**DEVELOP AND EVALUATE A PREDICTION MODEL USING VARIOUS DATA SCIENCE TECHNIQUES**

**CIS7031 - Programming for Data Analysis**

**Module Leader: Dr Ambikesh Jayal**

**Submitted by**

Rohith Vurukonda

ST20195136

Table of Contents

[Introduction 3](#_Toc73440007)

[Data preparation and Exploration 3](#_Toc73440008)

[Develop and evaluate prediction mode 4](#_Toc73440009)

[Critical review and analysis of techniques used. 13](#_Toc73440010)

[Conclusion 14](#_Toc73440011)

[References 15](#_Toc73440012)

# Introduction

Data evaluation is very essential and important for the data prediction method. This process helps to identify the significant data gaps within the selected database. In this case, a database has been selected for the prediction. The confusion matrix, accuracy score and also hyperparameters using grid search cv and random search cv is going to be visualized by using a random forest classifier. The entire data preparation and exploration are going to be discussed in this assignment. Also, the methods and the function of this data prediction are going to be described in the body of the report. One of the major parts of this assignment is a critical evaluation and analysis parts are also going to be discussed in the body of the report.

# Data preparation and Exploration

Data preparation is used to transform raw data within a form and this is a very unique process for data modeling. This preparation data is one of the most essential and important parts of a predictive modeling. It is also helping to make the selected database more suitable for machine learning. This data preparation is helping to establish the right data collection mechanism and it is known as the most common procedure of the entire machine learning. In that case, there are various types of steps are present here within this data preparation and these steps are given below,

1. Collecting an appropriate dataset
2. Discovering and assess the selected database
3. Cleanse and validate this appropriate dataset
4. Transform and enrich the dataset
5. Store data

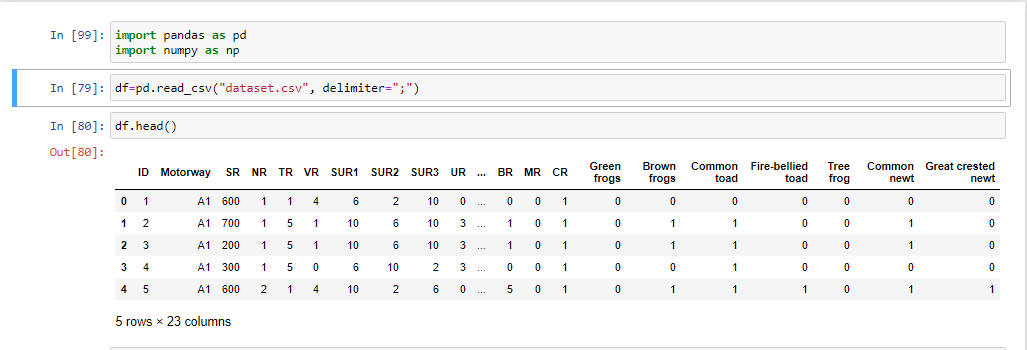
In the case of this data preparation, loading data, transforming data format, cleaning data, rearranging data are the different types of strategy for this. This data preparation is also helping to store various types of data within one data frame.

Data exploration is a very essential and important step for the data science and machine learning process. In that case, it is known as the initial step of data analysis and this data exploration is helping to explore the large dataset in an unstructured process to uncover initial patterns, various characteristics and points of interest (Zadeh*et al.* 2018). Also, it is used for the combined purpose of every manual method and the automated tools such as initial reports, data visualization, and project charts. This data exploration is helping to structure an easy and simple visualization rather than pouring within the various types of unstructured data. In that case, there are various types of procedures are present here within the data exploration and these steps are also given below,

1. Setting an appropriate environment
2. Implementing pandas python library
3. Divide the range of the selected dataset
4. Displaying the entire data type
5. Implementing pandas data structure
6. Accessing the entire series elements
7. Implementing data frame element
8. Querying the entire selected dataset
9. Grouping the dataset
10. Aggregating dataset

These steps of this data exploration are helping to combine large amounts of data and this procedure is also helping to gather the necessary information within the selected dataset. The most important techniques of this data exploration are histogram, correlations, summary statistics and outliers. This entire process is helping to cut down the massive data and converting this selected database to a manageable size. Also, this exploring dataset is helping to develop a deep understanding of the selected database and it is one of the most important skills with the data science and machine learning process (Konstantinou*et al.* 2019).

# Develop and evaluate prediction mode



**Figure 1: code-1**

In order to perform the data evaluation and development in the jupyter notebook the first step is to input some basic yet important libraries. In the above figure, two libraries are imported; one is the pandas, this library allows the dataset to get structured or stored in a two-dimensional Table. Another Library is numpy, this library is generally used for the analysis and evaluation process. In the above figure a dataset is also inputted into a variable named df with the help of pandas so that the dataset is inputted in a two-dimensional structure. Lastly, by using the head () the values in the df function are extracted.

**Pandas**

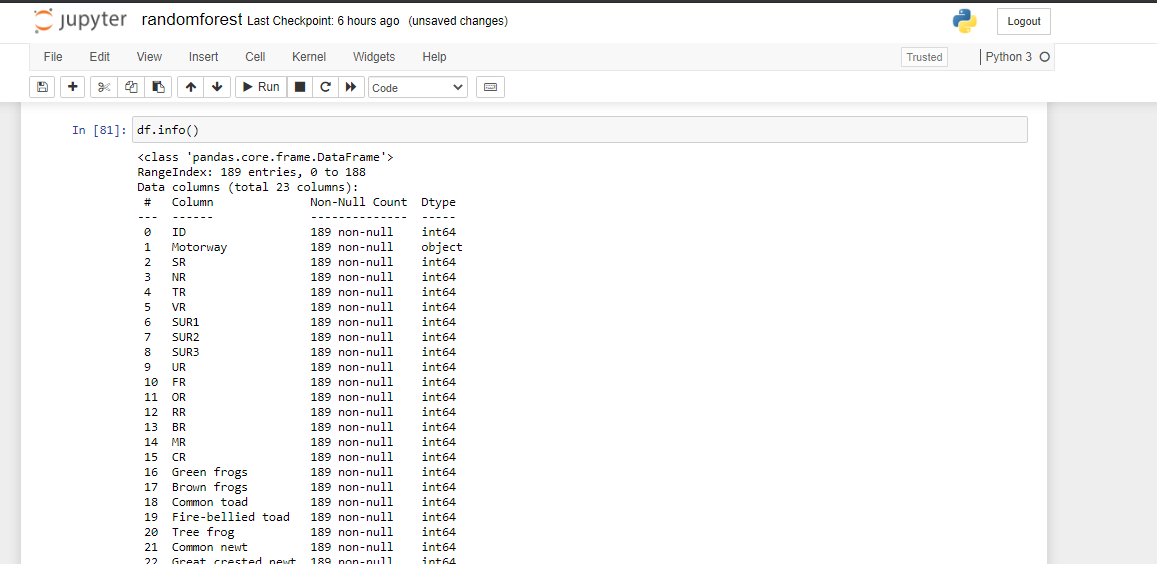
Pandas are known as the software library for the python programming language and this pandas library is mainly helping with data manipulation and analysis. In that case, there are some steps are present to use this pandas library within python and these steps are given below,

|  |  |
| --- | --- |
| Step 1 | Converting a python's list, dictionary and Numpy array into a pandas data frame |
| Step 2 | Open a local file such as a CSV file by using this pandas library |
| Step 3 | It is not allowed to determine the test file and Excel file |
| Step 4 | Allow an open remote file such as CSV and JSON on a website through a URL |

The command form of this pandas is like Pd.dataframe() and this command is helping to read the selected database. This library is also able to provide a fast, flexible and structured dataset to complete this assignment.

**Numpy**

This Numpy library is used for working with the arrays of a given database and various functions are present here within this library such as linear algebra, matrices and Fourier transform. The working function of this library is very easy and simpler rather than the other libraries within python (Zhou *et al.* 2021). This numpy contains different types of multi-dimensional arrays and data structure and it can be utilized to perform a number of mathematical operations such as algebraic routines. This function is also able to handle arithmetic functions and complex numbers and NumPy.mean () function returns the arithmetic mean of the element within the array.

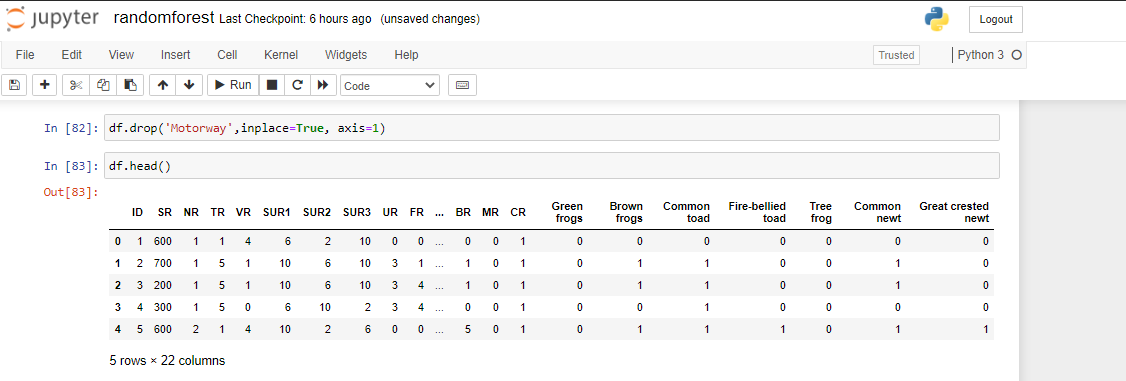


**Figure 2: code-2**

In the above figure, the info() function is being used in order to check for the null values. This null values identification is important to identify as these null values are to be removed if present as this can cause calculation problems or inconsistent outputs.

**Info**

This info() function is used to get a concise summary of the selected data frame and it is also helping to do the exploratory analysis of the given data to get the fastest overview of the dataset. In that case, this info() function is allowed to learn the shape of the object within the selected database. The main object of this function is printing the entire information of a selected database such as index dtype, non-null values, column dtype and also memory usages.



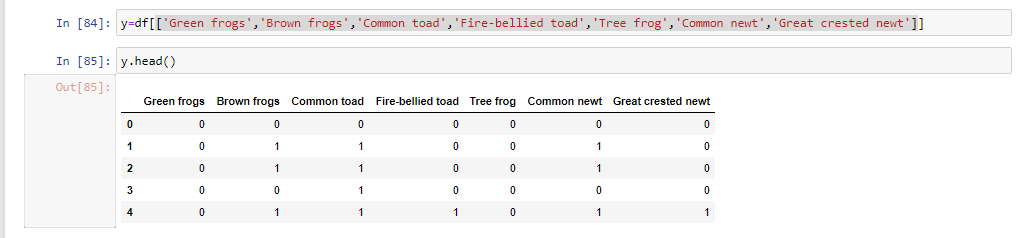
**Figure 3: code-3**

The above figure shows that the motorway column has been removed. This motorway column is removed by using the drop() function. This drop function drops the whole column at once.

**Drop**

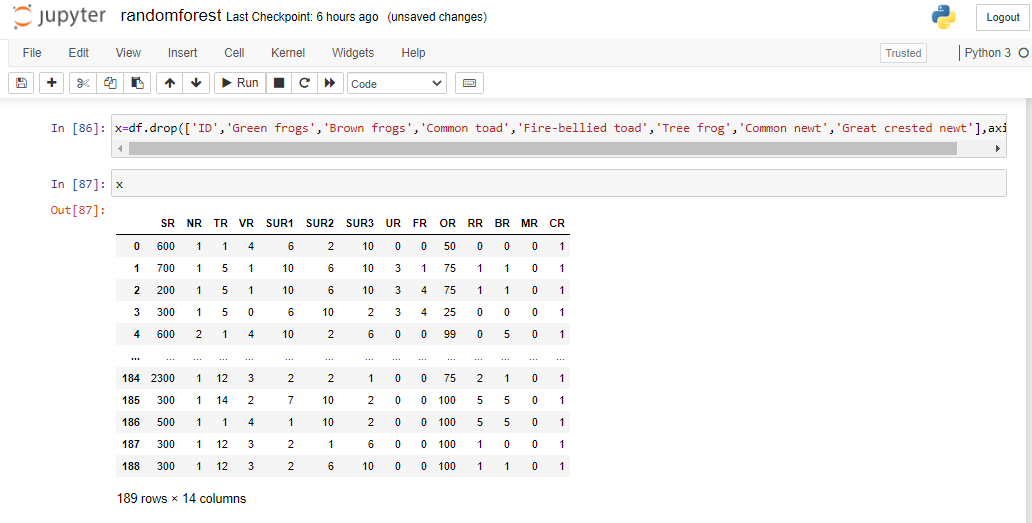
This drop function is used to specify the labels of various types of rows and columns within this given dataset. In that case, it is also helping to remove the null values of rows and columns from the selected database. In this case, there are three steps are present here within this functions and these steps are given below,

|  |  |
| --- | --- |
| Step 1 | Syntax |
| Step 2 | Parameter |
| Step 3 | Return type |



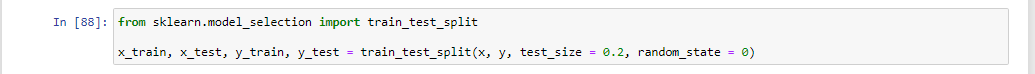
**Figure 4: code-4**

In the above figure, the y value is being settled which contains multiple columns. As this y value contains multiple columns these columns are inputted all at once with the help of []. The y value is the part that provides the result of the calculation



**Figure 5: code-5**

In this figure, the x values are setted. The x values are those values which are to be calculated into order to get the results. In the above figure all the y values are dropped by using the drop function. Apart from the y values the ID column is also dropped as the values of the ID will also be calculated.



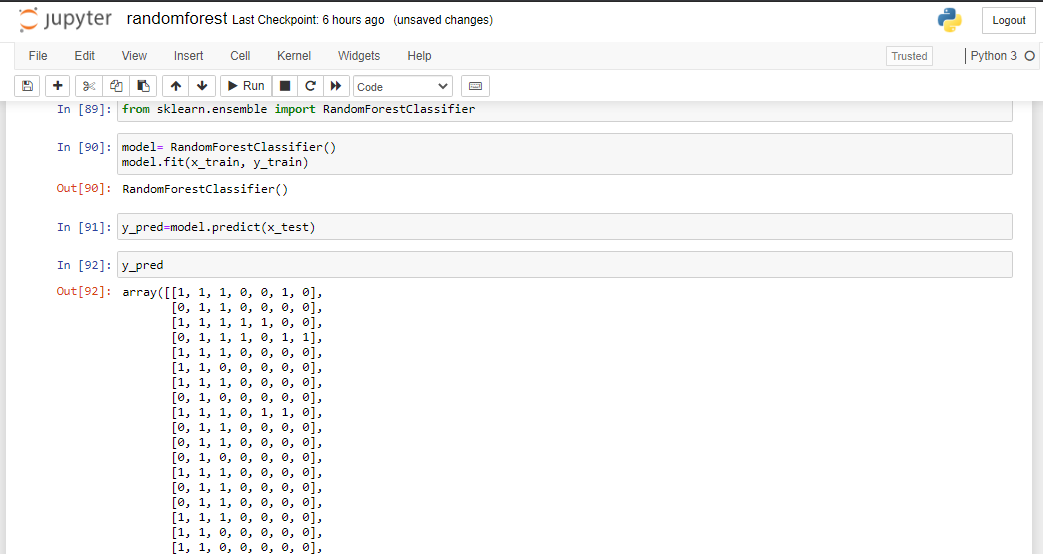
**Figure 6: code-6**

In the above figure, a function of the sklearn library has been inputted into the process. This sklearn library contains a huge number of functions and processes that allows to perform different types of analysis into the dataset. One of the first and most important parts in the data evaluation process is splitting the dataset. The dataset is to be split into two parts or models named train and test. In order to perform this process, the train test split function is inputted. This function splits the data into two different models and as the y and x are defined in the previous steps the respective x and y values will also get generated. After the process, the train model will contain x train and y train, and the test model will contain x test and y test.

**Test train split in SK learn.**

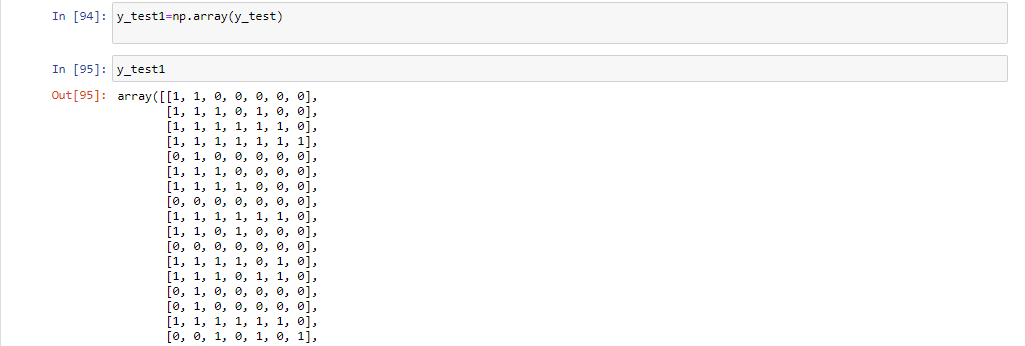
This train test split in SK learns model is used for splitting data arrays within two subsets and this function is also able to make a random partition for two subsets within the given database. In that case, there are some steps present here to split the entire dataset or column and these steps are given below,

|  |  |
| --- | --- |
| Step 1 | Import the entire selected database by using pandas |
| Step 2 | Split the database by using train\_test\_split from the SK learn |
| Step 3 | Controlling the size of two subsets with the help of parameters such as train\_size and test\_size |
| Step 4 | Determining randomness of the split data with the help of random\_state parameter |
| Step 5 | Obtaining a stratified split with stratify parameter |



**Figure 7: code 7**

In the above figure, it has been inputted from the sklearn library which is a random forest classifier. This function allows the categorical data to get analyzed and evaluated. After importing the function into the process the next step is to input the function into a variable so that the variable can be inputted with multiple values that perform as per the function process. After the variable is created with the function the values of the x train and y train are inputted into the variable. After that a variable is created where the y test is predicted by using the predict function into the model. After that the values of y pred are provided.

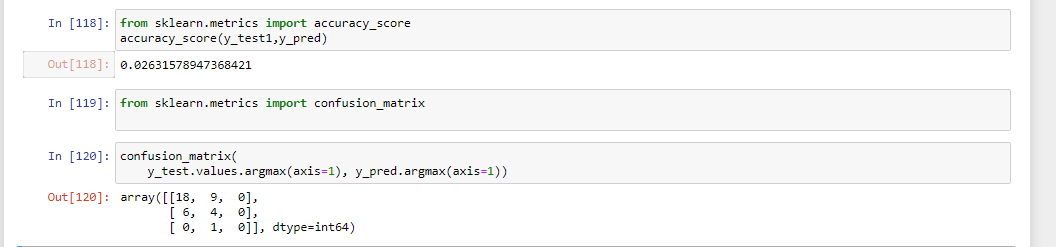


**Figure 8: code 8**

**Random forest classifier**

Random forest classifier is used to build multiple types of decision trees and also merge these decision trees to get appropriate and also stable predictions. This classifier is used for the classification and regression task for handling the missing or null values of the given database. In that case, this process is able to maintain and understand the accuracy of a complex proportion of data (Shang-Fu *et al.* 2021). There are various types of steps are present here to use this classifier and these steps are given below,

|  |  |
| --- | --- |
| Step 1 | Choosing random samples from the given database |
| Step 2 | Structuring a decision tree for every individual, sample to get an appropriate prediction result for every decision tree |
| Step 3 | Performing a vote for every predictive result of each decision tree |
| Step 4 | Selecting the prediction result as a final prediction of the given dataset |
| Step 5 | Analyzing the prediction method |



**Figure 9: code-9**

In the above figure, at first a function of the sklearn which is the accuracy score is inputted. This function has been inputted in order to check the accuracy score of the y pred and the y test1 variables. This process will provide the information about the predicted result and the real output results and provides the information of how much the accuracy is maintained. After getting the accuracy score confusion matrix is to be extracted. In order to extract the confusion matrix at first a function named confusion matrix is to be inputted.

**Accuracy score**

This accuracy score is used to compute the subset of accuracy. This accuracy score is defined as the percentage of correct prediction of a given dataset and it can be structured or calculated easily by dividing the number of correct predictions within the given database. In that case, accuracy is representing the number of accurate data instances over the total number of data instances.

**Confusion matrix**

The confusion matrix is known as a tabular summary of the correct and incorrect data within the data prediction by the classifier. In that case, it is used to evaluate the performance rating of the classification model via the structure of performance metrics such as recall, accuracy and precision (Lumbanraja*et al*. 2021). There are various types of forms are present here within this confusion matrix and these forms are given below,

|  |  |
| --- | --- |
| Step 1 | Accuracy= TP + TN / TP + TN + FP + FN |
| Step 2 | Misclassification = FP + FN / TP + TN + FP + FN |
| Step 3 | Precision = TP / TP + FP |

Graphical user interface, text, application

Description automatically generated

**Figure 10: code-10**

**Hyperparameter**

Hyperparameters is the parameters that can have a direct effect on machine learning algorithm training. As a result, it is critical to understand how to optimize them in order to achieve maximum performance.

* **Grid Search cv:** This is the traditional method of looking for the best performing. hyperparameter among a set of manually specified hyperparameters. Make use of that number.
* **Random Search cv:** Similar to grid search, uses random search instead of exhaustivesearch. When just a limited number of hyperparameters are used to optimize the algorithm, this can outperform grid search.

According to bergstra, Trials on a grid are less effective for hyper-parameter optimization than trials picked at random (Bergstra, 2012).now we are going to perform both random search cv and grid search cv to check execution time.

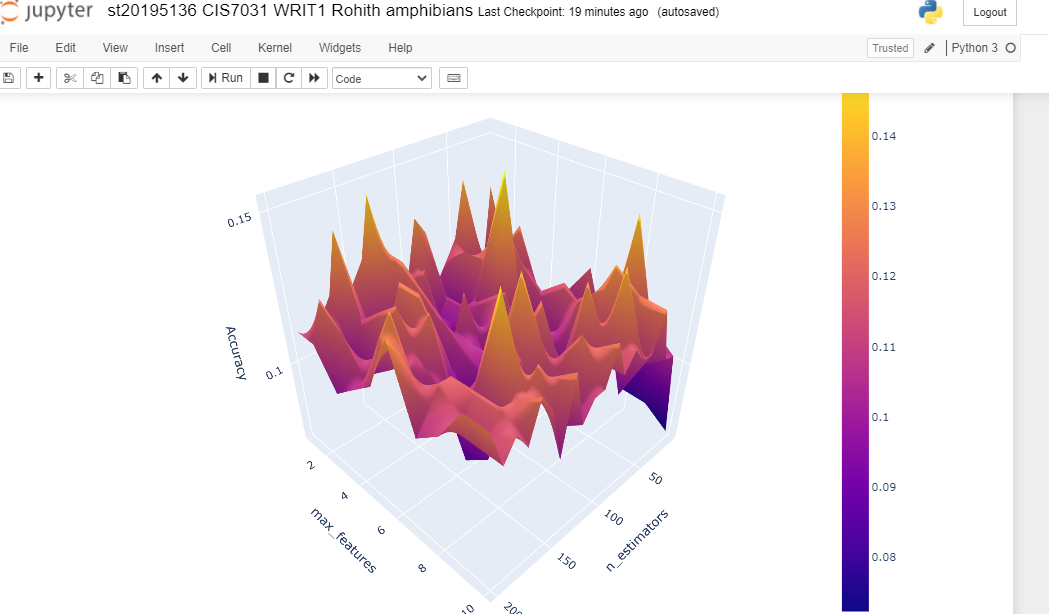
Graphical user interface, text, application, email

Description automatically generated

**Figure11:code-11**

Graphical user interface, text, application, email

Description automatically generated

**Figure12:code-12** 

**Figure 12: 3d surface plot**

# Critical review and analysis of techniques used.

The data evaluation and development is one of the most important and an essential part as this allows an individual or an organization to study the data and its performance. In this process, the data evaluation and development is done in a jupyter notebook that provides the input and output at the same platform and in doing this process python language is used. In order to perform the process the first and most important part is the dataset. In this process a dataset is identified and a proper study of the dataset is done so that the x and y values are to be identified as per the requirements. In order to input the dataset into the processes at first the dataset is to be inputted perfectly this can be done with the help of pandas library that inputs the dataset into a two dimensional table. This will make the dataset to be studied and understood properly. After checking the dataset the next step is to check for the null values. If the null values are present it has to be eliminated as this can make the calculation issue and can provide wrong results. After that the unwanted columns present in the dataset are removed which are the motorway and the ID. Fate removing the columns the next step is to input the values into the x and y. In x, the Calculation parts are to be stored and in y the results are to be stored. After the x and y are fixed the dataset is to be separated in two models named train and test. When the two models are ready the evaluation is to be started. Here the random forest classification is to be done as the Result is categorical. Accuracy score and the confusion matrix are provided here to make it more accurate.  we will be performing the tuning of hyperparameters on Random Forest model. The hyperparameters that we will tune includes **max features** and the **n estimator. At** this stage we will compare execution time both randomized search cv and grid search cv. lastly, we found that execution time taken by randomized search cv is less than grid search cv. after that we will be exporting the grid search parameters and their resulting accuracy scores into a dataframe and prior to making contour plots, we will have to reshape the data into a compatible format so that will be recognized by the contour plot functions. To start plotting 3D surface plot, we'll divide the data into groups using Pandas' groupby() method and the two hyperparameters max features and n estimators. The data is reshaped by pivoting it into a m by n matrix, with max features and n estimators as rows and columns, respectively. Let's give the plot a third dimension, and we'll have a three-dimensional surface plot. You may rotate the graph, which is a neat feature of this display.

we found that model is **Overfitting**, because it is performing good on training set data (**100%**), but it failed to perform well on the test set data.

# Conclusion

According to this assignment, it is based on the data prediction of the given database. All the data preparation and exploration is already discussed in the body of the report to complete this assignment. The entire development of prediction methods is also discussed within this assignment to do the proper data prediction of this given database. All the methods and functions such as pandas, NumPy, info, random forest classifiers are also described in this assignment. The accuracy score, the confusion matrix, and also the hyperparameters is already discussed within this assignment. The entire analysis and critical evaluation part are also discussed in the body of this report for accurate data prediction. All the screenshots of this data analysis are already given in this critical evaluation and analysis part to complete this assignment.

# References

Konstantinou, N., Abel, E., Bellomarini, L., Bogatu, A., Civili, C., Irfanie, E., Koehler, M., Mazilu, L., Sallinger, E., Fernandes, A.A.A., Gottlob, G., Keane, J.A. And Paton, N.W., 2019. VADA: an architecture for end user informed data preparation. *Journal of Big Data,***6**(1), pp. 1-32.

Lumbanraja, F.R., Fitri, E., Ardiansyah, Junaidi, A. And Prabowo, R., 2021. Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science Journal). *Journal of Physics: Conference Series,***1751**(1),

Shang-Fu, C., Dias, R., Evans, D., Salfati, E.L., Liu, S., Wineinger, N.E. And Torkamani, A., 2020. Genotype imputation and variability in polygenic risk score estimation. *Genome Medicine,***12**, pp. 1-13.

Zadeh, A.H., Schiller, S., Duffy, K. And Williams, J., 2018. Big Data and The Commoditization of Analytics: Engaging First-Year Business Students with Analytics. *The e - Journal of Business Education & Scholarship of Teaching,***12**(1), pp. 120-137.

Zhou, T., Chen, L., Guo, J., Zhang, M., Zhang, Y., Cao, S., Feng, L. And Wang, H., 2021. MSIFinder: a python package for detecting MSI status using random forest classifier. *BMC Bioinformatics,***22**, pp. 1-14.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research 13, 281–305.