | 16 |
| 4.3.1. Correlation Matrix of Dataset | 16 |
| 4.3.2. Graphs (columns) | 17 |
| 4.3.3. Scatter And Density Plot | 19 |

| | |
|--|----|
| 4.4. Dataset split | 20 |
| 4.5. System Design | 21 |
| 5. Implementation of Machine Learning Algorithms | 22 |
| 6. Mission Marks Machine Learning Prediction on new Data(snapshots): | 24 |
| 7. Discussion and Conclusions | 27 |
| 8. Future work | 29 |
| References | 30 |
| Appendices | 31 |
| Glossary | 32 |

List of Figures

| | |
|---------------------------------------|----|
| Figure 1. Supervised Learning model | 5 |
| Figure 2. Unsupervised Learning model | 6 |
| Figure 3. Workflow of model | 13 |
| Figure 4. Correlation matrix | 16 |
| Figure 5. Graph MS-I | 17 |
| Figure 6. Graph MS-II | 17 |
| Figure 7. Graph Internal | 17 |
| Figure 8. Graph External | 18 |
| Figure 9. Graph EndSem | 18 |
| Figure 10. Graph Dataset | 18 |
| Figure 11. Scatter and Density plot | 19 |
| Figure 12. System Design | 21 |

List of Tables

| | |
|---------------------------------|----|
| Table 1. Dataset by head() | 14 |
| Table 2. Description of dataset | 15 |

1. Introduction

Machine learning is the ability of a system to automatically learn from past experience and improve performance. Now a days machine learning for education gains more attention. We are going to use Multiple Linear Regression algorithm to predict student performance or final marks, also we are going to use software tools like anaconda Navigator, Jupyter Notebook, Visual Studio Code and using Python libraries like numpy, pandas for analysis the data, pyplot of matplotlib and seaborn for Visualizing the data. and sklearn(scikit-learn) for prediction of result. API is created in Flask(micro web frame).

Machine learning is used for analyzing data based on past experience and predicting future performance. Machine learning algorithms is a subset of artificial intelligence. It determines the behaviour of the dataset and maximizes its performance. The intention of our project to create a website to analyse, visualize and predict performance of student that will help to improve the marks of the student and to learn a more about machine learning. ML is a pretty multidisciplinary field and it mainly involves programming and mathematics (mostly working with probabilities and density functions). Also, as it is somewhat new and rather complicated, it requires good research skills.

This document goes from the basic explanation of what ML is and what types are there, to the applying of one of ML algorithms(Multiple Linear Regression), going through the analysis of the data provided and the algorithm that have been tested. Whenever it has been possible, graphics and flowcharts have been provided to clarify the explanations, along with some examples. Finally, to avoid it being dense, all the code and scripts are provided in the appendix section.

2. Machine Learning (ML)

2.1. A little bit of history

It might seem that this is a pretty new technology, but, in fact, it isn't. The first ML-related work dates from 70 years ago, in 1950.

In the early days of AI, ML research used mainly symbolic data, and algorithm design was based on logic. At about the same time, Frank Rosenblatt proposed the *Perceptron*, a statistical approach based on empirical risk minimization. However, this approach remained unrecognized and undeveloped in the following decades.

The real development of statistical learning came after 1986, when David Rumelhart and James McClelland proposed the nonlinear backpropagation algorithm. AI, pattern recognition, and statistics researchers became interested in this approach and nowadays is highly used in deep learning Neural Networks.

For the sake of curiosity, below there is a chronological list of the most relevant events in this field since the “starting point” in 1950

1950 – Alan Turing creates the “Turing Test”. This test determined whether a computer had real intelligence or not.

1952 – Arthur Samuel writes the first computer learning program. It played checkers, and the computer was able to improve at the game the more it played, studying which moves made up to winning games.

1957 – Frank Rosenblatt designed the first neural network for computers (the perceptron), which simulates the thought processes of the human brain.

1967 – The “Nearest Neighbours” algorithm was written, allowing computers to begin using very basic pattern recognition.

1979 – Students in Stanford University invent the “Stanford Cart”, which can

navigate obstacles in a room on its own.

1981 – Gerald Dejong introduces the concept of Explanation Based Learning (EBL), in which a computer analyses training data and creates a general rule it can follow by discarding unimportant data.

1985 – Terry Sejnowski invents NetTalk, which learns to pronounce words the same way a baby does.

1986 – David Rumelhart and James McClelland propose the nonlinear backpropagation algorithm.

1990s – Work on Machine Learning shifts from a knowledge-driven approach to a data-driven approach. Scientists begin creating programs for computers to analyse large amounts of data (Big Data) and draw conclusions (or “learn” from the results).

1997 – IBM’s Deep Blue beats the world champion at chess.

2006 – Geoffrey Hinton coins the term “deep learning” to explain new algorithms that let computers “see” and distinguish objects and text in images and video.

2011 – Google Brain is developed and its deep neural network can learn to discover and categorize objects much the way a cat does.

2012 – Google’s X Lab develops a machine learning algorithm that is able to autonomously browse YouTube videos to identify the videos that contain cats.

2014 – Facebook develops *DeepFace*, a software algorithm that is able to recognize or verify individuals on photos to the same level as humans can.

2015 – Amazon launches its own machine learning platform.

2015 – Microsoft creates the Distributed Machine Learning Toolkit, which enables

the efficient distribution of machine learning problems across multiple computers.

2016 – Google’s *AlphaGo*, an artificial intelligence algorithm, beats a professional player at the Chinese board game Go, which is considered the world’s most complex board game and is many times harder than chess. It managed to win five games out of five.

2.2. What is Machine Learning?

Machine learning has become one of the mainstays of information technology in the past two decades and thus, an important, but hidden, part of our life. The increasing amount of data that is being generated (and stored) daily by individuals and corporations, demands a smart analysis. It is here where machine learning comes to stage as a necessary ingredient for technological progress .

As the word stands for, machine learning is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. It focuses on achieving that programmable devices and “machines” learn automatically, by themselves. Basically, it is all about systems learning from data.

The field is closely related to pattern recognition and statistical inference. It works with data and processes it to discover patterns that can be later used to analyse new data. It usually relies on specific representation of data, a set of “features” that are understandable for a computer. For example, if text had to be represented, it should be through the words it contains or some other characteristics such as length of the text, number of emotional words, etc. This representation depends on the task that one is dealing with and is typically referred to as “feature extraction” .

2.3. Types of ML

All Machine Learning tasks can be classified in several categories; the main ones are:

2.3.1. Supervised learning

Relies on a training set where some characteristics of the data are known, typically the labels or classes (the variables to predict). For example:

A computer has to be taught to distinguish pictures of cats and dogs. Some pictures of cats and dogs might be tagged with “cat” or “dog”, respectively. Labelling is usually done by human annotators to ensure high quality of data. Having these true labels of the pictures, they can be used to “supervise” the algorithm in learning the right way to classify images. Once it has learned how to classify them, it can be used on new data and predict labels (“cat” or “dog” in this case) on previously unseen images

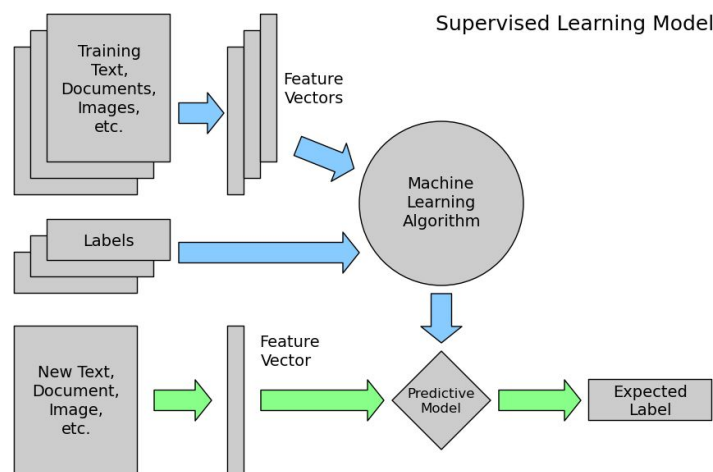


Figure 1. Supervised Learning model

2.3.2. Unsupervised learning

As it can be guessed from the name, in unsupervised ML the algorithm is deprived of the labels used in the previous one. It is just provided with a large (usually huge) amount of data and characteristics of each observation (i.e. a single piece of data).

Its aim is usually finding patterns among this data. For example:

Taking the last example, imagine that someone forgot to label the images of cats and dogs. However, they have to be split into two categories as well. Unsupervised ML might be used (in this case a clustering technique) to separate images in two groups based on some inherent features (characteristics) of the pictures.

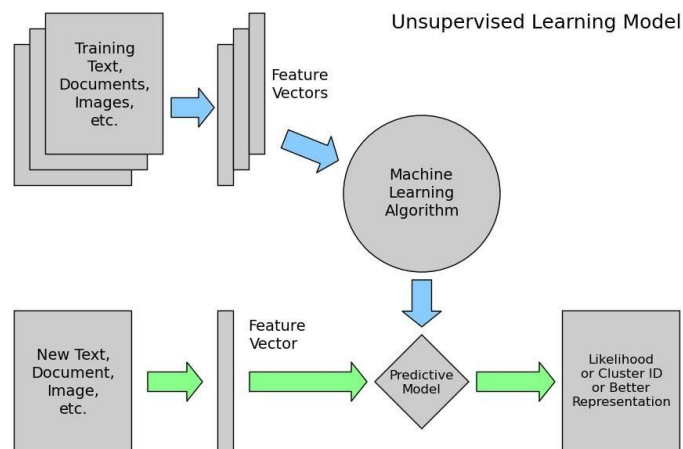


Figure 2. Unsupervised Learning model

2.3.3. Reinforcement learning

The algorithm learns to react to an environment (or to some conditions) by giving positive rewards to “satisfactory” behaviours, and negative or none to “unsatisfactory” ones.

This can be easily illustrated by an example of learning to play chess. The input for the algorithm in this case is the information about whether a game played was won or lost. It does not have to have every move in the game labelled as successful or not, but only the result of the whole game. Therefore, the ML algorithm can play lots of games, and each time it gives bigger “weights” to those moves that resulted in a winning combination, and less to those leading to a lost game.

2.3.4. Lazy learning

It is a learning method in which generalization beyond the training data is delayed until a query is made to the system. They are called lazy because they wait as much as they can to create the model .

They learn rapidly, but they classify slowly and require large space to store the entire training dataset.

K-Nearest Neighbours is an example.

2.3.5. Eager learning

As opposed to the previous one, the system tries to construct a general, input independent target function during the system training .

The main advantage is that it requires much less space than a lazy learning system.

Also, they deal much better with noise in the training data. However, the model creation might be slow.

Artificial Neural Networks (ANN) or Support Vector Machines (SVM) are an example.

2.4. State of the art

Internet browsers have facilitated the acquisition of large amounts of information. This technology's development has greatly surpassed that of data analysis, making large chunks of data uncontrollable and unusable. This is one reason why so many people are enthusiastic about ML.

In fields such as molecular biology and the Internet, if the concern is only about the ease of gene sorting and distributing information, there is no need of ML. However, when it comes to advanced problems such as gene functions, understanding information, and security, Machine Learning is inevitable.

As it has been seen before in section 2.1 though, ML is not a new thing and it has been continuously growing through years. Despite this fact, its foundations have remained the same. They are, basically:

- **Statistics.** ML's aim is to estimate a model from observed data, so it must use statistical measures to evaluate the model's performance and estimate it. It also needs statistics to filter noise in the data.
- **Computer science algorithm design** methodologies, for optimizing parameters and executions.

Bearing this in mind, a few of the problems that Machine Learning can solve will be explained, giving real world applications when possible. Of course, it does solve a wide variety of problems, but below are the most relevant.

- **Classification:**

In this case, having an input dataset, the algorithm's goal is to, for each new unclassified data sample, be capable of assigning it to a category after performing some type of operation on it.

A real application is email spam detection, where each element is represented as a vector of features (e.g. the number of times a specific word is repeated) and different algorithms trained with other already classified emails are applied.

- **Prediction:**

With an already done classification, if the events to predict are similar to the previous ones, it can be determined to which class the new event belongs to. A real application is market trend prediction for hedge funds. For example, using certain data from tweets written in a certain country

- **Pattern recognition:**

Its aim is to extract repetitive structures or common characteristics between the data samples and form certain patterns.

Facial recognition is a clear example of this application. Though faces are not

always the same, they all have a generic structure (i.e. eyes, eyebrows, mouth, ears...).

- Clustering:

This is an unsupervised procedure and consists in figuring out the existence of groups among the data. Thereby, objects with the same characteristics would be grouped under a same cluster or group, represented by its centre.

Grouping a company's customers into several clusters (by age for example) so that adverts could be personalized, would be a real application.

- Regression:

It is similar to classification, in this case though, the class or output variable is continuous. For example: trying to predict the housing prices under huge databases of the real-estate market.

Although it is capable of solving many problems and is a very promising field, there are also some areas where ML still performs poorly nowadays and where most of the investigation is centered. These are:

- Transfer learning:

It is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned .

For example: a vision system that can learn to detect objects invariant to things like lighting changes, rotation, etc. by learning useful features after observing somewhat unrelated/different objects from different points of view and in different lighting conditions.

What is interesting about this is that, even after the system has learnt, it can still improve as it learns more invariant features from different but related objects.

- One-shot learning:

Whereas most ML based object categorization algorithms require training on hundreds or thousands of images and very large datasets, one-shot learning aims to learn information about object categories from one, or only a few, training images .

This can be achieved using architectures with augmented memory capacities, such as Neural Turing Machines, which offer the ability to quickly encode and retrieve new information and avoid having to inefficiently relearn their parameters (as gradient-based networks do).

3. Literature Survey

1. A Machine Learning Approach for Analysis and Predicting Student Performance in Degree Programs. Novel machine learning method is used in this project. Data driven(college combined dataset) approach is used to predict students' final marks or performance. Input parameters given to this system are students' marks of mid sem-I , mid sem-II , Internal and External which are predicting future marks(End-Sem) using the past results of those parameters.
2. Student performance prediction using machine learning: In this project We used Multiple Linear Regression algorithms to predict students performance or Final marks .
3. We used different software tools like Anaconda Navigator(Jupyter Notebook), Visual Studio code, python IDLE to develop the machine or website which is based on a Machine learning approach.
4. We used Python Programming Language for writing code. we basically implement Machine learning model with python.
5. We used different python libraries like numpy for Scientific calculations, pandas for analysis the data, pyplot of matplotlib and seaborn for Visualizing the data. and sklearn(scikit-learn) for prediction of final result.
6. We created an API of the Machine learning model which is working on the algorithm of Multiple linear regression by using Flask(micro web frame).

4. Mission Marks

4.1. Description of the project

Our project aims to analyze, visualize the previous result dataset and train and test the machine by using machine learning algorithms (Multiple linear regression) for predicting student marks on the basis of previous results.

4.1.1. How is the plan going to be solved?

Multiple linear regression algorithms will be used to come up with a good result in this contest. a model on multiple linear regression will be explained, tried and tested, and finally we will get to see the working of the model.

Step #1 : Data Pre Processing

- Importing The Libraries.
- Importing the Data Set.
- Encoding the Categorical Data.
- Splitting the Data set into Training Set and Test Set.

Step #2: Fitting Multiple Linear Regression to the Training set

Step #3: Predicting the Test set results.

4.1.2. workflow

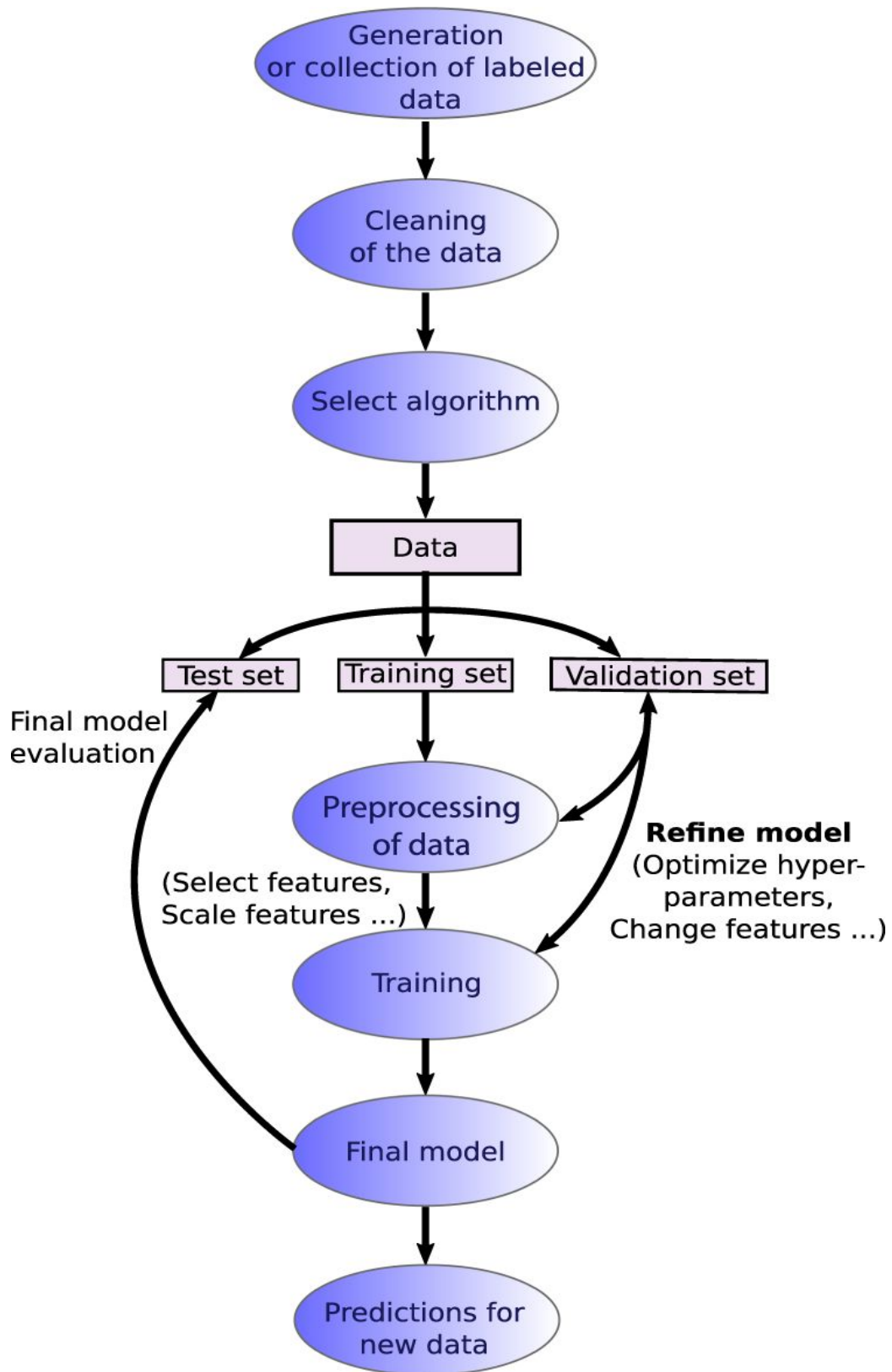


Figure 3. Workflow of model

4.2. Dataset Analysis

The provided dataset has different “features”, each one being of a different relevance. In this chapter we will proceed to analyse this database and extract the useful information out of it.

There are a total of 415 rows and 5 columns in the dataset.

Each record contains the following information:

- **Mid Sem - I** : First Mid sem marks out of 20
- **Mid Sem - II**: Second Mid sem marks out of 20
- **Internal**: Internal marks out of 10
- **External**: External marks out of 30
- **End Sem**: Final marks out of 70

4.2.1. Dataset: by head() Function

```
In [7]: df= pd.read_csv("C:\\Users\\Agnim S\\Desktop\\FINALSTUDENTDATA.csv")
```

```
In [12]: df.head()
```

Out[12]:

| | MS-I | MS-II | INTERNAL | EXTERNAL | ENDSEM |
|---|------|-------|----------|----------|--------|
| 0 | 15 | 10 | 8 | 22 | 55 |
| 1 | 9 | 15 | 7 | 30 | 30 |
| 2 | 8 | 7 | 6 | 22 | 65 |
| 3 | 12 | 15 | 10 | 30 | 65 |
| 4 | 20 | 20 | 10 | 30 | 56 |

Table 1. Dataset by head()

4.2.2. Info of Dataset:

- ❑ class 'pandas.core.frame.DataFrame'
- ❑ RangeIndex: 415 entries, 0 to 414
- ❑ Data columns (total 5 columns):
- ❑ MS-I 415 non-null int64
- ❑ MS-II 415 non-null int64
- ❑ INTERNAL 415 non-null int64
- ❑ EXTERNAL 415 non-null int64
- ❑ END SEM 415 non-null int64
- ❑ dtypes: int64(5)
- ❑ memory usage: 16.3 KB

4.2.3. Description of Dataset:

| | MS-I | MS-II | INTERNAL | EXTERNAL | ENDSEM |
|--------------|------------|------------|------------|------------|------------|
| count | 415.000000 | 415.000000 | 415.000000 | 415.000000 | 415.000000 |
| mean | 10.397590 | 11.722892 | 7.380723 | 25.231325 | 41.108434 |
| std | 5.244596 | 4.039399 | 1.646139 | 2.914976 | 14.791312 |
| min | 0.000000 | 0.000000 | 5.000000 | 10.000000 | 7.000000 |
| 25% | 7.000000 | 9.000000 | 6.000000 | 23.000000 | 30.000000 |
| 50% | 11.000000 | 12.000000 | 7.000000 | 25.000000 | 42.000000 |
| 75% | 14.000000 | 15.000000 | 9.000000 | 28.000000 | 52.000000 |
| max | 20.000000 | 20.000000 | 10.000000 | 30.000000 | 70.000000 |

Table 2. Description of dataset

4.3. Data Visualization

4.3.1. Correlation Matrix of Dataset

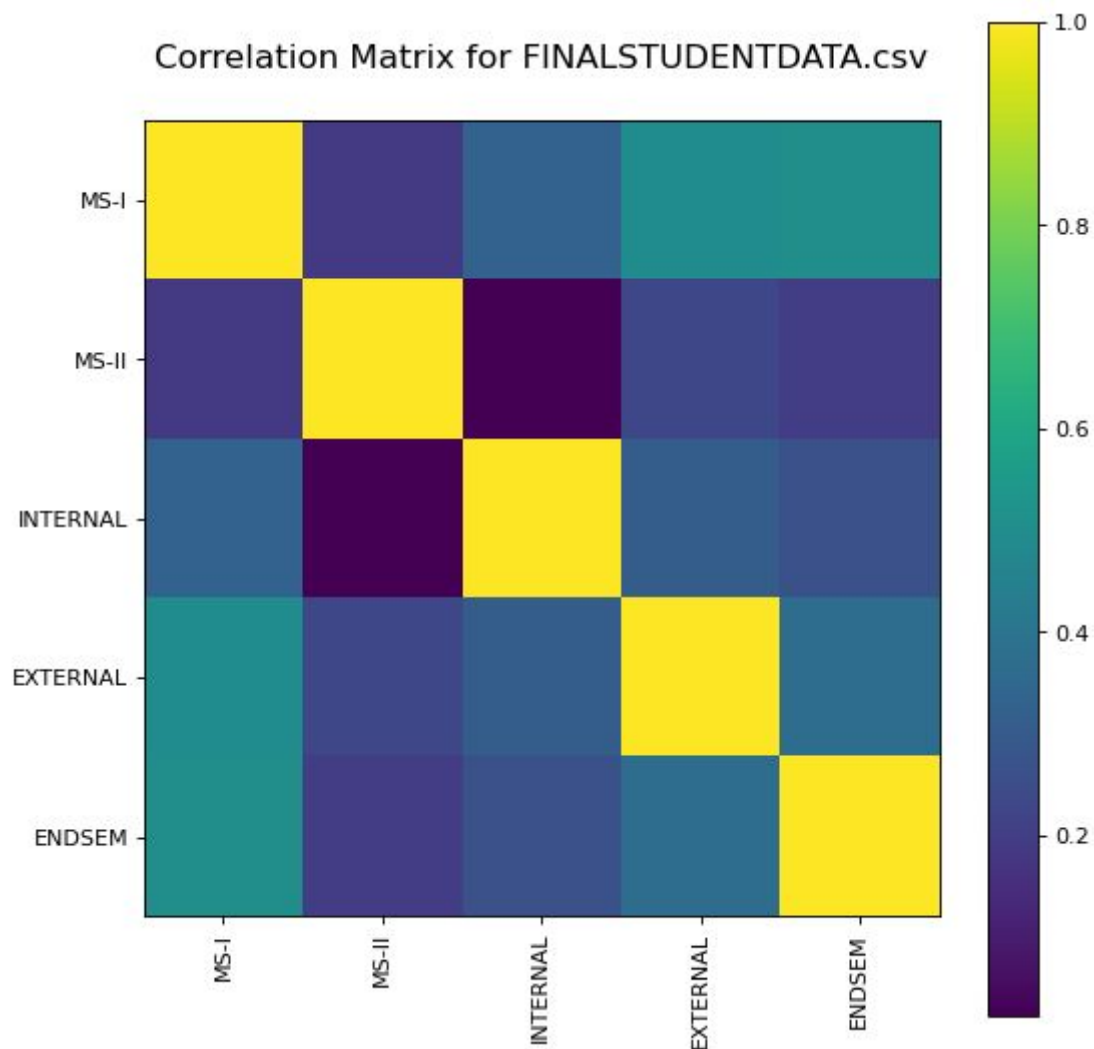


Figure 4. Correlation matrix

4.3.2. Graph of all columns
#midsem-I

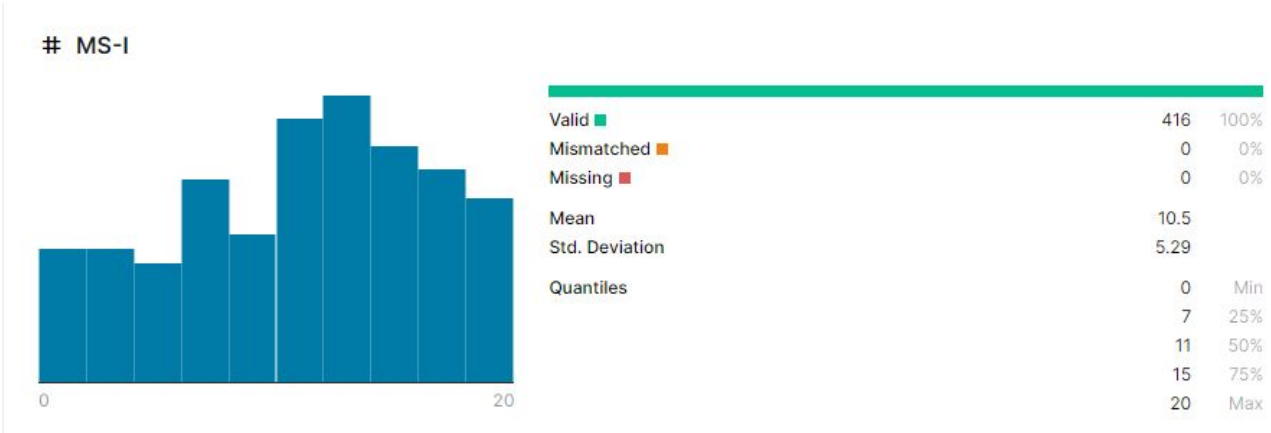


Figure 5. Graph MS-I

#Mid Sem-II

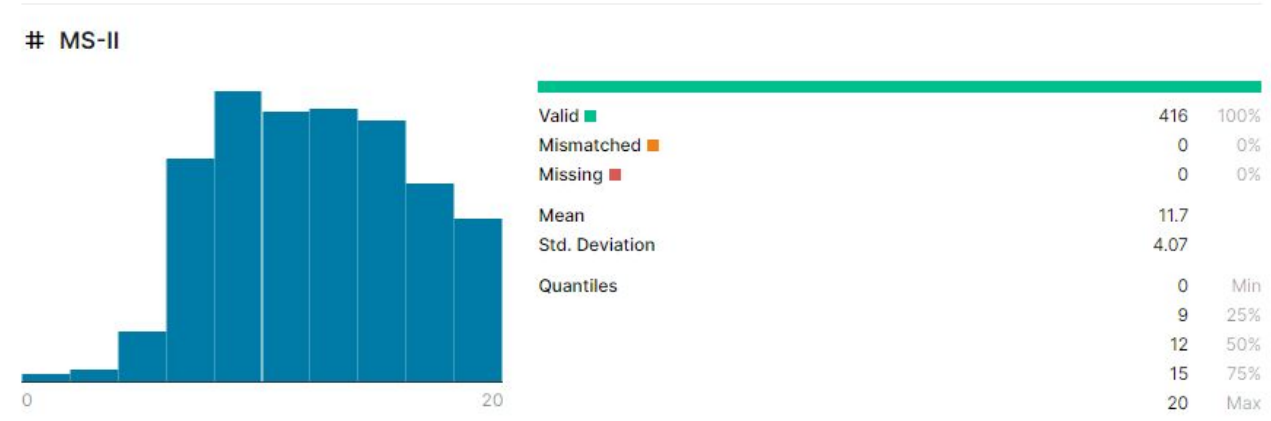


Figure 6. Graph MS-II

#Internal Marks

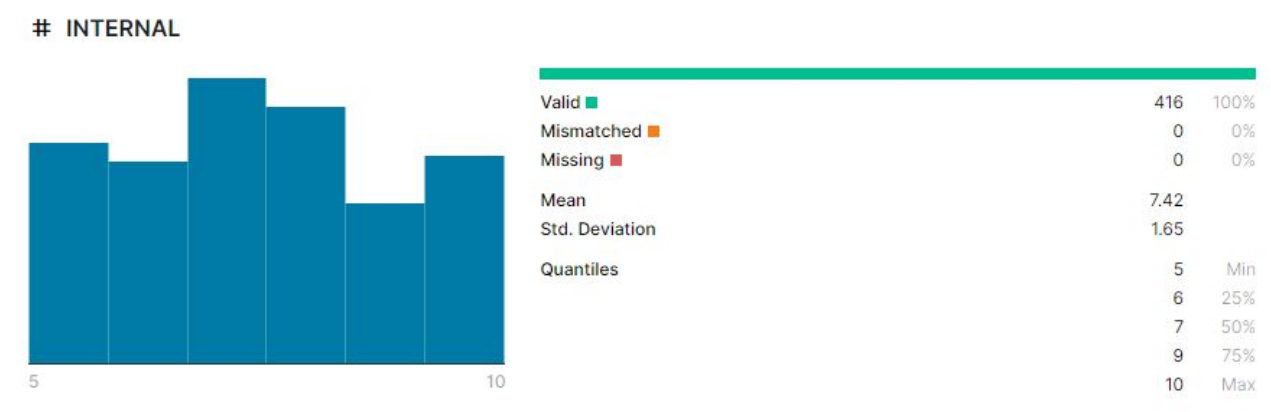


Figure 7. Graph Internal

#External Marks

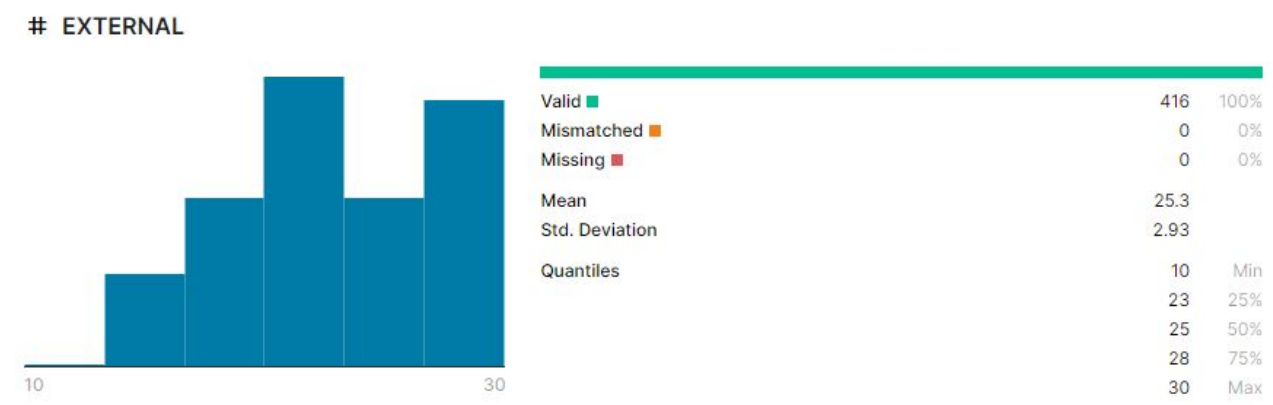


Figure 8. Graph External

#EndSem(Final Marks)

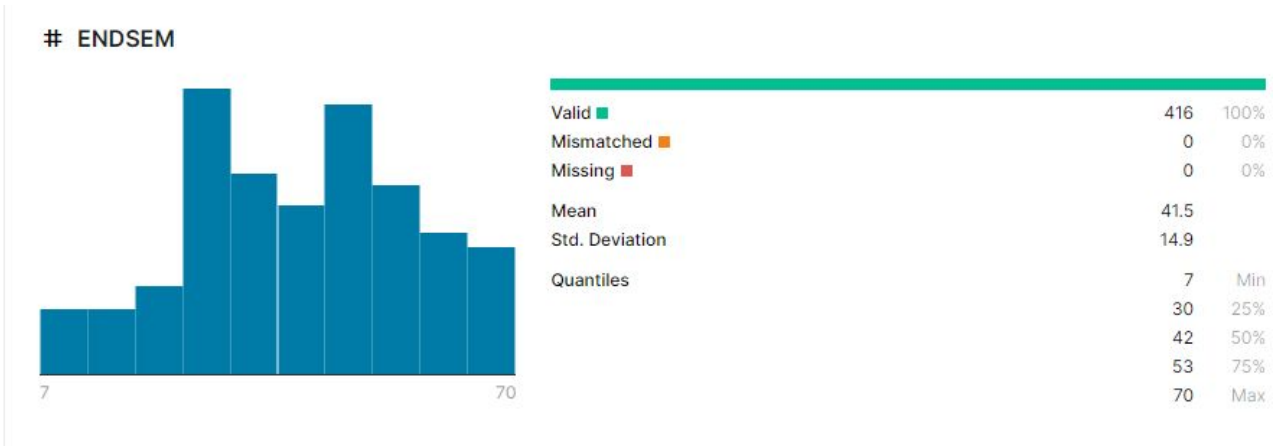


Figure 9. Graph Endsem

#Combine all Dataset



Figure 10. Graph Dataset

4.3.3. Scatter And Density Plot

Scatter and Density Plot

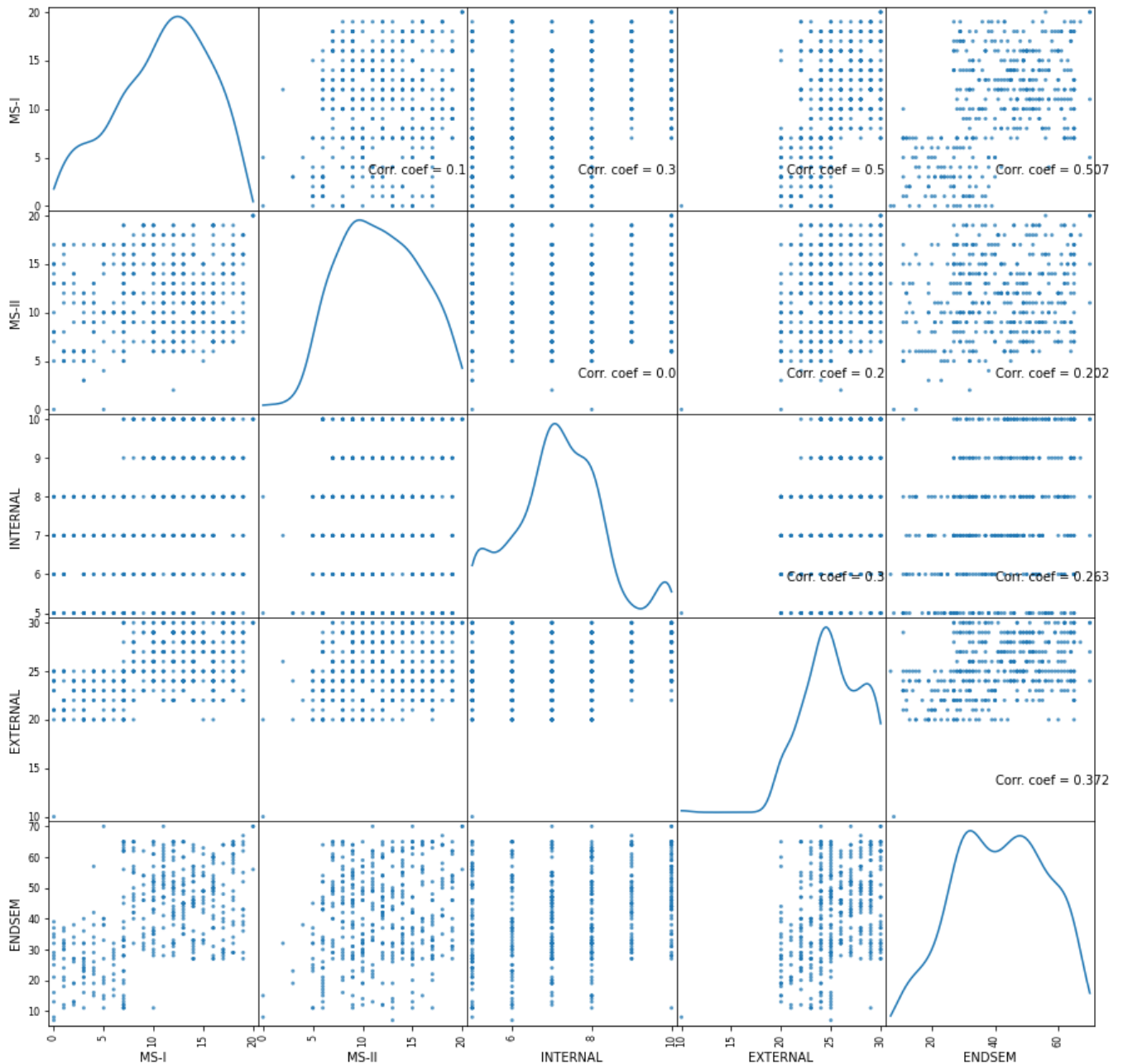


Figure 11. Scatter and density plot

4.4. Dataset split

Split dataset into 2 parts to train, test the algorithms.

```
train,test = data_split(data, 0.2)
```

80% of the dataset is going to be used for training the model and the remaining 20% of the dataset are going to be used for testing the model.

we created function to split the dataset into 2 part

function:

```
def data_split(data, ratio):  
    np.random.seed(42)  
    shuffled = np.random.permutation(len(data))  
    test_set_size = int(len(data)*ratio)  
    test_indices = shuffled[:test_set_size]  
    train_indices = shuffled[test_set_size:]  
    return data.iloc[train_indices], data.iloc[test_indices]
```

```
train,test = data_split(data, 0.2)
```

4.5. System Design

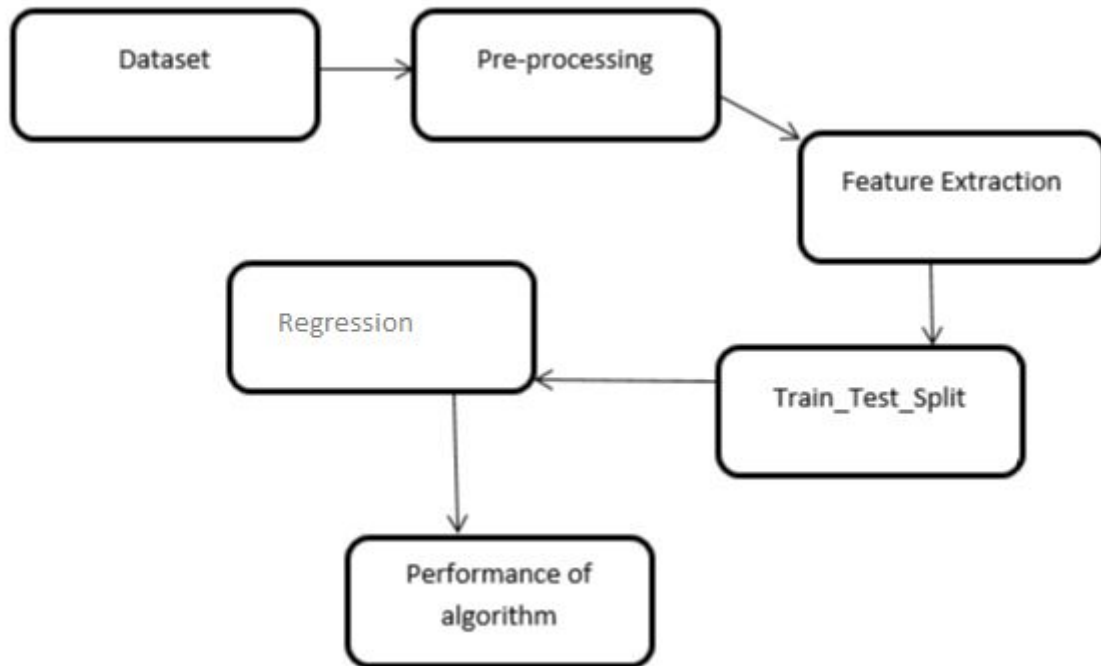
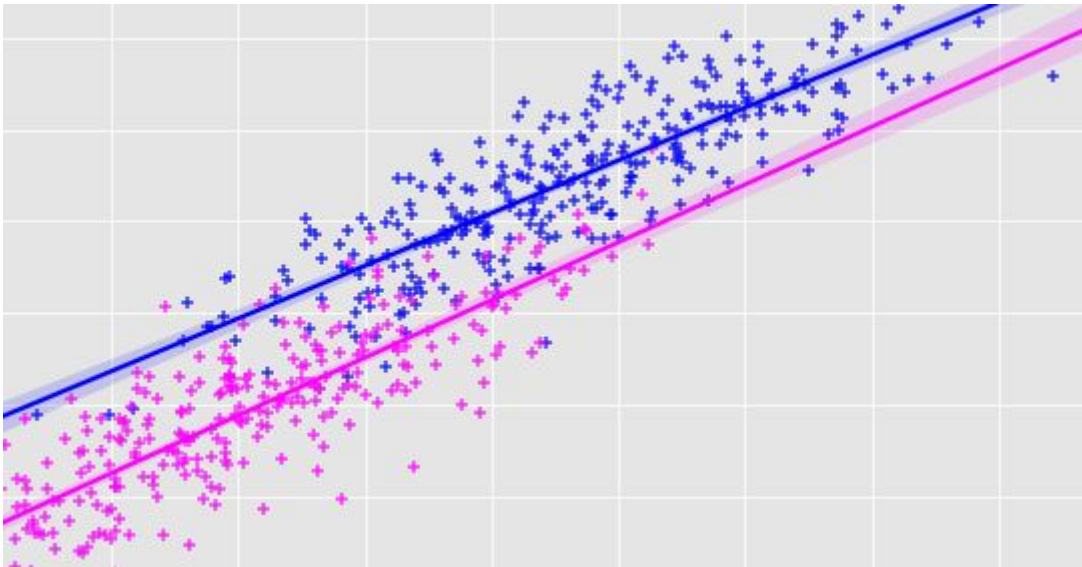


Figure 12. System Design

5. Implementation of Machine Learning Algorithms

Multiple Linear Regression :



Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relation between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

In multiple linear regression, multiple lines of best fit are used to obtain a general equation from the training dataset which can then be used to predict the values of the testing dataset. The general equation can be of the form: $y = ax + bx_2 + \dots + c$ where y is the predicted value, a and b are the gradients of the lines connecting the independent variables to the dependent variable and c is the point at which the line strikes the y-axis

$$Y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \dots + b_n * x_n$$

$Y =$ Dependent variable and $x_1, x_2, x_3, \dots, x_n =$ multiple independent variables

Import Multiple Linear regression model in the project :

```
from sklearn.linear_model import LinearRegression
```

Split dataset into test and train for implement Multiple Linear Regression:

```
train,test = data_split(data, 0.2)

x_train = train[['MS-I','MS-II','INTERNAL','EXTERNAL']].to_numpy()
x_test = test[['MS-I','MS-II','INTERNAL','EXTERNAL']].to_numpy()

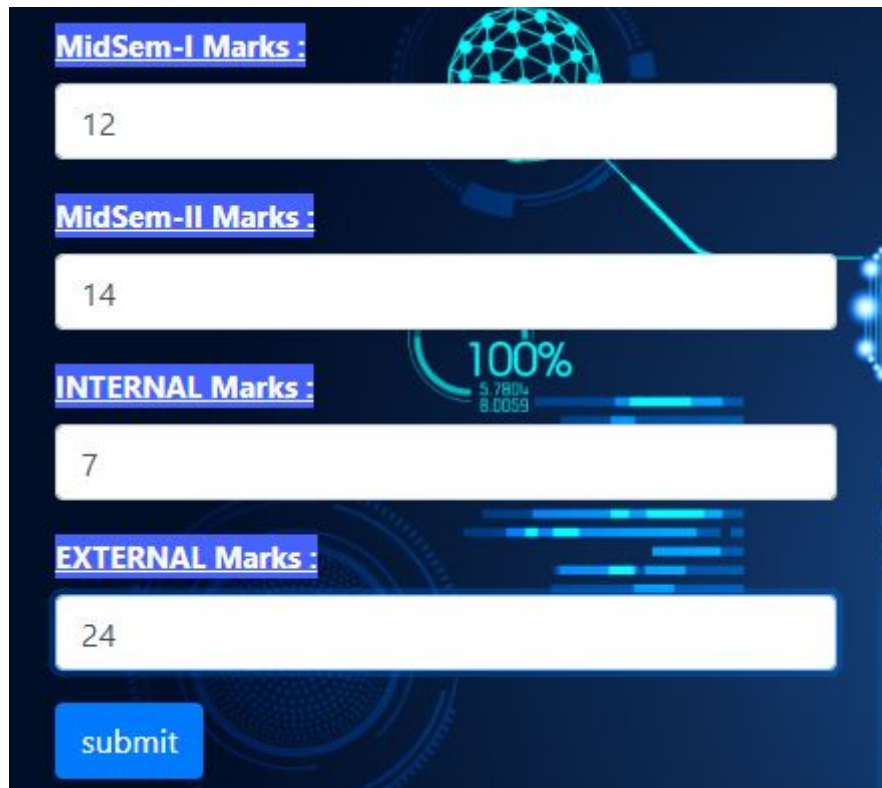
y_train = train[['ENDSEM']].to_numpy().reshape(332,)
y_test = test[['ENDSEM']].to_numpy().reshape(83,)
```

Use Multiple Linear regression model in the project:

```
clf = LinearRegression()
clf.fit(x_train, y_train)
```

6. Mission Marks Machine Learning Prediction on new Data(snapshots):

Input[1] for predicting final marks:



MidSem-I Marks : 12

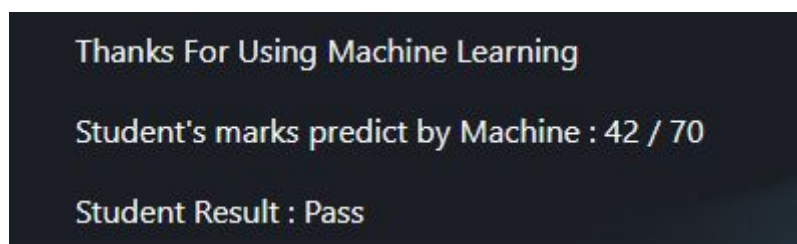
MidSem-II Marks : 14

INTERNAL Marks : 7

EXTERNAL Marks : 24

submit

Marks predict final marks by mission marks model:

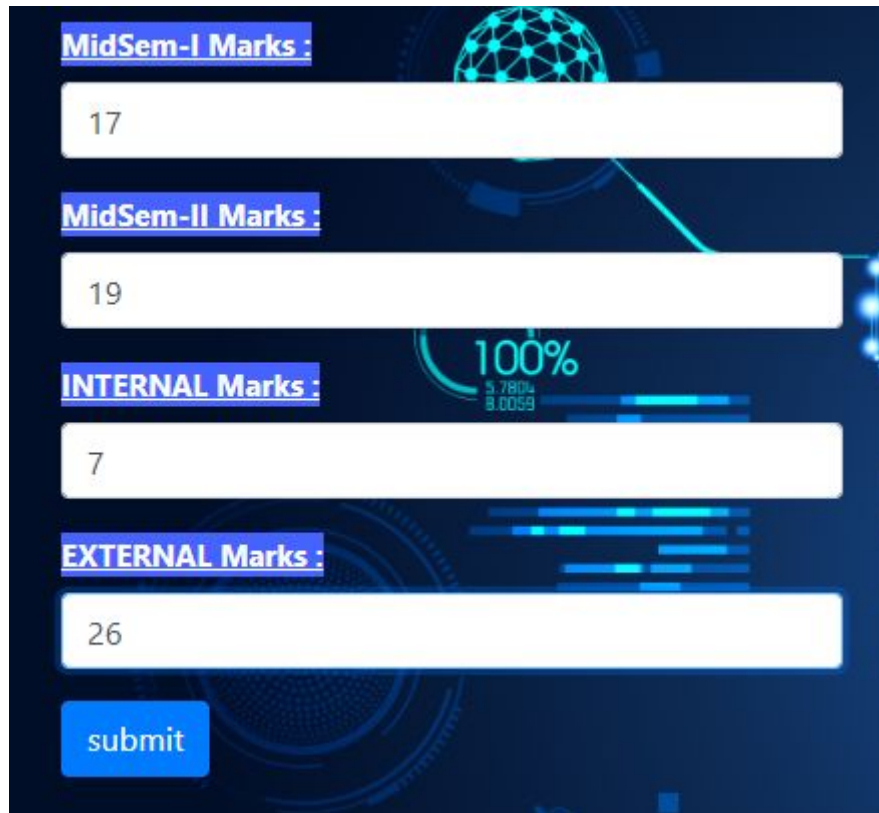


Thanks For Using Machine Learning

Student's marks predict by Machine : 42 / 70

Student Result : Pass

Input[2] for predicting final marks:



The image shows a web interface for a machine learning project. It has a dark blue background with a futuristic, glowing network pattern. There are four input fields, each with a label above it in a blue box: 'MidSem-I Marks :', 'MidSem-II Marks :', 'INTERNAL Marks :', and 'EXTERNAL Marks :'. The values entered in the fields are 17, 19, 7, and 26 respectively. Below the fields is a blue 'submit' button. In the background, there is a circular progress indicator showing 100% completion.

| Category | Marks |
|-----------------|-------|
| MidSem-I Marks | 17 |
| MidSem-II Marks | 19 |
| INTERNAL Marks | 7 |
| EXTERNAL Marks | 26 |

Marks predict final marks by mission marks model:

Thanks For Using Machine Learning
Student's marks predict by Machine : 50 / 70
Student Result : Pass

Input[3] for predicting final marks:

MidSem-I Marks :
8

MidSem-II Marks :
9

INTERNAL Marks :
7

EXTERNAL Marks :
21

submit

Marks predict final marks by mission marks model:

Thanks For Using Machine Learning

Student's marks predict by Machine : 35 / 70

Student Result : Pass

7. Discussion and Conclusions

The success of machine learning in predicting student performance relies on the good use of the data and machine learning algorithms. Selecting the right machine learning method for the right problem is necessary to achieve the best results. However, the algorithm alone can not provide the best prediction results. Feature engineering, the process of modifying data for machine learning, is also an important factor in getting the best prediction results.

Machine learning techniques can be useful in the field of student performance prediction considering that they help to identify marks after external exams that will help students for preparation of the examination and professor to help weak students just before examination. The aim of this paper is to apply machine learning algorithms for prediction of student performance. An early manual analysis of students having poor performance helps the management take timely action to improve their performance through predicting their academic details. Accurately predicting student performance based on their ongoing academic records is predicted. Also we conclude that the proposed system is helping us to make the student performance better. In this project machine learning can prove to be a powerful tool and multiple linear regression algorithms really helps students to predict their marks on the basis of a combined dataset of college results.

The results of this project indicate that feature engineering provides more improvement to prediction results than method selection. Despite feature engineering being done in a limited capacity, it made a bigger difference in prediction performance. Furthermore, the biggest leap in improvement was made in

the case of decision trees, where both feature selection and feature modification is applied to the data. When trying to improve the prediction of student performance, the modification of input data is an important factor besides selecting the right method for the data

8. Future work

This project has certain limitations that must be noted. There was no access to a dedicated student data set, and the study relies on public data sources. In addition, both data sets were small, having less than thousand records. A project that has access to more comprehensive data may offer more conclusive results

Another area that future work can improve is the variety of the machine learning methods. This project used multiple linear regression. Other methods, such as decision trees, clustering and artificial neural networks can be used to have a better Prediction of the Student final Marks.

Final area that can be improved is the process of feature creation. Since the data is limited, the amount of feature modification that can be made is also limited. Both data sources used in this research consists of a single table, and custom variables were created using variables from the same table. With a more comprehensive data set that spans multiple tables, there will be more potential to create new custom variables, while keeping in mind that the more a custom variable is, the more difficult it is to interpret the relation between it and the dependent variable.

References

- ❑ <https://www.udemy.com/course/machinelearning/>
- ❑ <https://pandas.pydata.org/docs/>
- ❑ <https://numpy.org/doc/>
- ❑ https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
- ❑ <https://matplotlib.org/>
- ❑ <https://www.geeksforgeeks.org/machine-learning/>
- ❑ <https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/>
- ❑ https://www.tutorialspoint.com/machine_learning_with_python/index.htm
- ❑ <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/>
- ❑ https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_regression_algorithms_linear_regression.htm
- ❑ https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_data_preprocessing_analysis_visualization.htm

Appendices

Appendices consist of all the code scripts that have been used during this project.

Below there is a reference of all of them, explaining what they are intended to do:

- **main.py:** connect all other python and html code
- **mytraining.py:** Data Preprocessing were held and implemented Multiple linear regression algorithms.
- **analysis.py:** dataset analysis .
- **visualization.py:** visualize different graphs of dataset..
- **index.html:** write front end code for mission marks.
- **show.html:** html page where prediction will show.
- **analysis.html:** shows the analysis of analyis.py.
- **visual.html:** show the graphs of visualization.py.
- **dataset.html:** show whole dataset.

Glossary

ML = Machine Learning

KNN = K-Nearest Neighbours NN = Neural Networks

ANN = Artificial Neural Networks

MS-I = MidSem - First

MS-II = MidSem - Second

EndSem = Final Semester Marks

CSE = Computer Science and Engineering

dept = Department