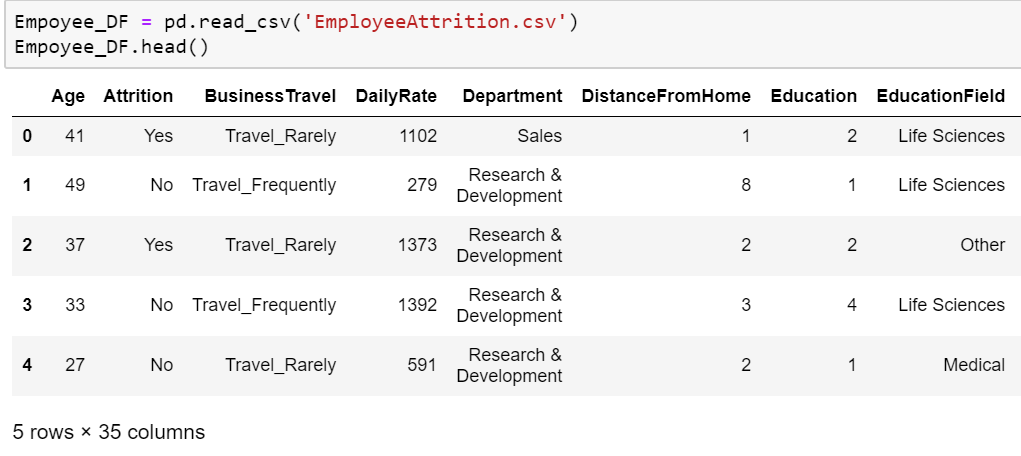
Predicting Employee Attrition

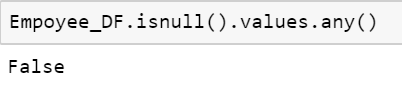
Done By - Group 3 (Nagesh, Saioni, Sakshi, Srishti, Alekhya, Vishnu, Aditya)

**Exploratory Data Analysis**

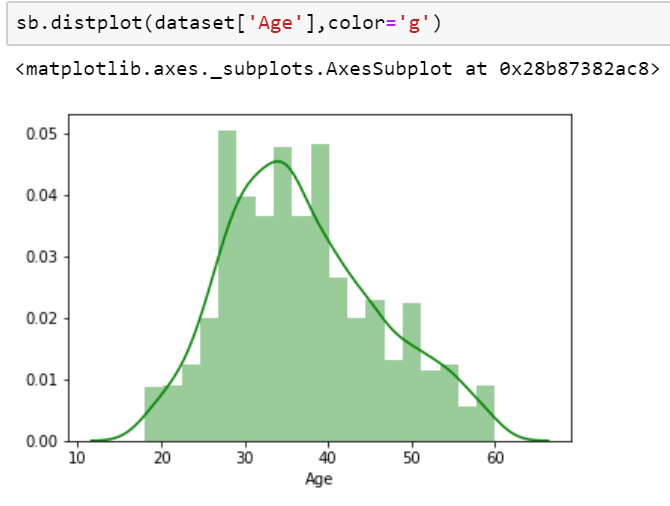
Reading the Dataset, the dataset consists of 1470 rows and 35 columns



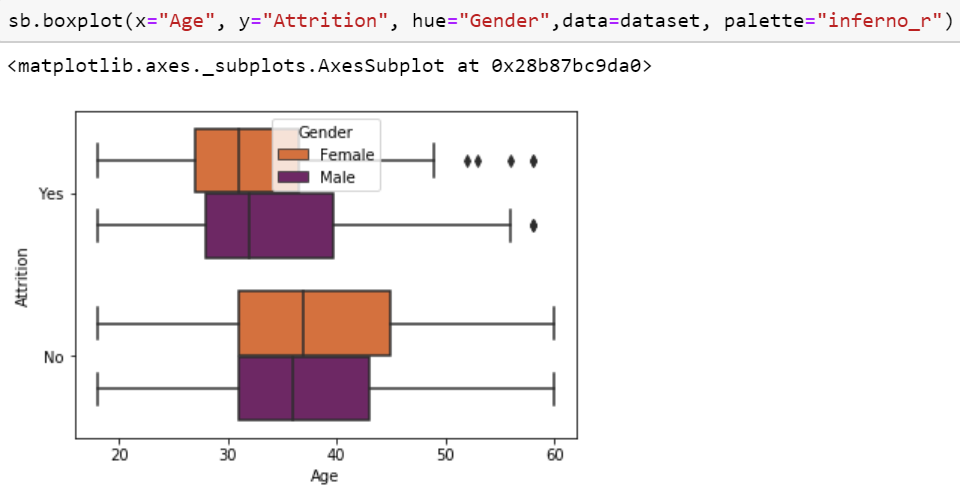
Checking if any of values are Null



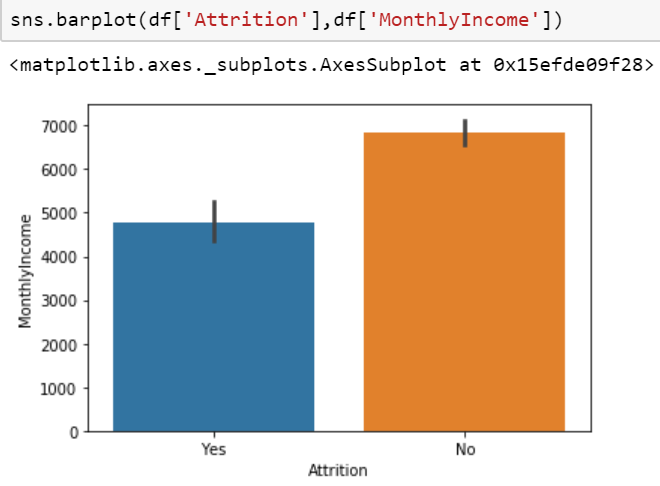
We notice here that most of the employees working here are within 25-40 years age



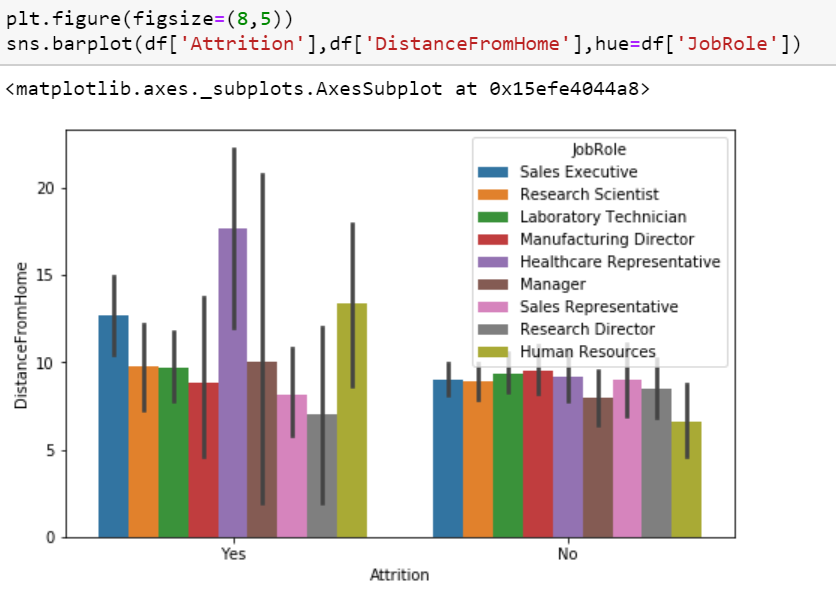
Attrition is more likely for the male employees around 30 years of age



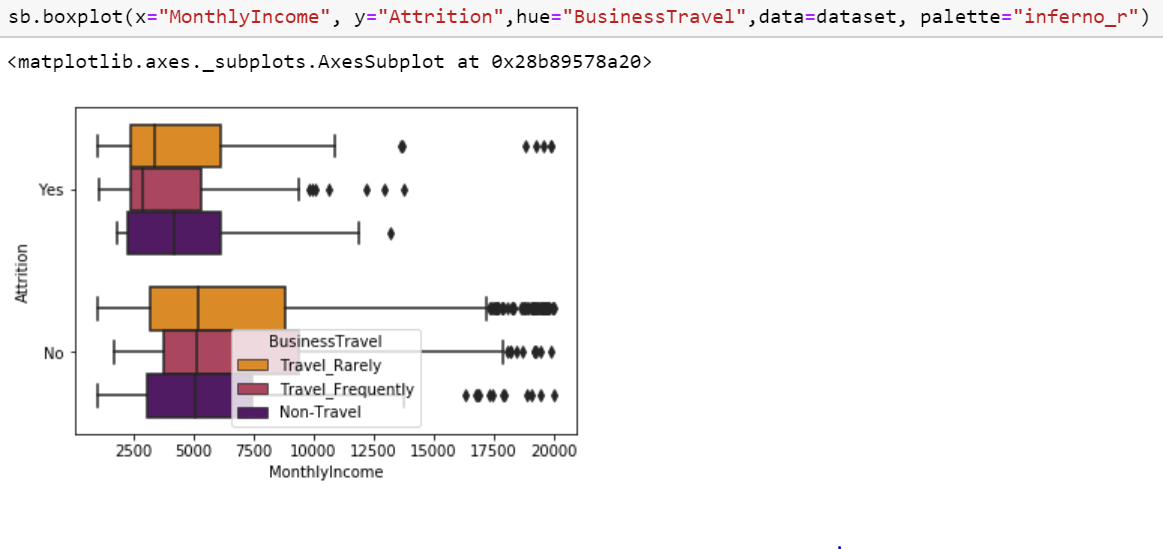
Employees with lower monthly income tend to leave more



Employees from HR department who’s staying far tend to leave



Employees tending to leave has comparatively lower monthly income and travel rarely or not travel at all

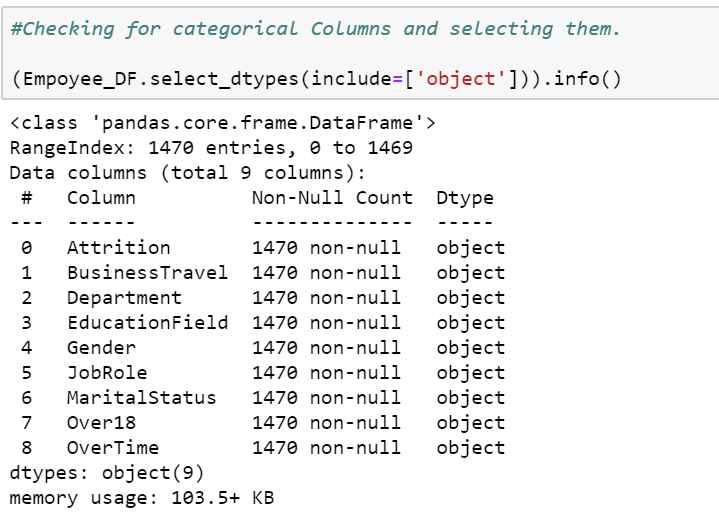




**Conclusions drawn from EDA**

* Most of the employees working in the company are within 25-40 years age
* Younger employees are likely to leave the company
* Attrition is more likely for males with age around 30 years
* Attrition is more likely for employees who travel rarely or not travel at all
* Attrition is higher for employees with higher education
* Employees with low job involvement tend to leave more
* Attrition likely for employees working overtime
* Employees having monthly income less than 5000 tend to leave
* Most of the employees reside 1-5 km far from the office
* Attrition is high for who employees who’s a healthcare representative, in Marketing education field
* Employees with more than 8 years of experience tend to leave

**Pre – Processing**



All the values in this column were same, so this attribute won’t help in predicting attrition

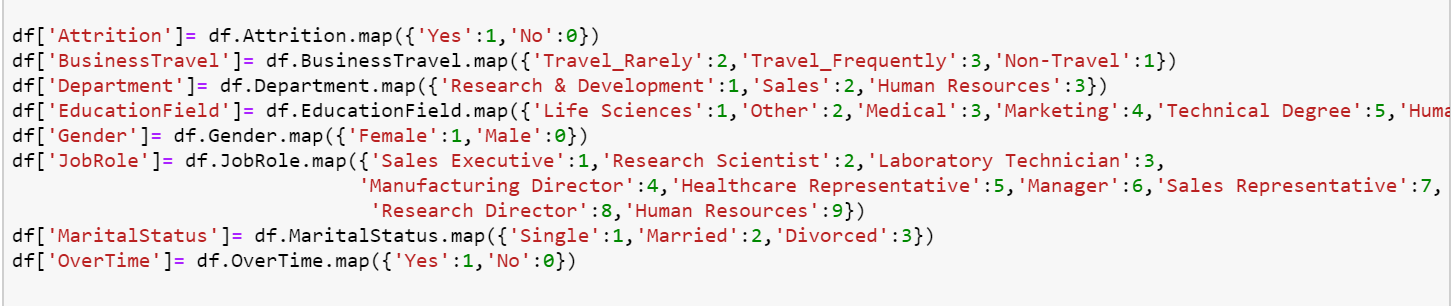


Age attribute gives much more valid and required information than over18

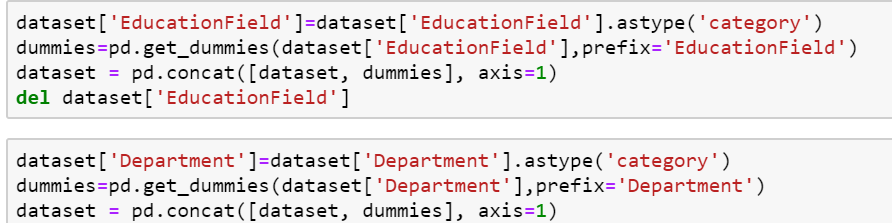


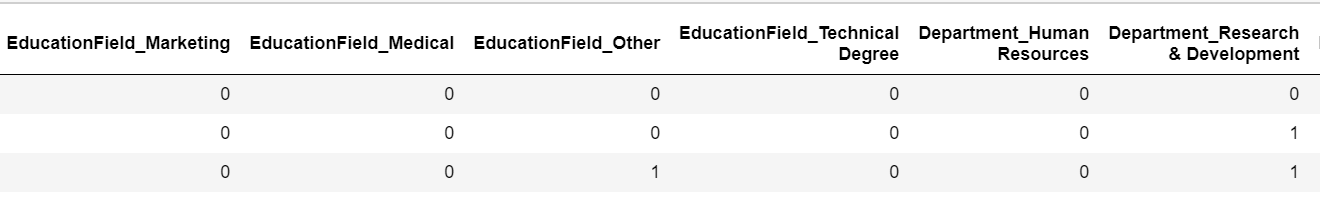
Converting categorical data to numerical one’s

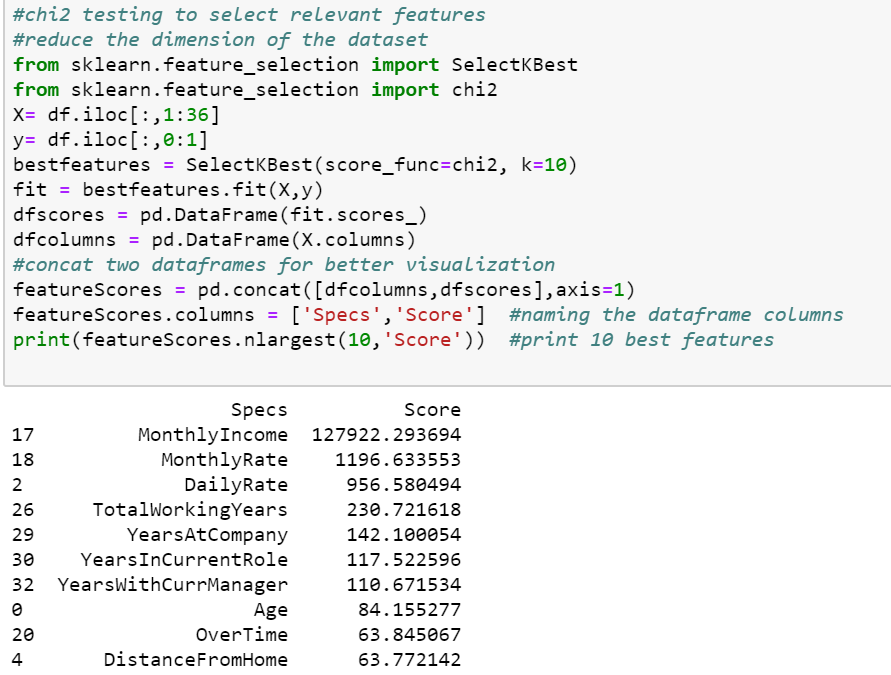
(One of the approaches)



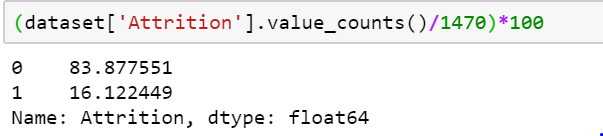
One hot encoding (Another Approach)







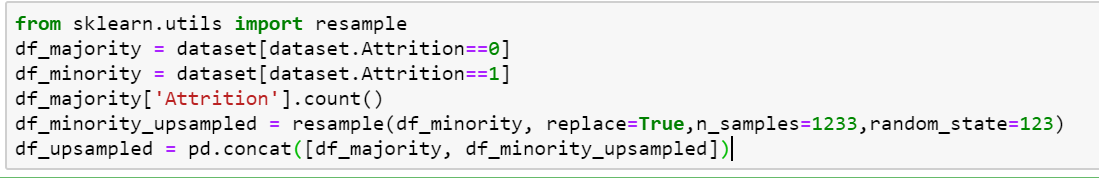
Its noticed here that our target attribute is highly imbalanced with ~84% of attrition value as Yes and only 16% as No

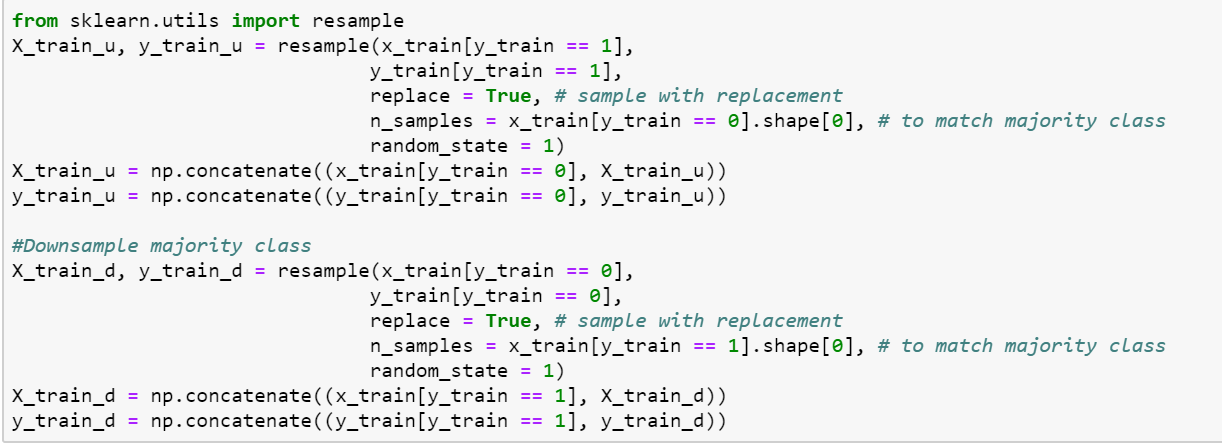


So, dataset needs to be balanced before modelling

Up-sampling the dataset,

Up scaling the minority class to equal the majority class and then combine the majority class with up-sampled minority class

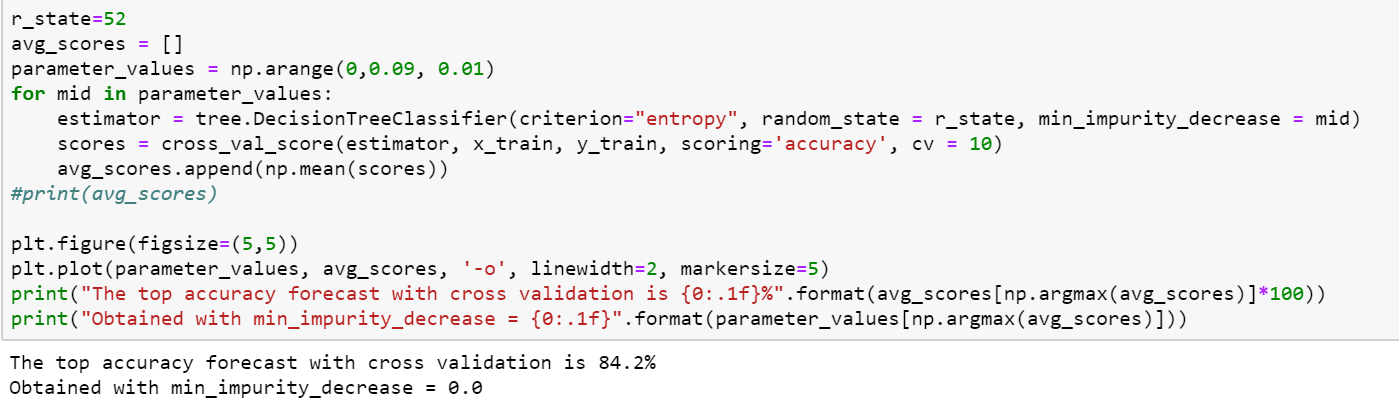


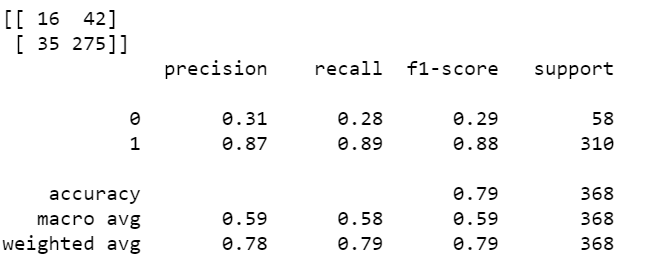


**Modelling**

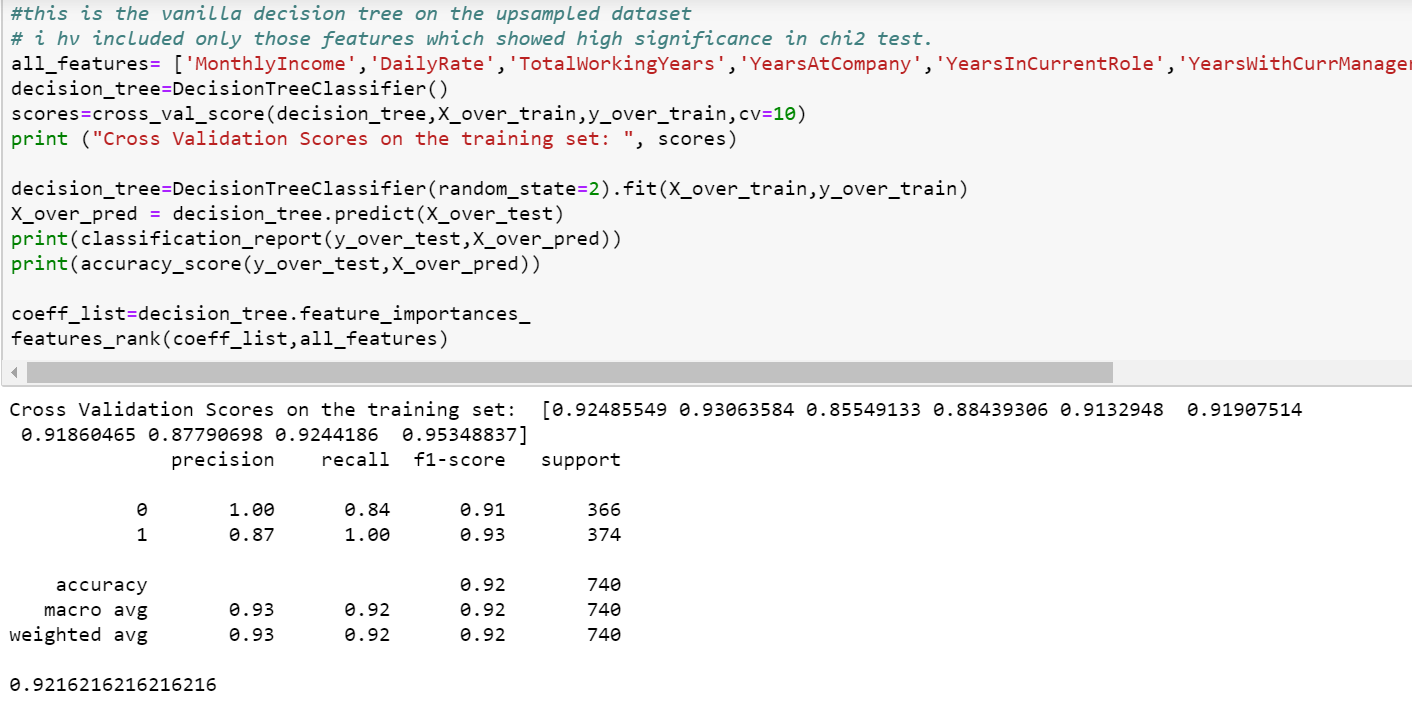
* **Decision Tree**

Modelling was carried out first on imbalanced one and accuracy achieved was ~79%, performed with cross-validation



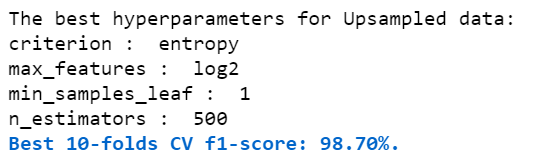


Vanilla decision tree on up-sampled dataset

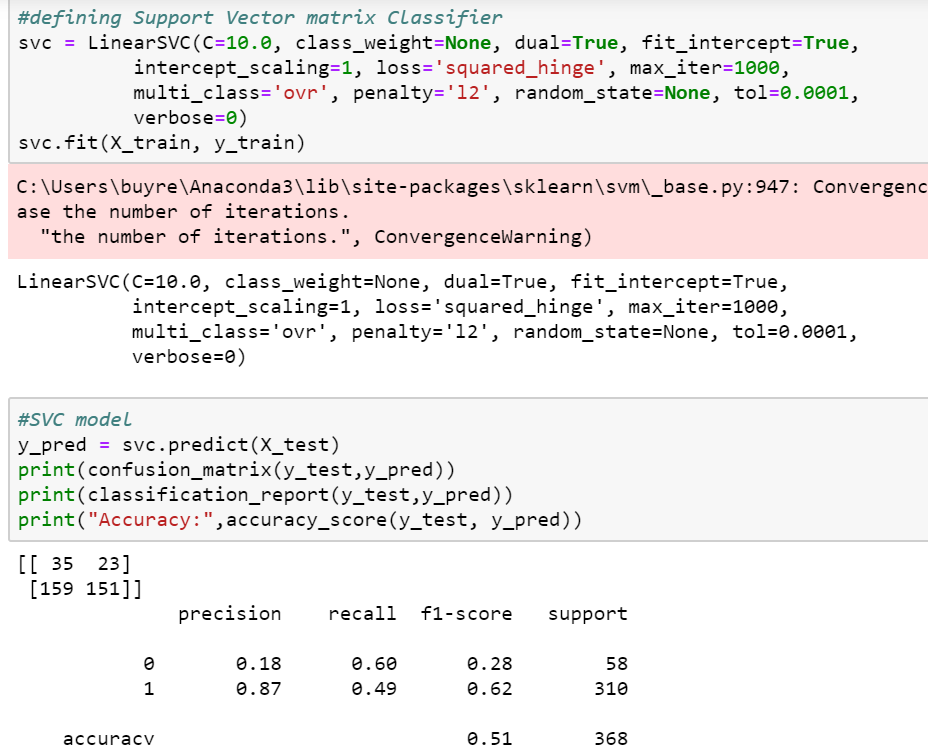


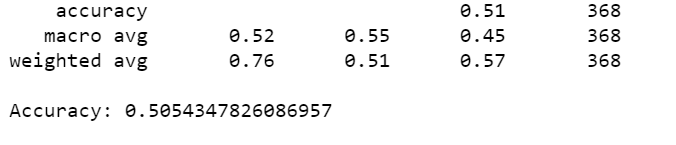
* **Random Forest**

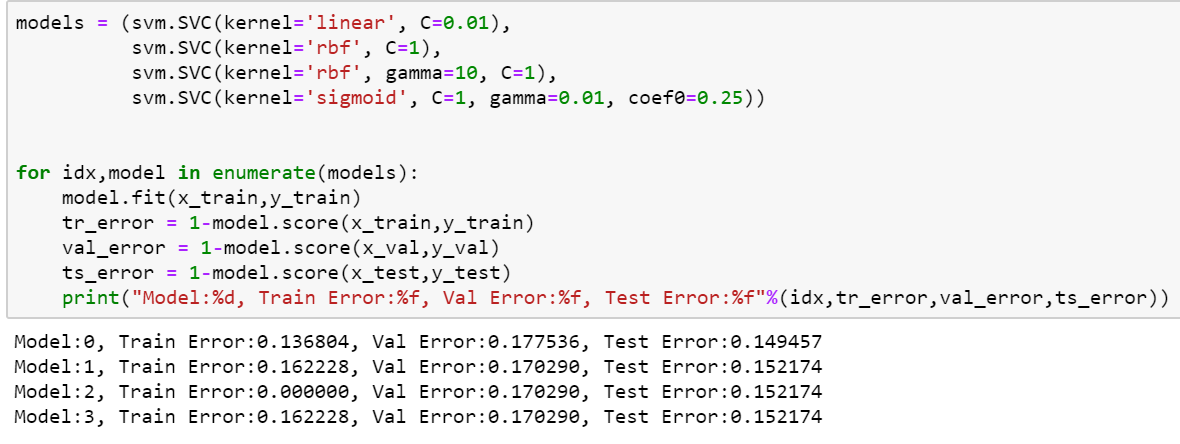
Pipeline was created here to run multiple processes (Standard Scalar and Random Forest). The purpose here was to assemble several steps that can be cross-validated. Grid search, a parameter tuning module that selects best parameters that would maximise the accuracy of the model



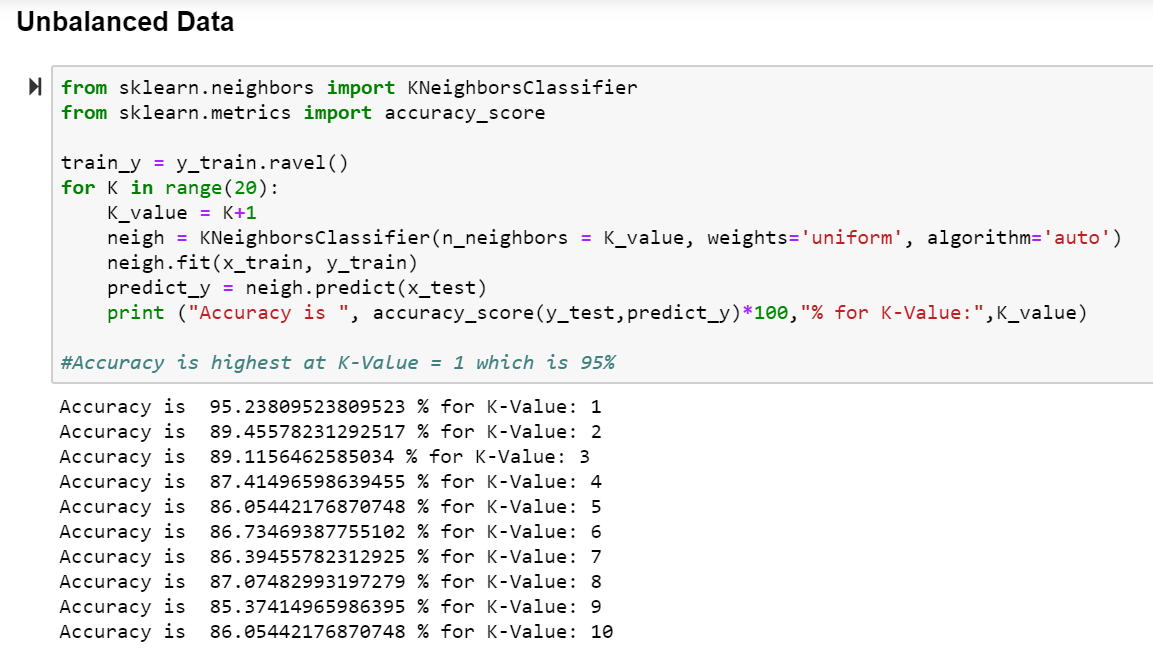
* **SVM**

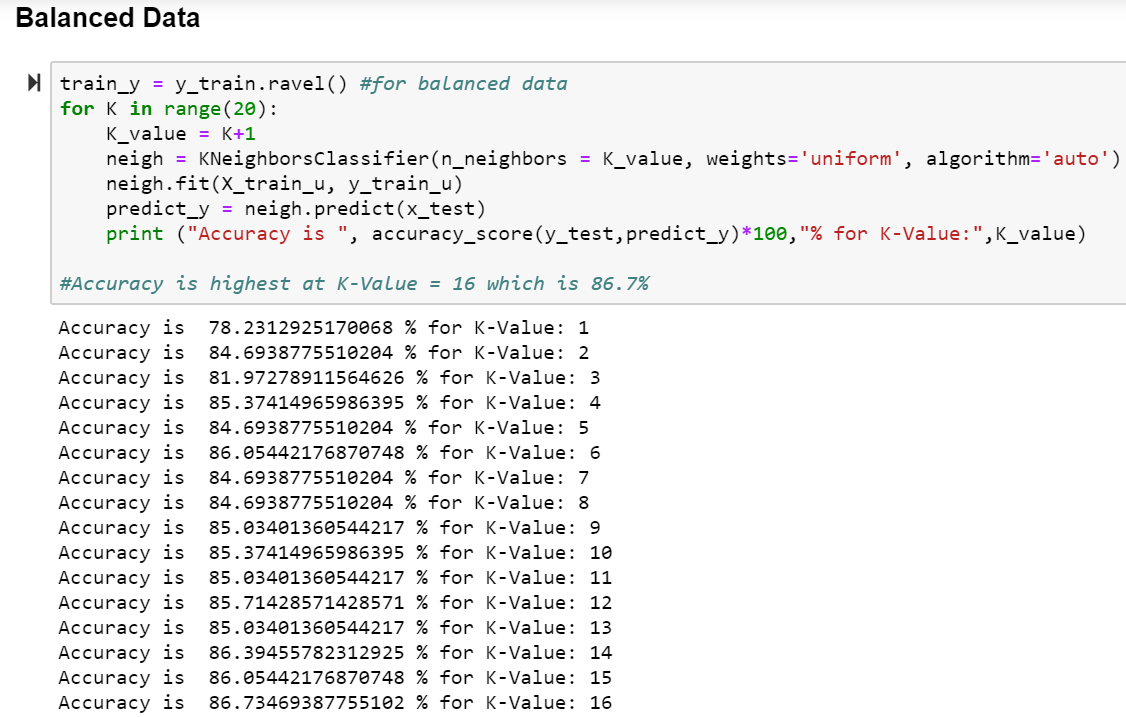




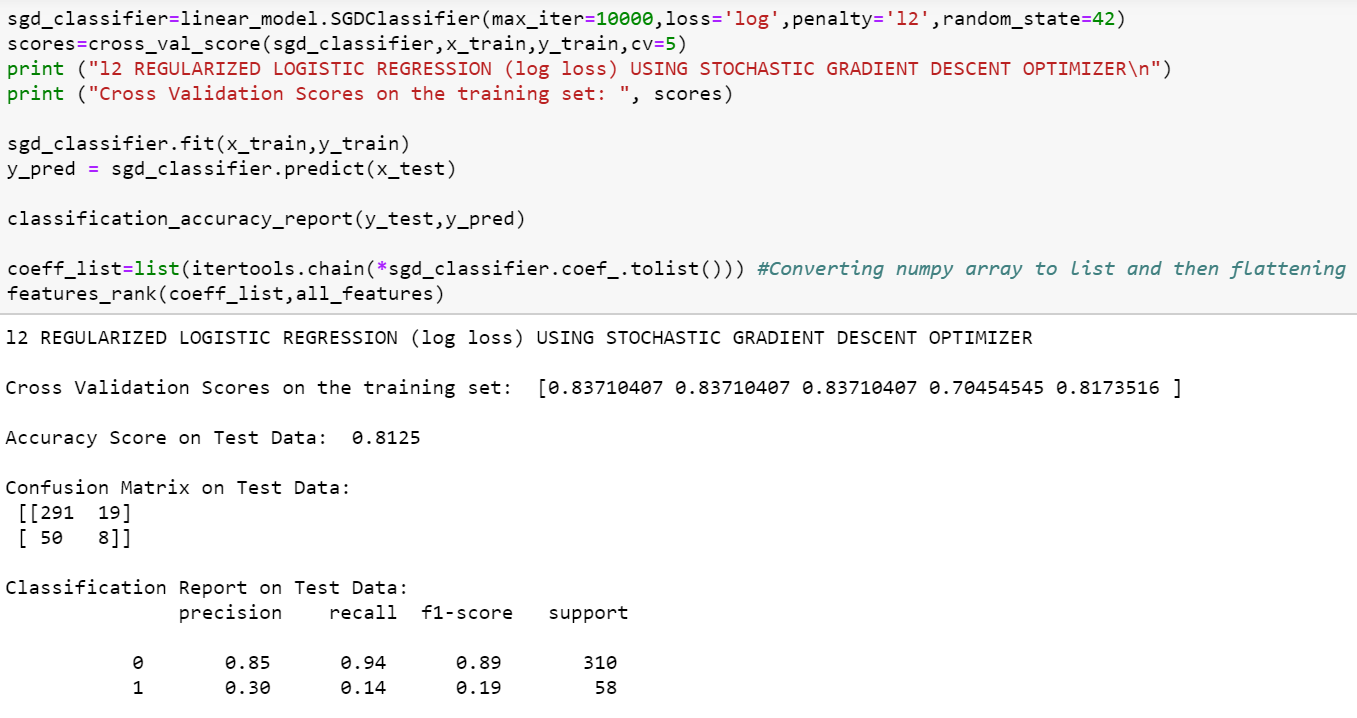


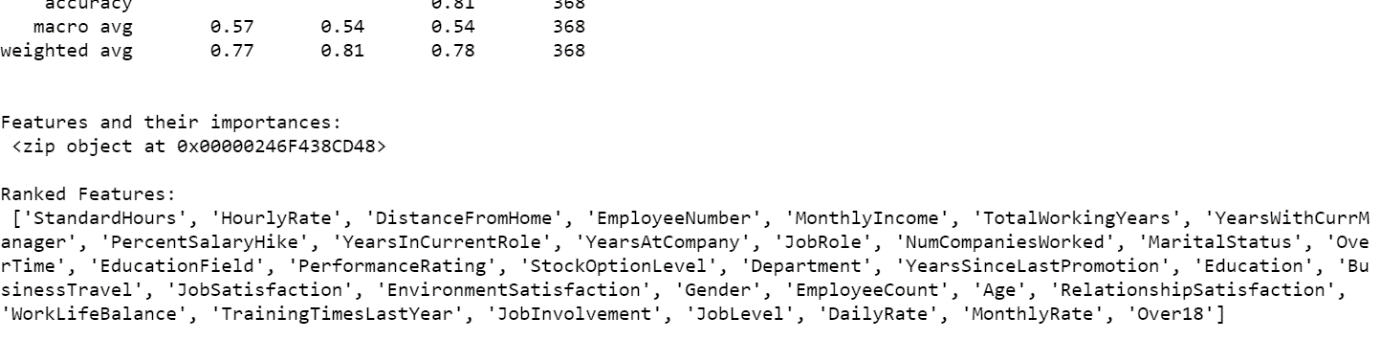
* **KNN**

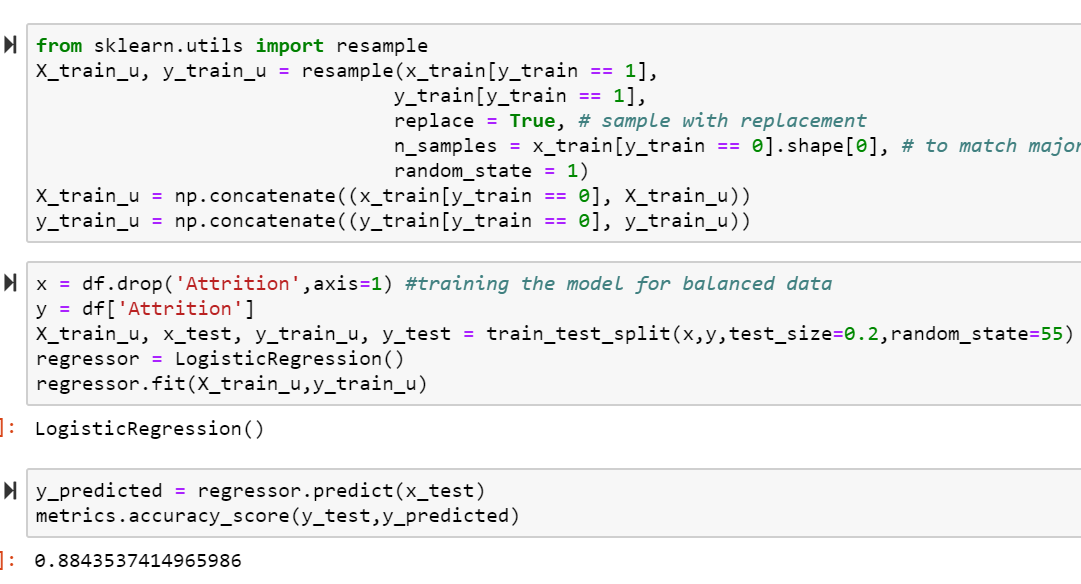




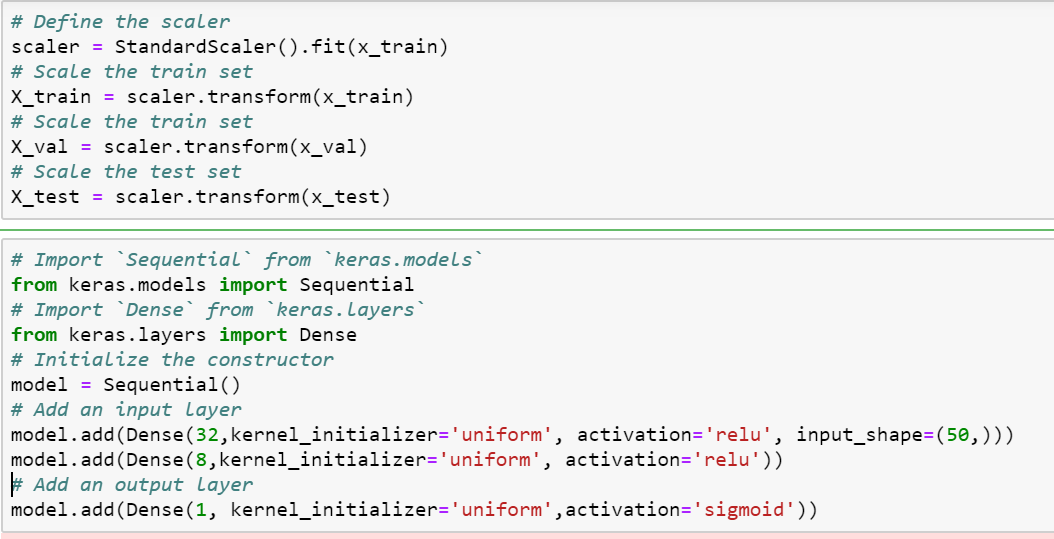
* **Logistic Regression**

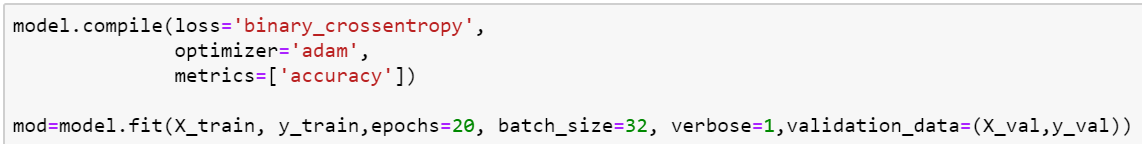


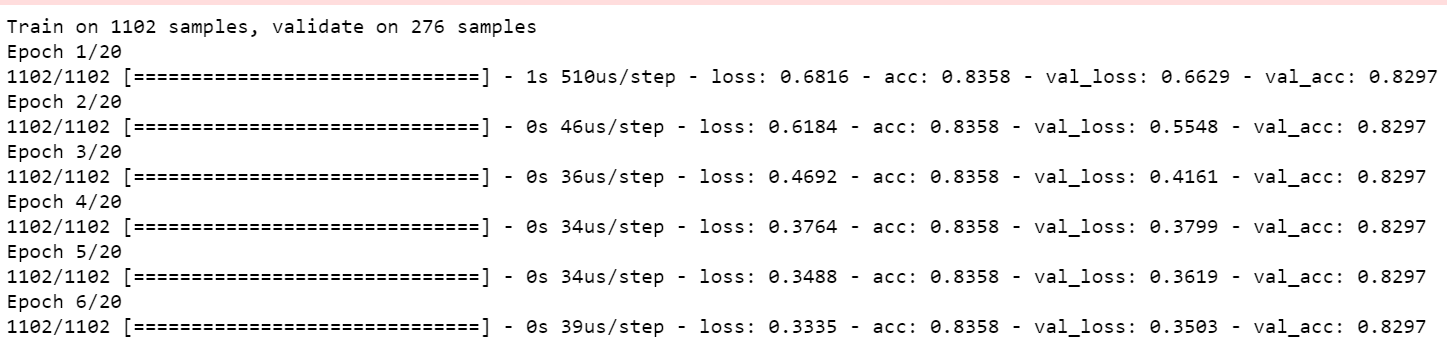


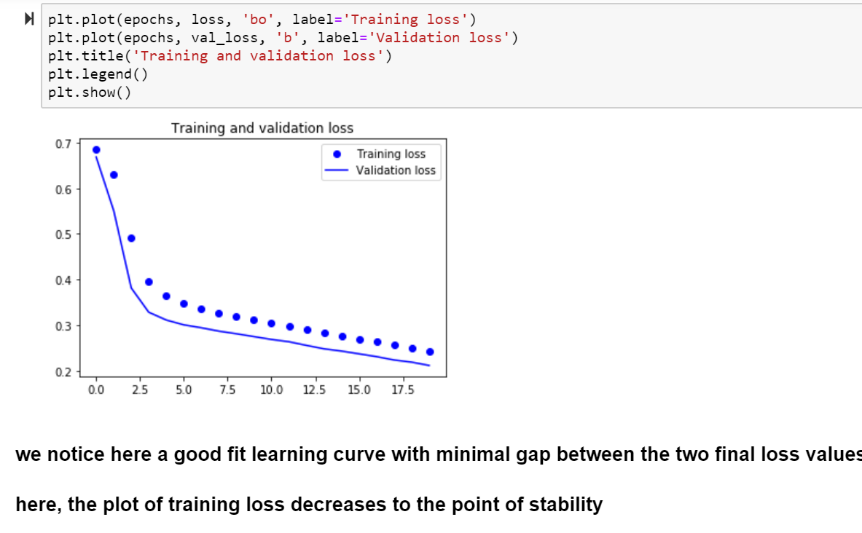


* **ANN**









Final Classification Report (Best accuracy values on balanced dataset is chosen)

