

**Micro-Credit Defaulter Model**

Submitted by:

M.KAVITHA

**ACKNOWLEDGMENT**

I HAVE DONE THIS PROJECT AS PER MY KNOWLEDGE LEARNT FROM THE COURSE.

INTRODUCTION

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income.

* Business Problem Framing

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

* Conceptual Background of the Domain Problem

Data science can be used to assist minimize costs, discover new markets and make better decisions. Understanding the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They arecollaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days.

* **Review of Literature**

There are many algorithms and methods that data experts analyse to help the managers and directors extend their companies bottom line and strategic positioning. Data science techniques are incredible at spotting abnormalities, optimizing constraint problems, predicting and targeting.

* **Motivation for the Problem Undertaken**

With better tools, comes better understanding of how we should be doing analytics and how we should use the resulting information, all of which is evolving extremely fast. It satisfies curiosity.

What really motivate me to be a data analyst, is the potential I usually see in the big data which is currently available in the internet. ... In addition to job creation, another motivating factor is my love for technology and doing research**.**

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem
* We will explore a Data set and perform the exploratory data analysis. The major topics to be covered are below:
* – Handle Missing value  
  – Removing duplicates  
  – Outlier Treatment  
  – Normalizing and Scaling( Numerical Variables)  
  – Encoding Categorical variables( Dummy Variables)  
  – Bivariate Analysis

* Data Sources and their formats

#### Basic Data Exploration

In this step, we will perform the below operations to check what the data set comprises of. We will check the below things:

– head of the dataset  
– the shape of the dataset  
– info of the dataset  
– summary of the dataset

1. The head function will tell you the top records in the data set. By default, python shows you only the top 5 records.

The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The cars data set has 209593

observations and 37 variables in the data set.

info() is used to check the Information about the data and the datatypes of each respective attribute.

**Data Preprocessing Done**

* Data Analysis: Data Analysis is the statistics and probability to figure out trends in the data set. It is used to show historical data by using some analytics tools. It helps in drilling down the information, to transform metrics, facts, and figures into initiatives for improvement.
* Data Inputs- Logic- Output Relationships
* The described method will help to see how data has been spread for numerical values. We can clearly see the minimum value, mean values, different percentile values, and maximum values.
* State the set of assumptions (if any) related to the problem under consideration

Since we have number of duplicate rows: (1, 36)

we will remove this from the data set so that we get only distinct records. Post removing the duplicate, we will check whether the duplicates have been removed from the data set or not

.Hardware and Software Requirements and Tools Used

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **random**

**import** **matplotlib.pyplot** **as** **plt**

%matplotlib inline

**import** **warnings**

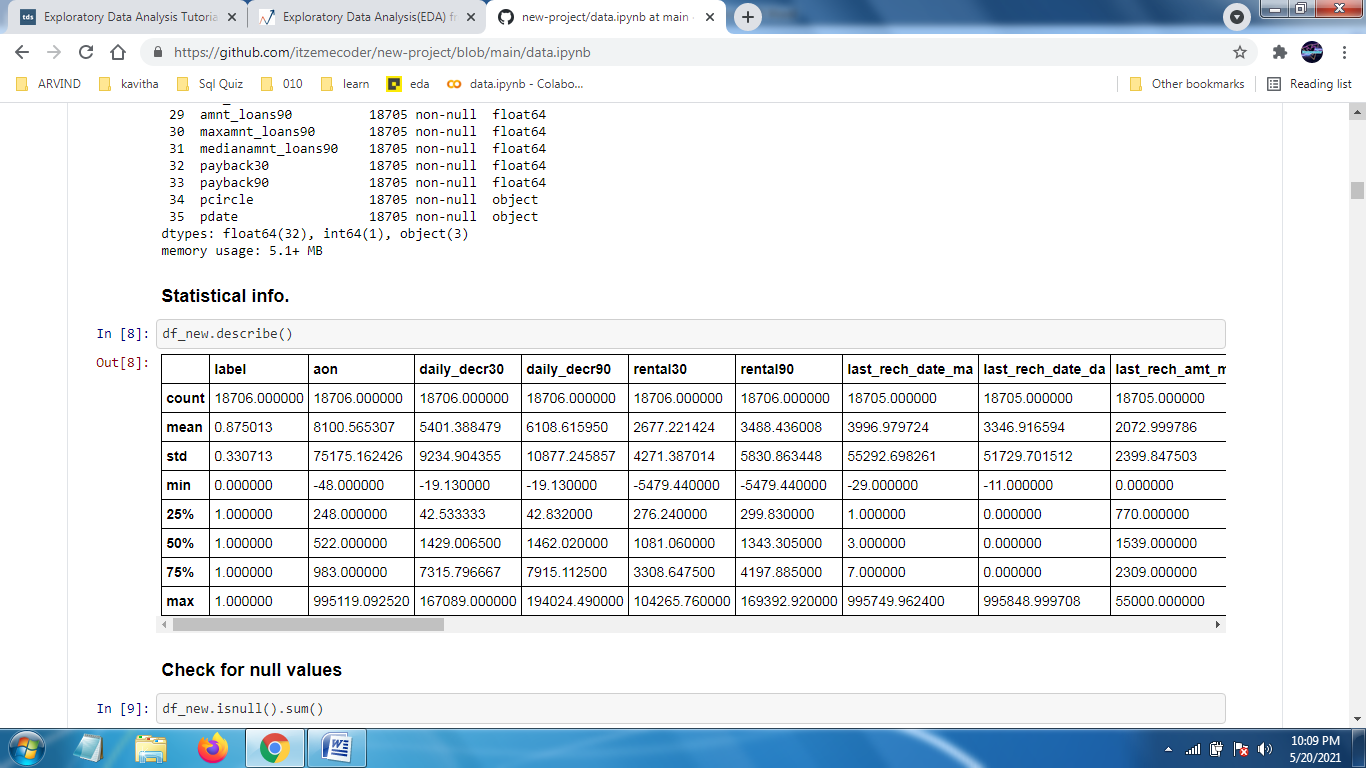
warnings.filterwarnings("ignore")

**from** **sklearn.metrics** **import** classification\_report, confusion\_matrix

**from** **sklearn.neural\_network** **import** MLPClassifier

**from** **sklearn.model\_selection** **import** GridSearchCV

**Model/s Development and Evaluation**

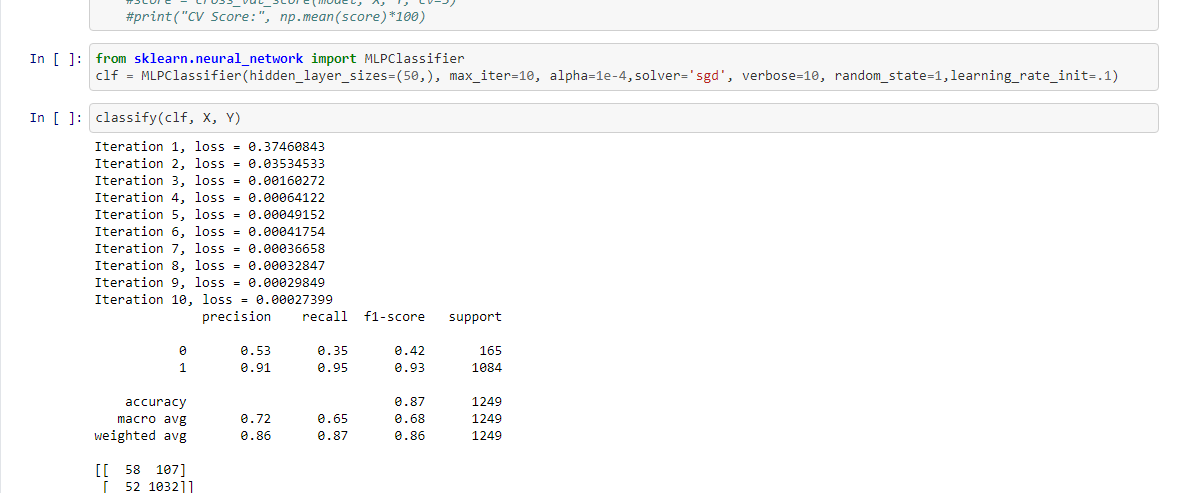
* Identification of possible problem-solving approaches (methods)
* Statistical Calculation: Mean is less than the median.
* ****
* Testing of Identified Approaches (Algorithms)

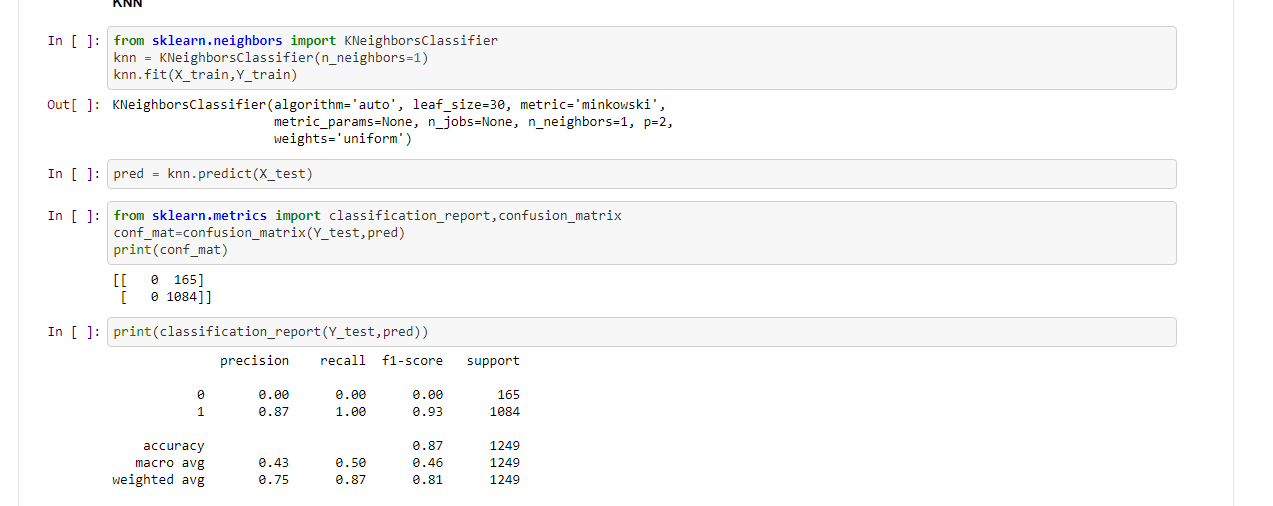
Algorithms used for the training and testing. :

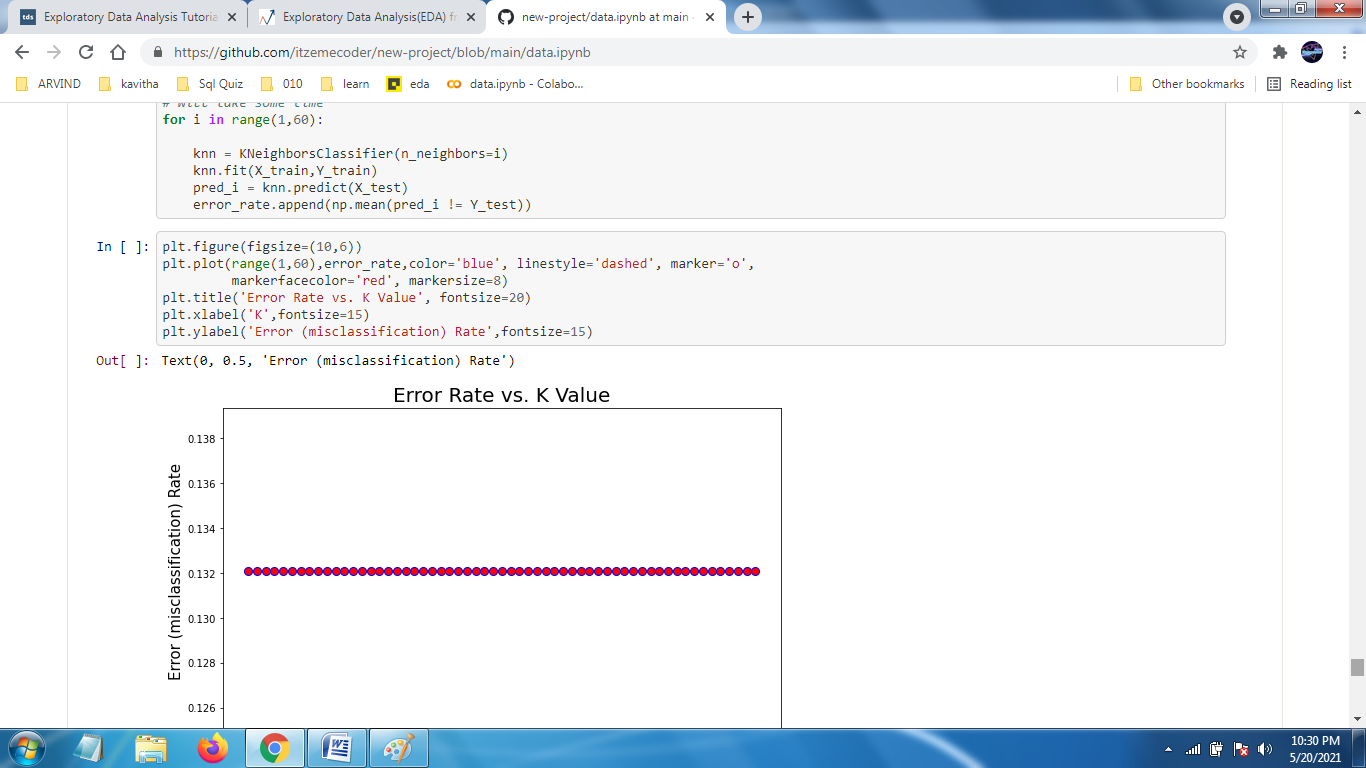
* K Nearest Neighbour Regressor
* Neural Network Classifier
* Random Forest Regressor
* Run and Evaluate selected models:

### Neural network classifier





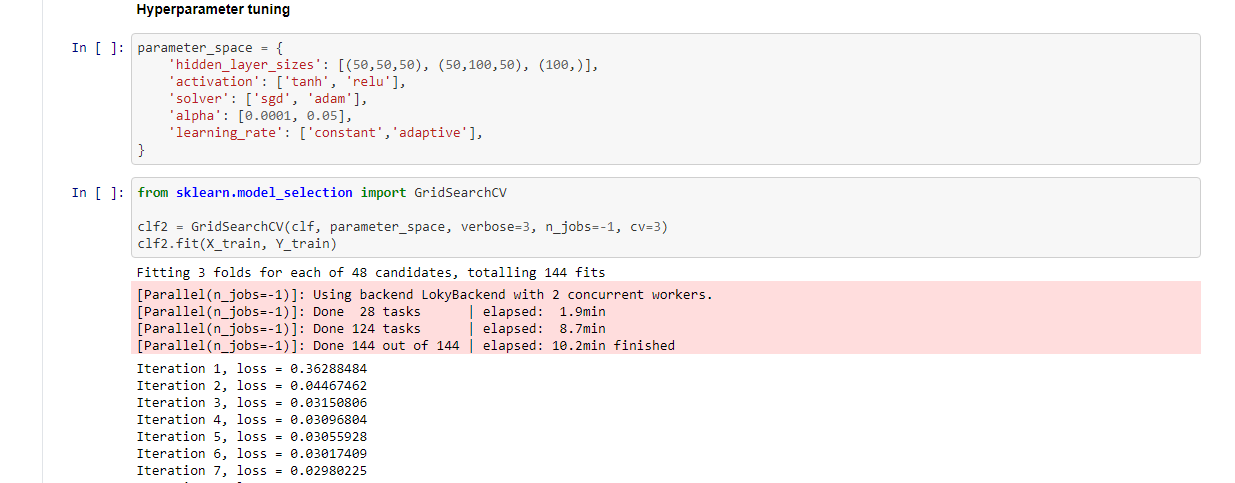






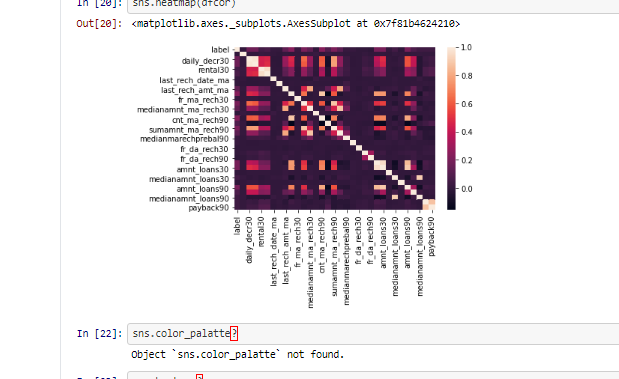
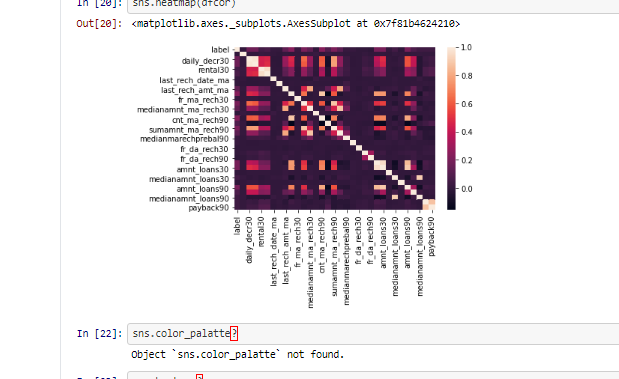
* Key Metrics for success in solving problem under consideration

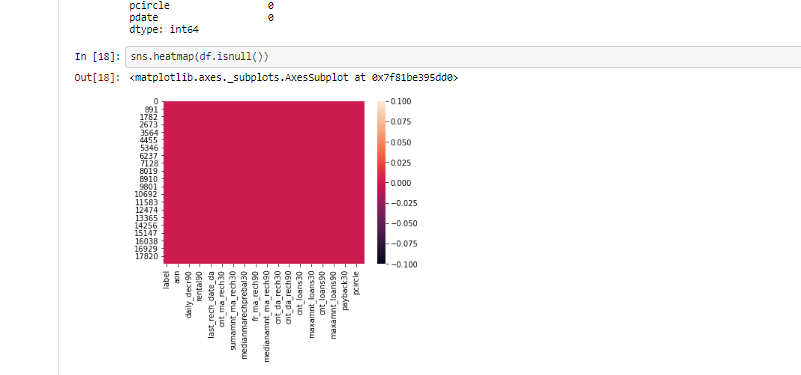
What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

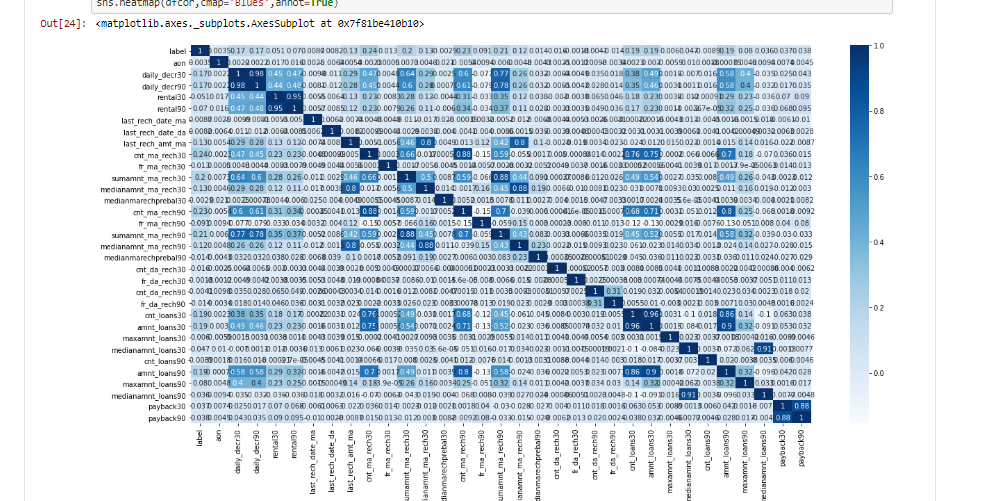




* Visualizations

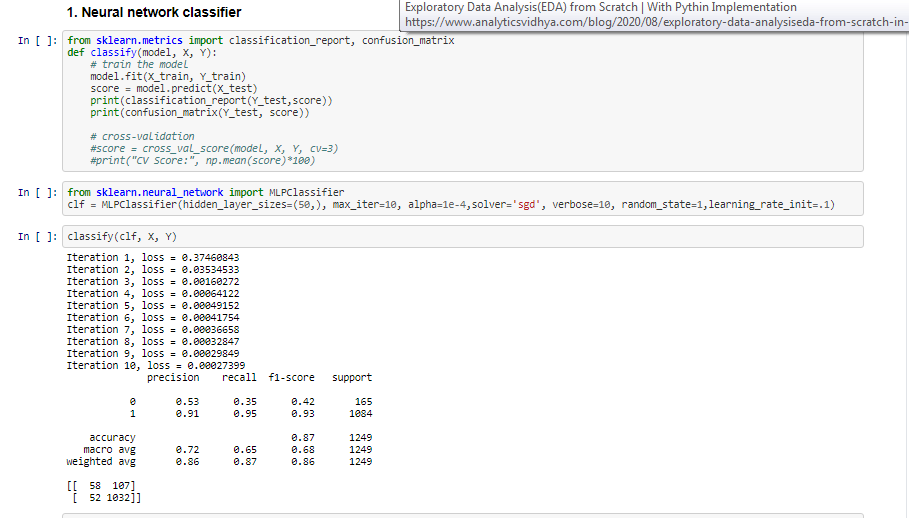






**CONCLUSION**

* Key Findings and Conclusions of the Study
* Data Analysis: Data Analysis is the statistics and probability to figure out trends in the data set. It is used to show historical data by using some analytics tools. It helps in drilling down the information, to transform metrics, facts, and figures into initiatives for improvement.
* K nearest neighbour regressor
* Neural network classifier
* Random forest regressor

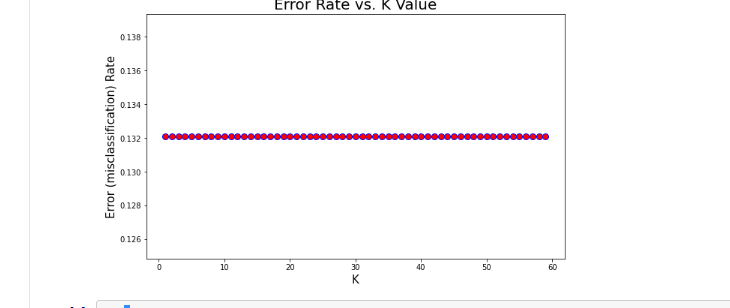


Models are commonly evaluated using resampling methods like k-fold cross-validation from which mean skill scores are calculated and compared directly. Although simple, this approach can be misleading as it is hard to know whether the difference between mean skill scores is real or the result of a statistical fluke.

Statistical significance tests are designed to address this problem and quantify the likelihood of the samples of skill scores being observed given the assumption that they were drawn from the same distribution. If this assumption, or null hypothesis, is rejected, it suggests that the difference in skill scores is statistically significant.

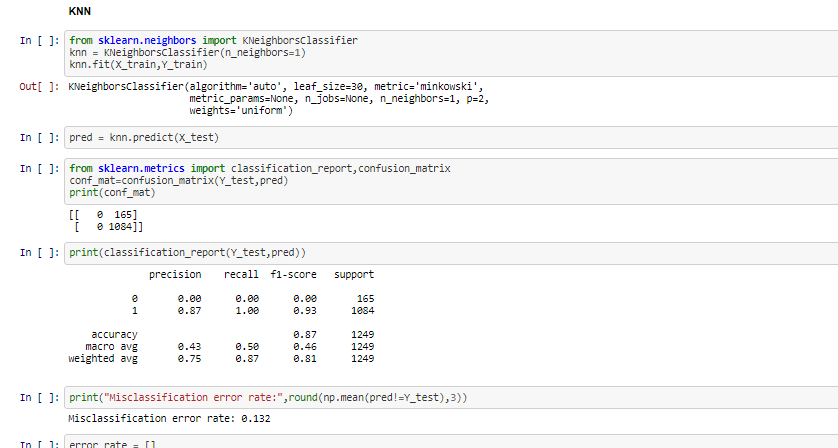
The assumptions of the paired t-test are violated in the case of random resampling and in the case of k-fold cross-validation (as noted above). Nevertheless, in the case of k-fold cross-validation, the t-test will be optimistic, resulting in a higher Type I error, but only a modest Type II error. This means that this combination could be used in cases where avoiding Type II errors is more important than succumbing to a Type I error.

Misclassification error rate: 0.132



MSE:0.12395916086920328

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.



Misclassification error rate: 0.132

**Random forest regressor**

Generally, Random Forests produce better results, work well on large datasets, and are able to work with missing data by creating estimates for them. However, they pose a major challenge that is that they can’t extrapolate outside unseen data.

****

**{'max\_depth': 15, 'n\_estimators': 30}**

**MSE:0.07328779931393382**