**Course Difficulty Prediction Using Different**

**Classification Methods**

**B. Om Sudha, INT 247, 11703536, Lovely Professional University**

**Abstract:**

***Machine studying (ML) has turn out to be the most important methodology that indicates good effects in the classification and prediction domains. Predictive systems are being employed to are expecting occasions and its outcomes in almost every stroll of lifestyles. In this mission we are going to expect the tough concern that scholars are facing in Bachelor of Technology (Electrical and Electronic Engineering) by the usage of some category models of Machine Learning. The field of prediction is gaining significance in many approaches in pur actual lives. Moreover, many are suffering for Machine getting to know models that might be used for formulating many actual life examples. Numerous factors such as results of each and each scholar in all of the examinations that the university would had conducted are required to construct these models. This paper gives an analysis of such key models that specialize in utility of machine gaining knowledge of algorithms to predict the result . The effects acquired helped us to clarify the excellent combination of feature choice and type algorithms that render maximum accuracy in course prediction.***

**KEY WORDS:**

Machine Learning, Prediction, Supervised learning, Classification, Feature Selection, Course Prediction

**INTRODUCTION:**

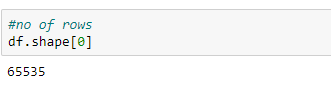
Machine learning provides a way for identifying patterns in data and to make use of them to automatically take good decisions or make predictions. One such common task under machine learning, which involves prediction of the target variables in a previously unrevealed data, is termed as classification. The ultimate purpose of classification is to predict something called a class which is nothing but the target variable, by training the classification model with training dataset and then using the same to predict the class with testing data. This type of processing falls under the category called supervised learning as it tries to find out the relationship between input attributes that are otherwise called independent variables and a target attribute that are otherwise called dependent variable. Few applications of classification include medical diagnosis, email filtering, internet traffic interception, click stream analysis and many more.

Prediction of course is usually considered as a classification problem, as it would be to predict one from the difficult course that the students are facing It has been always challenging in predicting the outcomes of such events. It involves a collection of large number of features like MTT, ETE, ETP, CA like that. This will be useful when any new student wants to know which course is difficult in that particular field.

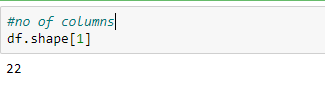
**Data Preprocessing:**

Before training data on a model we need to preprocess the data. Preprocessing of data is known as analyzing and making the data perfect for training.

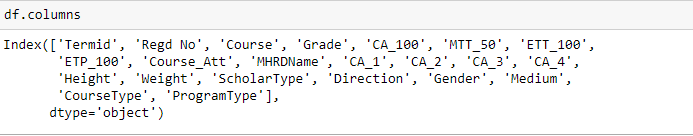
We have total 65535 rows



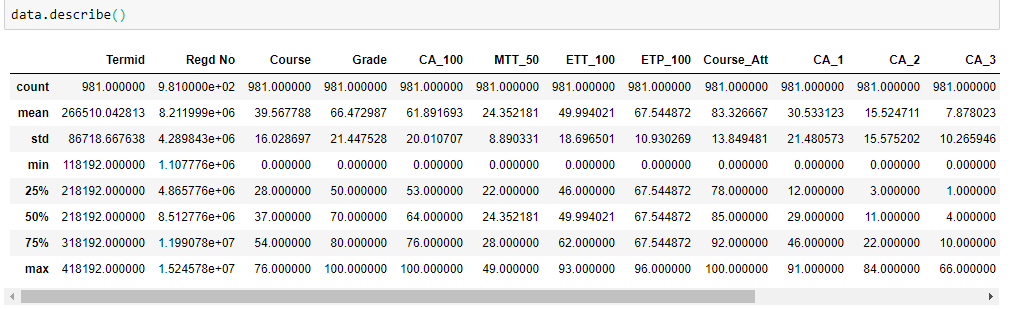
We have total 22 columns



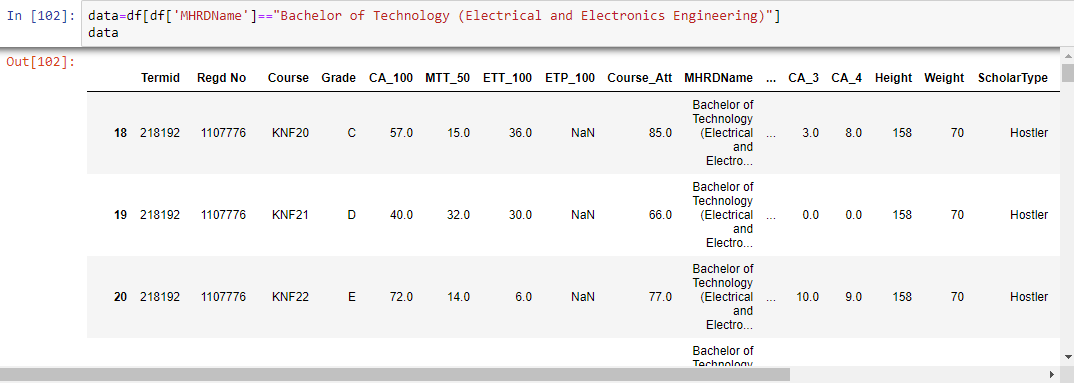
The columns of the data set are below mentioned:



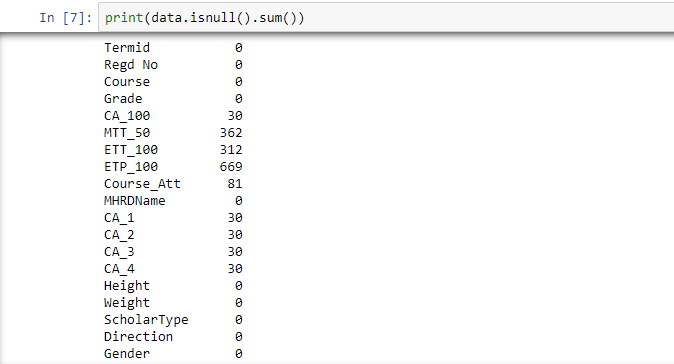
Describing the data:



From the given data extract the one which needed



Null values from the extracting data:



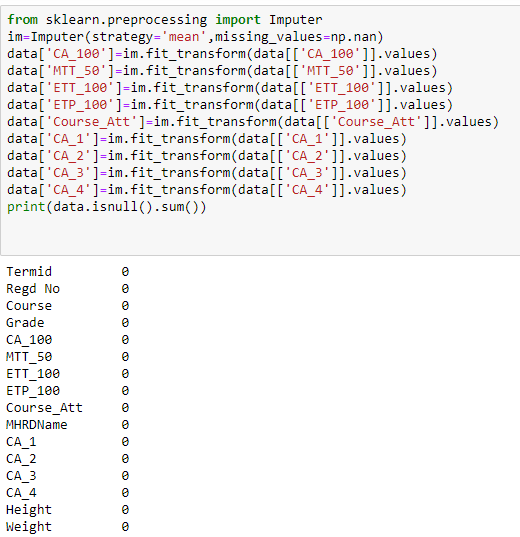
**Encoding:**

**Label Encoding:** Label Encoder encode labels with a value between 0 and n\_classes-1,where n is the number of distinct labels.

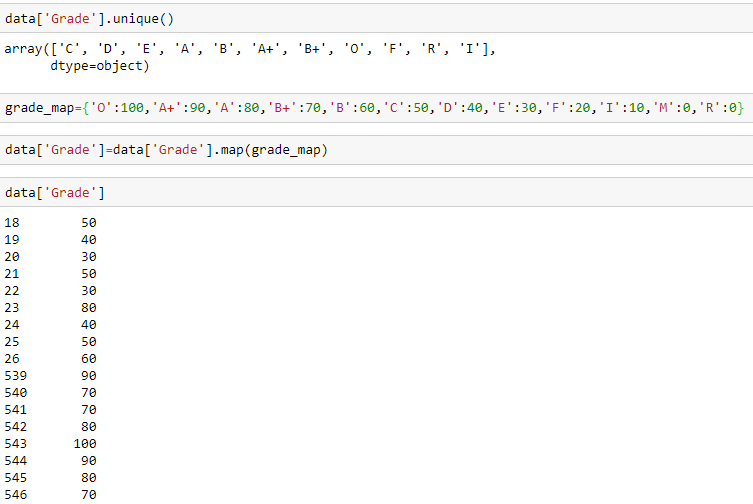
A **One hot encoding** is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values.

Since from our data Label Encoder in the best way to encode the categorical data.

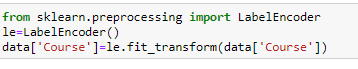
Removing the null values: By using Simple Imputer



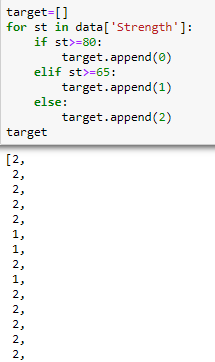
Now encode the grade column manually with out importing library:



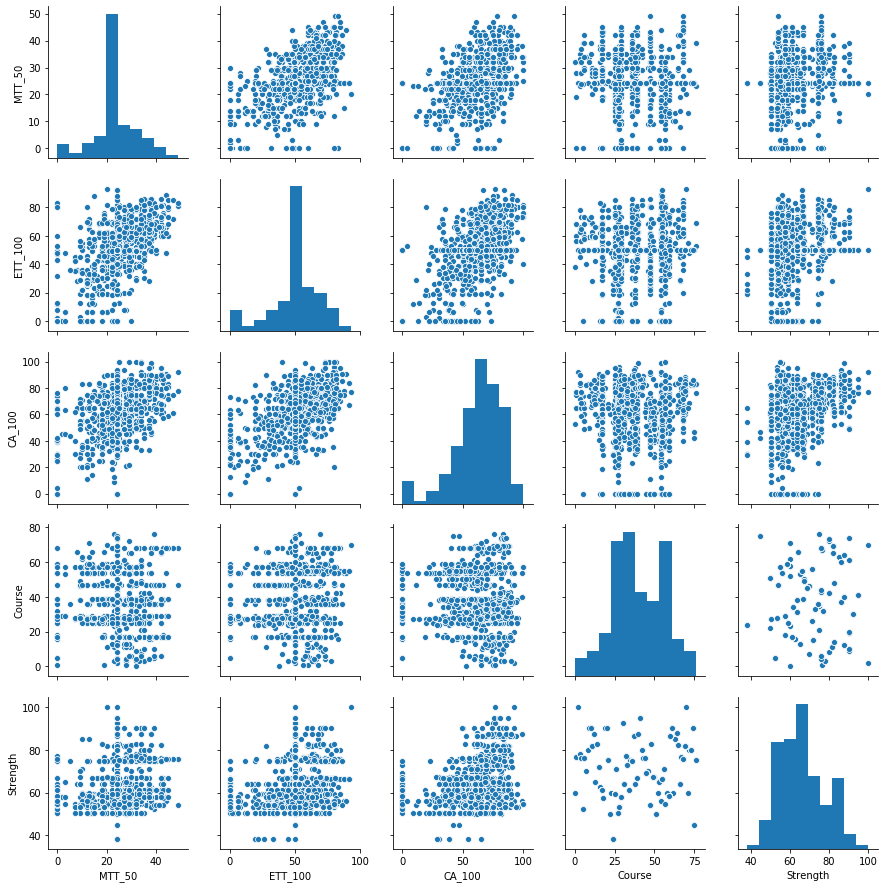
Label encoding of course, since it is used for fitting the data.



From the dataset for getting the particular course difficulty, calculate the mean of grades by grouping it and store it in ‘Strength’ column in data.



**Feature Selection:**

****

For fitting a data with high accuracy we need to select the data from the given data set such that it gives highest accuracy. For that we used the correlation graphs.

**Standardizing/Normalization:**

Since this dataset values are already in between the range(0 to 100),there is no need to normalize or standardize the data i.e making them to fall in the range(-1 to +1).

**Training:**

After preprocessing the data it need to be trained on a model to predict the new data points. For this, we need to:

1. Split the dataset into testing and training points.
2. Selecting an appropriate Training Model.
3. Check the accuracy.
4. Predicting the testing points (Unknown Points).

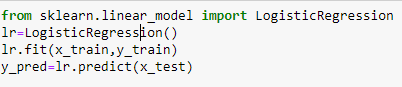
**Choosing a model:**

For choosing an appropriate model, we need to train with different models and by observing the metric we will choose the best fitting model. Let us go through some of the classification models:

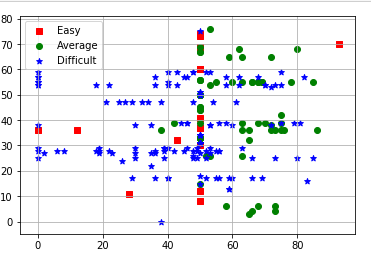
1. **Logistic Regression:**

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.

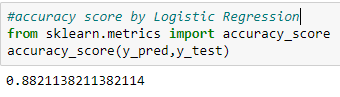
Logistic regression becomes a classification technique only when a decision threshold is brought into the picture



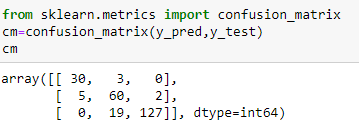
The visualization of the different classes in Logistic regression is:



The accuracy score that obtained by using this Logistic Regression is:



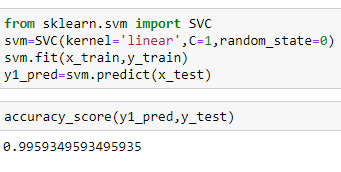
Confusion matrix in Logistic Regression:



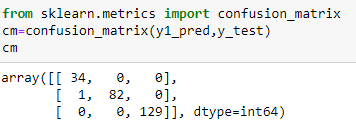
1. **Support Vector Machine:**

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they got refined in 1990.

SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.



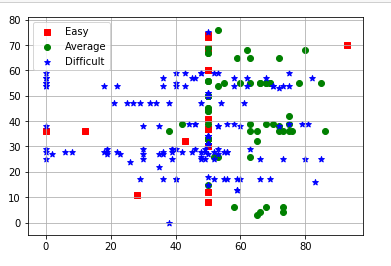
**Confusion matrix** in **SVM**:



In the above code, we can see that the accuracy score that we got by using SVM is more.

It means that SVM is giving us a very good accuracy score, which means that the data we trained is giving us a good accuracy score.

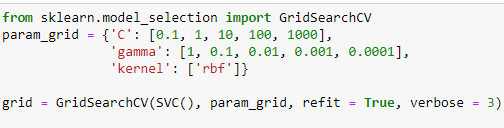
Visualization of the classes by using **SVM**:



**Hyperparameter Tuning:**

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. However, there are some parameters known as **Hyperparameters** and those cannot be directly learned.

**SVM** also has some hyper-parameters (like what C or gamma values to use) and finding optimal hyper-parameter is a very hard task to solve. But it can be found by just trying all combinations and see what parameters work best. The main idea behind it is to create a grid of hyper-parameters and just try all of their combinations (hence, this method is called **Grid search**, But don’t worry! we don’t have to do it manually because Scikit-learn has this functionality built-in with Grid Search CV.

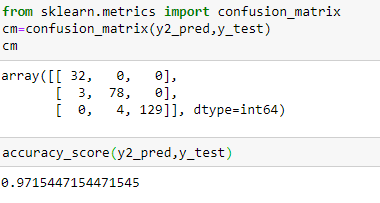
****

# Naive Bayes:

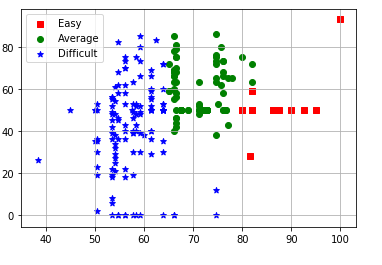
# Naive Bayes classifiers are a collection of classification algorithms based on  ****Bayes’ Theorem****. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

# 

Accuracy score and confusion matrix in Naïve bayes:



Visualizing the classes in naïve bayes:



**Comparative Study of three models:**

From the above all models we need to compare the accuracy score in all the three models.

The accuracy score of ***Logistic Regression***is**:0.88**

The accuracy score of ***SVM*** is**:0.99**

The accuracy score of ***Naive Bayes*** is**:0.97**

So, it is clearly observed that SVM model has given the maximum accuracy score.

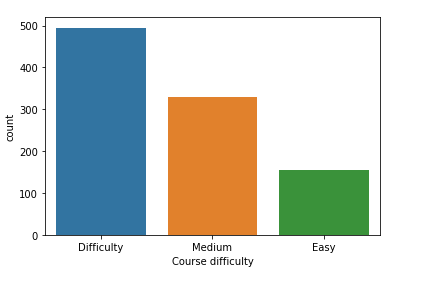
So, SVM is the best.

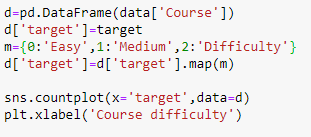
**Results and Discussions:**

From training and predicting the data, we have made three targets, such that all the features come under these targets. The targets are [‘Difficulty’ ,’Average’, ‘Easy’].

The following figure shows count of courses that come under these three different classes.

From this we can predict the easy course, difficulty course and medium course.





**Final results after training and predicting the difficult course**:

