**­­**

**Predictive analytics on Indian firms most likely to be involved in audit fraudulent activity using supervised classification models.**

**Big Data**

**2020-2021**

*I declare that all work submitted for this coursework is the work of Vrinder Chandar alone except where explicitly stated otherwise.”*

**­­**

**Signed by Student: Vrinder Chandar Date: 12/01/2021**

**Table of Contents**

[**1 Introduction** 1](#_Toc61343740)

[**1.1 Overview** 1](#_Toc61343741)

[**1.2 Problem Description** 3](#_Toc61343742)

[**2 Construction and tuning of Machine Learning Model** 4](#_Toc61343743)

[**2.1 Data Analysis** 4](#_Toc61343744)

[**2.2 Data Exploration** 8](#_Toc61343745)

[**2.3 Correlation Matrix** 9](#_Toc61343746)

[**2.4 Support Vector Machine (SVM)** 11](#_Toc61343747)

[**3 Testing Results** 13](#_Toc61343748)

[**3.1 Cross-Validation** 13](#_Toc61343749)

[**3.2 Confusion Matrix** 15](#_Toc61343750)

[**4 Discussion** 16](#_Toc61343751)

[**References** 18](#_Toc61343752)

**Table of Figures**

[**Figure 1: CSV Dataset upload** 4](#_Toc61343754)

[**Figure 2: First and last 5 rows of the dataset** 4](#_Toc61343755)

[**Figure 3: Data types of the dataset** 5](#_Toc61343756)

[**Figure 4: Drop column “LOCATION\_ID”** 5](#_Toc61343757)

[**Figure 5: Missing value in the ‘Money Value’ column** 6](#_Toc61343758)

[**Figure 6: Filing one null value with the mean** 6](#_Toc61343759)

[**Figure 7: No missing values in the dataset** 7](#_Toc61343760)

[**Figure 8: Histograms of all attributes** 8](#_Toc61343761)

[**Figure 9: Drop column ‘TOTAL’.** 9](#_Toc61343762)

[**Figure 10: Creation of correlation matrix** 9](#_Toc61343763)

[**Figure 11: Drop of “Detection\_Risk” column** 10](#_Toc61343764)

[**Figure 12: Data splitting into 60% training, 20% testing and 20% validation** 12](#_Toc61343765)

[**Figure 13: SVC model building** 12](#_Toc61343766)

[**Figure 14: Building Cross-Validation** 13](#_Toc61343767)

[**Figure 15: SVM model evaluation** 14](#_Toc61343768)

[**Figure 16: Confusion Matrix of SVM classifier** 15](#_Toc61343769)

[**Figure 17: SVC model building with linear kernel** 17](#_Toc61343770)

[**Figure 18: SVM model evaluation** 17](#_Toc61343771)

**Table of Tables**

[**Table 1: Audit Data Dataset** 1](#_Toc61343774)

[**Table 2: Major attributes of the dataset** 2](#_Toc61343775)

# 

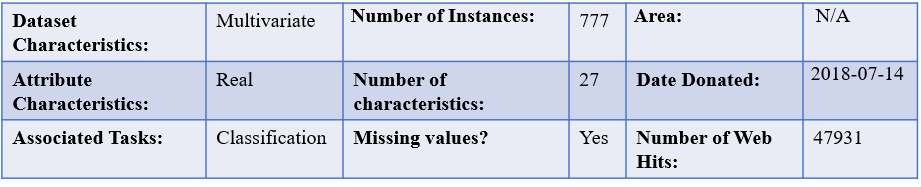
# **1 Introduction**

## **1.1 Overview**

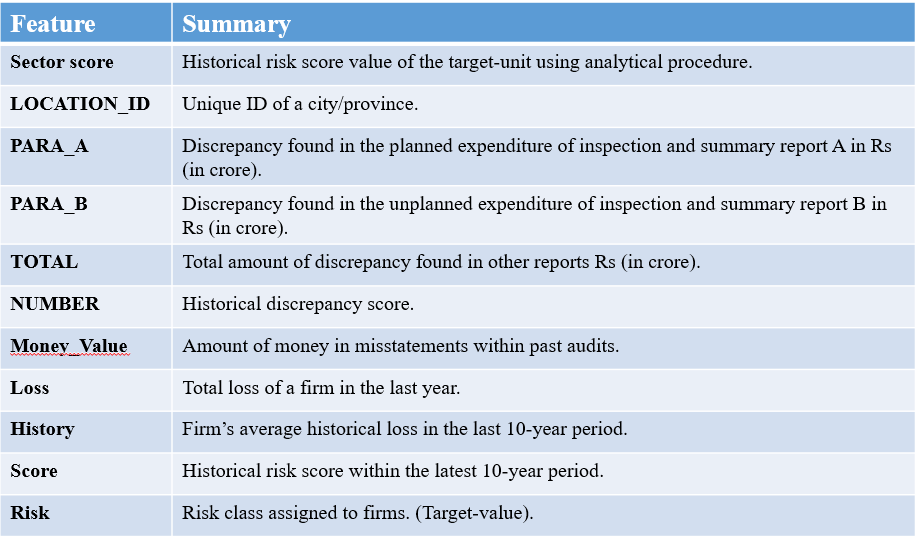
Auditors of India aim to predict firms involved in fraudulent activities with the use of a classification model. The model is developed based on present and historical data. Non-confidential data has been collected between 2015 and 2016 of 777 firms specialized in 14 different type of sectors. The Audit Data Dataset includes 27 numerical attributes, also presented in **Table 1**. Some of the significant values are illustrated in **Table 2**. Analytically, attributes like “**PARA\_A**” and “**PARA\_B**” analyse the discrepancies found in reports showing total firm expenditure. “**Total**” variable shows the total historical discrepancy score of each firm in rupees. “**Money\_value**” indicates the total amount of inaccuracies in historical audits. Both “**Sector score”** and “**Location\_ID”** state historical risk and the unique location id of each firm, respectively. Variables “**Loss**” and “**History**” show total financial loss in the year and the average of the last 10-year period. Lastly, “**Risk**” variable illustrates the firms highly likely to be involved in fraudulent activities denoted with 0’s and 1’s. This variable is used as the target variable. Each numerical attribute demonstrates major risk factors from various areas of the firms.

A supervised learning method – classification can identify the significant attributes to predict what firm has the highest probability of being involved in a fraud. For this case, the target variable is ‘**Risk’** which indicates a prediction.

**Table 1: Audit Data Dataset**



**Table 2: Major attributes of the dataset**



This report aims to investigate present and historical data of firms and predict those that are more likely to be involved in audit fraud. Initially data is explored to understand the type of values in the dataset. Next data-pre-processing has been applied to ensure data is clean. At this stage data is mined; data is modified to ensure it is without missing values and in an understandable format. In addition, data is analysed with the use of graphs and Pearson correlation matrix to identify critical variables for the ‘**Risk’** predictor. Support Vector Machine (SVM) model is developed for prediction purposes. Cross Validation is used to test the accuracy of the model. At the end, the model is evaluated and a discussion is followed for evaluation purposes.

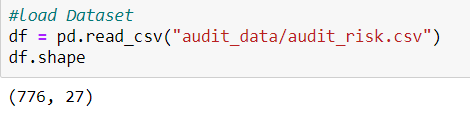
## **1.2 Problem Description**

A specific number of firms might be involved in fraudulent audit tasks based on risk factors. Identification of fraudulent firms requires many hours of intensive manual work on present and historical data. Audit fraudulent tasks may have a critical impact on firms’ stakeholders and particularly on the government. An example of this is presented by the Association of Certified Fraud Examiners that reported the US business are at a loss of $4 billion, annually *(Vona and Leonard, 1955).* Data mining processes have been utilised to focus on methodologies to detect fraudulent transactions. This type of data mining tasks can help the Auditor Office of India, predict suspicious firms that may have a high chance of being involved in suspicious audit fraud activities. Hence, classification machine learning algorithm has been decided to be implemented on audit data.

# **2 Construction and tuning of Machine Learning Model**

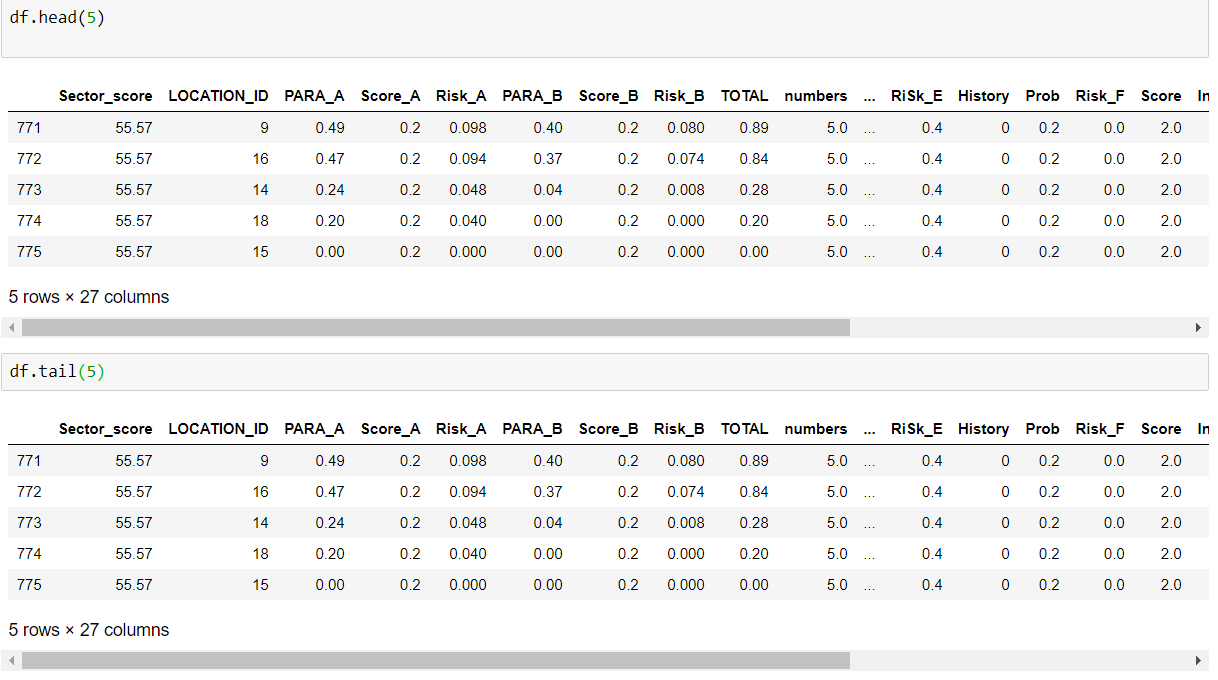
## **2.1 Data Analysis**

Initially, data understanding is an integral part of investigating the Audit Data dataset. Both NumPy and Pandas libraries are imported to manipulate the data and investigate the dataset further. With the use of Pandas library, the data is loaded into a data frame; “**df**” illustrated in **Figure 1**. A data frame is a two-dimensional data structure with both columns and rows. **“df.shape”** function is followed by demonstrating the total number of rows and columns in the dataset.



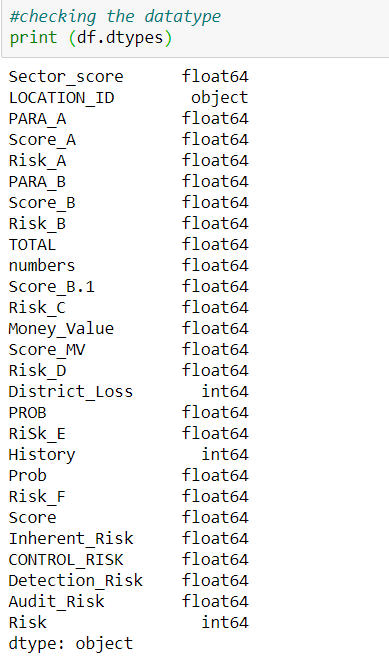
**Figure 1: CSV Dataset upload**

As soon as the data is loaded into a data frame, the dataset can be viewed. This is presented in **Figure 2** with the first and last 5 rows of the dataset.



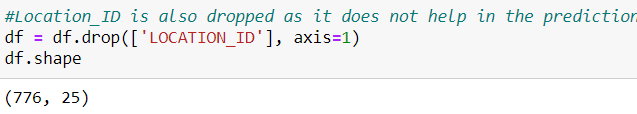
**Figure 2: First and last 5 rows of the dataset**

As it is evident, all the attributes of the dataset are of numerical type apart from the column **“LOCATION\_ID”,** also illustrated in **Figure 3**. The specific column is object type including both numerical and categorical types.

****

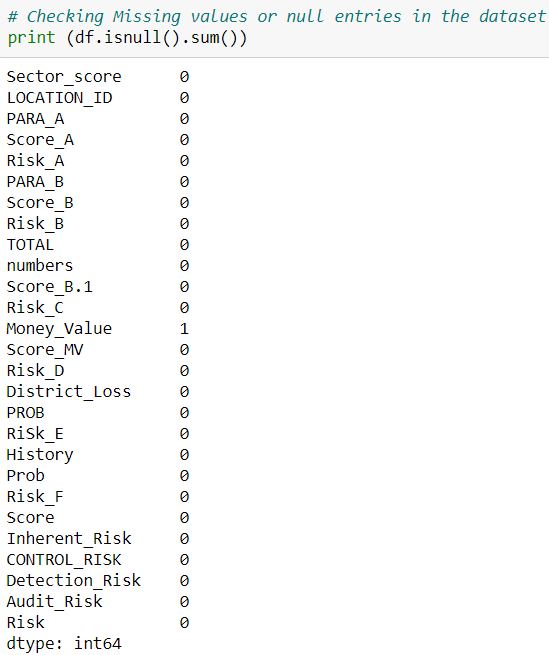
**Figure 3: Data types of the dataset**

At first it was thought that the **“LOCATION\_ID”** should be converted to numerical type as the classifier accepts numerical inputs. However, it was decided that the specific variable can be deleted from the dataset, as presented in **Figure 4**. The specific column demonstrates unique digits of the location of each firm, providing no value in the prediction.

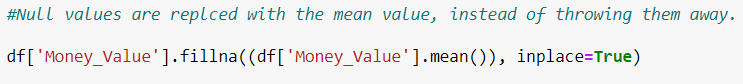
****

**Figure 4: Drop column “LOCATION\_ID”**

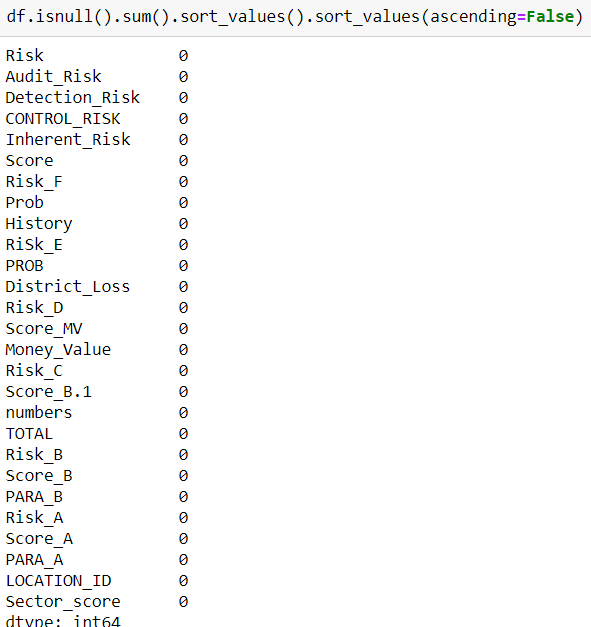
The next necessary task is the deletion of missing values. Pandas **“isnull().sum()”** function is utilised to count all the missing values showing 1 missing value in the “**Money** **Value**” column illustrated in **Figure 5.** This was dealt with by filling the null value with the mean of its attribute value as presented in **Figure 6**, hence leaving the dataset without any missing values as shown in the **Figure 7.**



**Figure 5: Missing value in the ‘Money Value’ column**



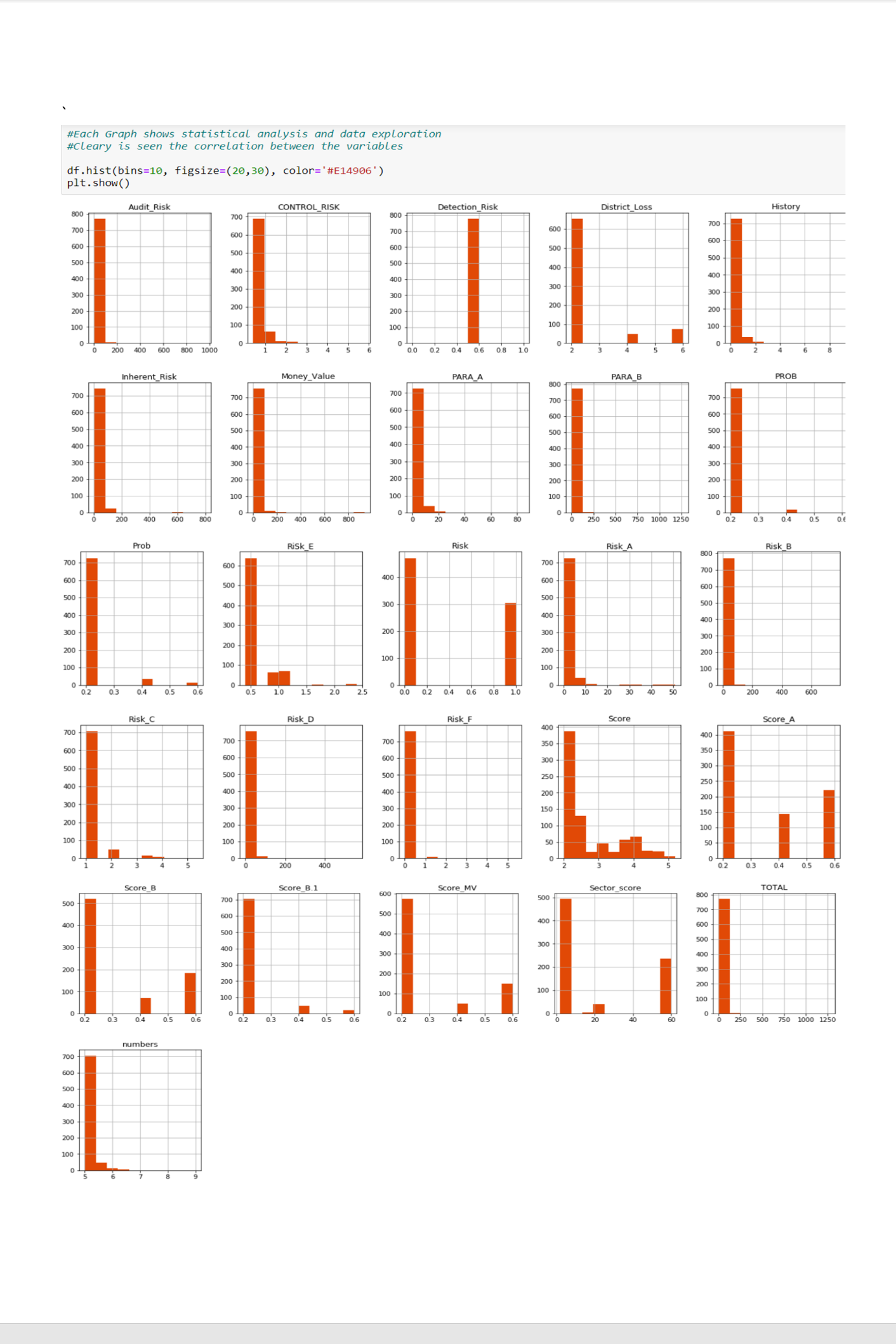
**Figure 6: Filing one null value with the mean**



**Figure 7: No missing values in the dataset**

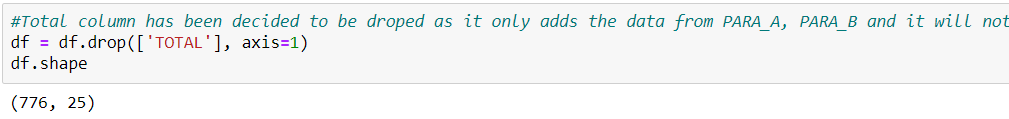
## **2.2 Data Exploration**

Data exploration involves data visualisation with a clear understanding of the data. Statistical graphs were produced using data visualisation libraries in Python. Histograms proved to show helpful and precise representations of the data within the dataset. Specifically, the histograms illustrate the frequency of every unique attribute in each column as presented in **Figure 8.**



**Figure 8: Histograms of all attributes**

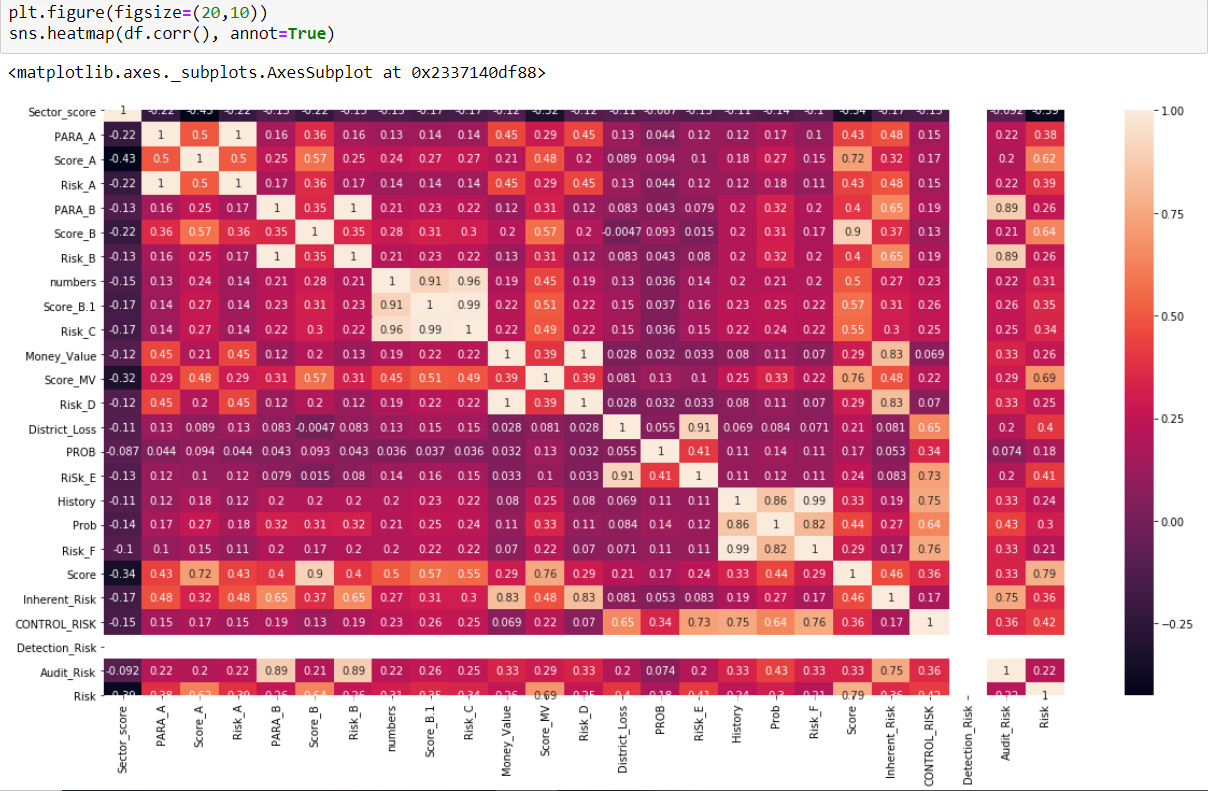
The histograms are clear depicting the range of the numerical values existing in the dataset. **“Detection\_Risk”** column, for instance, is one range column with a number set to 0.5. It seems that the ‘**TOTAL’** column only shows the summary of the columns named **‘PARA\_A’** and **‘PARA\_B’**, therefore it is logical that this column is dropped as it provides no value in the prediction of the fraudulent firms. The **Figure 9** shows how the specific column is dropped.

******

**Figure 9: Drop column ‘TOTAL’.**

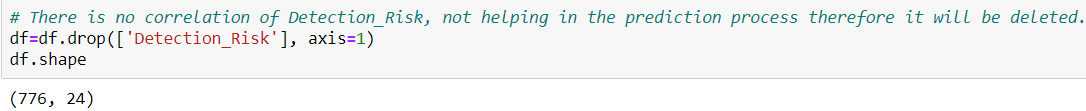
## **2.3 Correlation Matrix**

A correlation matrix is type of table that demonstrates the correlation coefficients between variables. Specifically, between target variable and the rest of the variables. The diagram shows each cell with a figure that represents the correlation between two variables. The diagram is utilised to summarize the data and provide better data insights.



**Figure 10: Creation of correlation matrix**

After creating the heatmap as presented in **Figure 10**, it is evident that most of the variables such as **“Inherent\_Risk”, “Score\_A”, “Score\_B”, “Score\_MV”** and “**Score”** variables seem to have a correlation above 0.5% with the target variable, apart from the “**Detection\_Risk**” variable. The specific variable seems to show no relation; hence it is shown as “**Nan**”. Due to the non-existence of variation within the specific variable, it has been decided that it is deleted as displayed in **Figure 11**. The deletion of the entire column is followed by **“df.shape”** function showing the dimensions of the dataset.



**Figure 11: Drop of “Detection\_Risk” colum****n**

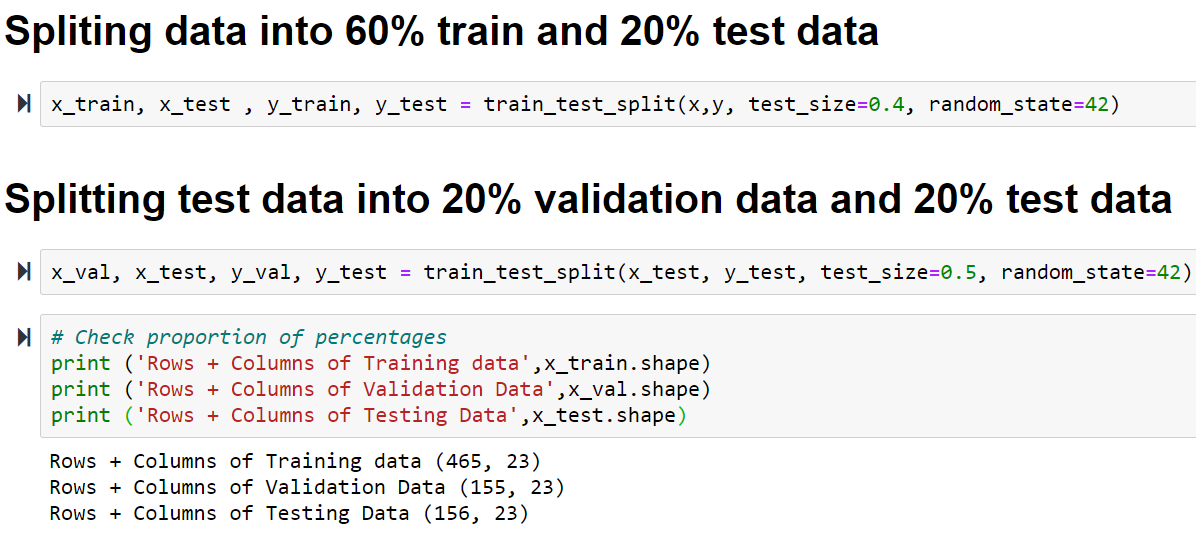
## **2.4 Support Vector Machine (SVM)**

At this stage, SVM classifier is created to predict firms that may have been involved in fraudulent activities. The Chinese government already applies machine learning algorithms, for example, Support Vector Machine (SVM), genetic algorithms, logistic regression, and neural networks with main purpose of identifying frauds in audit tasks of Chinese firms (*Ravisankar et al., 2011*). The SVM algorithm has been decided to be implemented as it can produce complex decision boundaries. In addition, it helps in detecting outliners since all datapoints are fit between the margins. However, in specific occasions of the dataset being too small, the specific algorithm would perform less efficiently since it would create false maximum margin classifier decision boundary.

Before building the algorithm, it is necessary to be aware of problems like overfitting and underfitting. Overfitting happens when model is trained too well; a big amount of unnecessary data is trained in the training data. Underfitting occurs when classifier cannot capture the entire structure of the data meaning the model is too simple. On both occasions, the classification model would entail to poor predictive performance. Therefore, to confront similar type of problems the dataset is divided into training, testing and validation data.

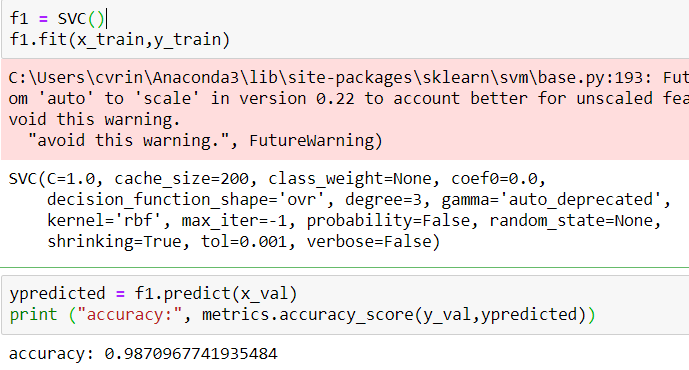
* *Training Data:* The actual data of the problem utilised to build the classification model. The classifier learns from the specific type of data and has outputs.
* *Validation Data:* Validation data is used for evaluation purposes. The specific type of data provides a model evaluation while tuning the classifier.
* *Testing Data:* Testing data is utilised as soon as the model has been trained. It is used for evaluation purposes to test the efficiency of the model prediction.

In detail, the original dataset is divided into 60% training data and 40% testing data. In the next phase, testing data with 40% of data is divided into 20% actual testing data and 20% validation data, as presented in **Figure 12**. “**Random\_state=42**” is used as the generator produces the same output with the first time the data is split. This is useful in the consistency of classifiers results.



**Figure 12: Data splitting into 60% training, 20% testing and 20% validation**

During the development stage of the classification algorithm, training data is utilised to build the SVC model from sklearn. A method, known as “**kernel trick**”, is utilised to identify the decision boundary between the outputs (*Boser, Guyon and Vapnik, 1992).* In this instance, the SVC model uses by default a kernel set to “**rbf**”, gamma set to “**scale**” giving a model accuracy of approximately 0.987%, as demonstrated in **Figure 13**. Gamma illustrates the coefficient of the kernel.

******

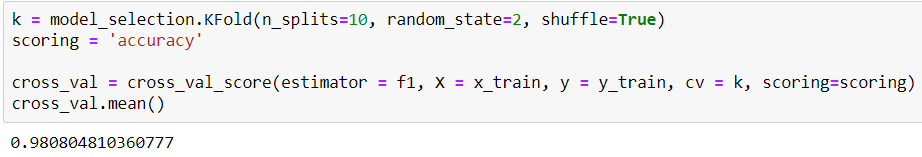
**Figure 13: SVC model building**

# **3 Testing Results**

At this stage of the report, the implemented machine learning algorithm; SVM is evaluated based on its accuracy, precision, recall. A confusion matrix is also analysed.

## **3.1 Cross-Validation**

Cross-validation is a statistical analysis with main goal of testing the accuracy of a predictive model. In the instance of predicting fraudulent firms, a 10-fold cross-validation has been used with the “**model\_selection.KFold**” function, as it is the most common method. The data is split into 10 consecutive folds with data being shuffled since it seems that the model has higher accuracy of 0.08% by randomly shaffling the dataset. Therefore, the variables of the method are set as “**n\_plits=10**”, “**random\_state=2”** and “**shuffle=True**” giving an accuracy of 0.987%, as presented in **Figure 14**. The “**random\_state**” controls the randomness of each fold and is set to 2, as the model has a higher accuracy.

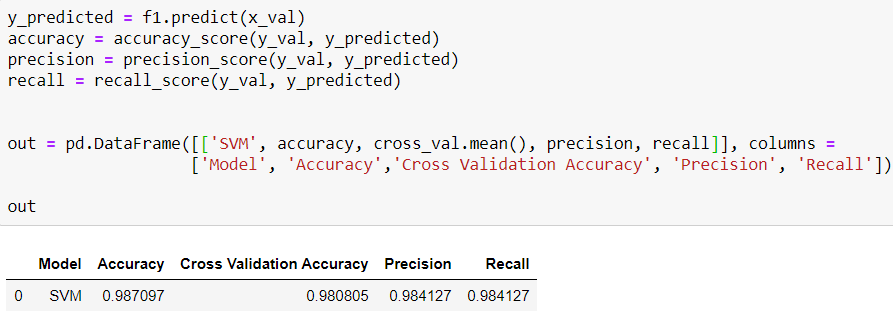
****

**Figure 14: Building Cross-Validation**

For testing purposes, the SVM classification model is evaluated. Model evaluation is achieved based on its precision, recall, cross-validation accuracy.

* *Precision*: Percentage of positive identifications predicted correct. Precision can be defined as the total amount of true positives divided by the amount of true positives plus the amount of false positives.
* *Recall*: Percentage of actual identifications identified correct. Otherwise the ratio of recall is the number of true positives divided by the number of true positives plus the number of false negatives
* *Cross-validation Accuracy*: The specific accuracy identifies an approximate estimate of how accurate a predictive model can be.

As presented in **Figure 15**, cross-validation, precision, and recall are equal to 0.980%, 0.984% and 0.984%, respectively. All percentages are very close to 100% making the specific algorithm highly accurate and precise. SVM’s model accuracy is equal to 0.987%.



**Figure 15: SVM model evaluation**

## **3.2 Confusion Matrix**

A confusion matrix, also known as error matrix, visualises the performance of the classification model. Every column of the matrix represents the instances of the predicted class and each column shows the instances in the actual class. In this confusion matrix, the model accurately predicted that 105/105 firms be 0 and 51/51 be 1. This is presented in the **Figure 16.** In summary, it seems that the confusion matrix is both accurate and correct.

****

**Figure 16: Confusion Matrix of SVM classifier**

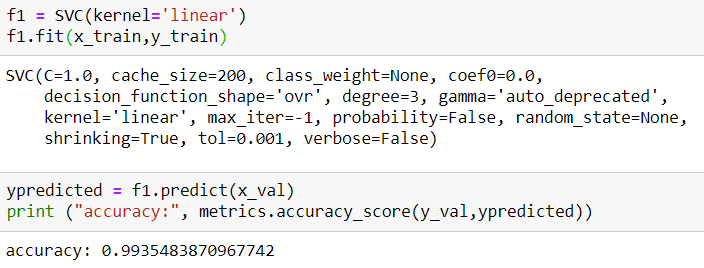
# **4 Discussion**

The goal of this report is the development of a classification algorithm by predicting the firms that may be involved in fraudulent activities using attributes demonstrating major risk factors from various areas of the firms. The attributes were highly related to one another, producing a classification algorithm highly accurate and precise.

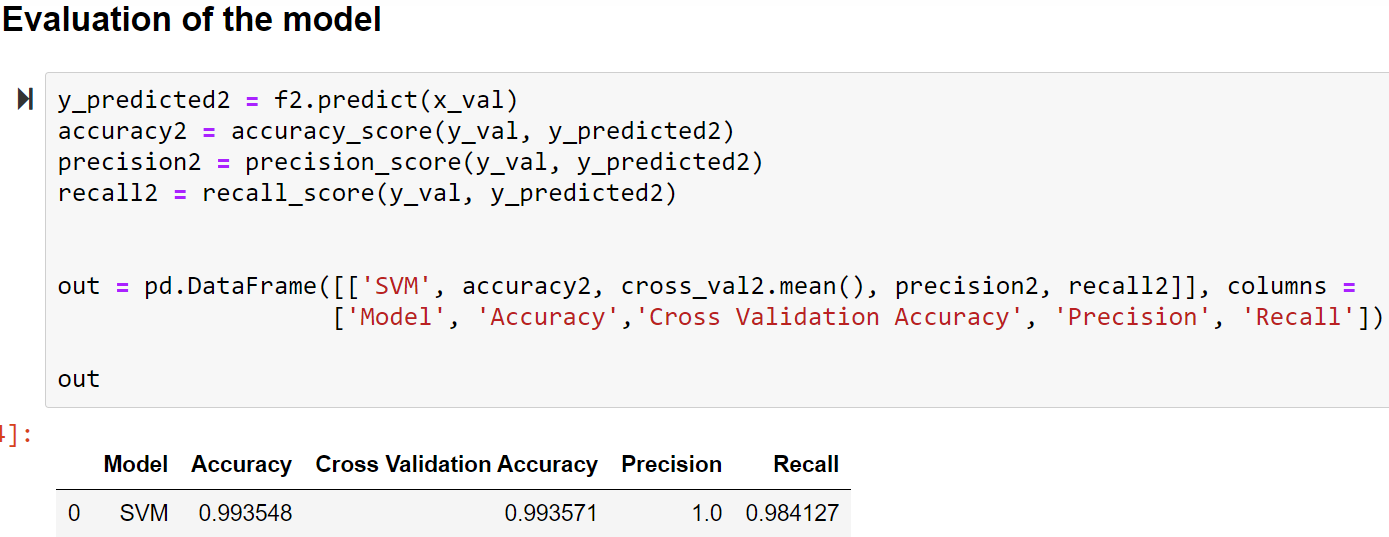
Data pre-processing has had an integral part in the development of the classification algorithm. Initially, the data was analysed by checking null values that were filled with the mean of each attribute value. Next, duplicate values and invalid data was removed before building the SVM classification model. The adoption of data pre-processing methods were important to ensure the data quality.

SVM is a popular machine learning algorithm with a goal of maximizing the quality output of training data. Main idea of the model is the creation of a flat hyperplane splitting the feature space into two parts. During the development stage of the model, the audit data dataset is split into train, test and validate. ”**rbf**” kernel method is utilised define the decision borderline between the outputs. Accuracy using the specific type of kernel is 0.987%.

Although the specific type of kernel has proved to be ideal for classification purposes when applied on generalised labeled data according to *Shankar et al. (2018),* it seems that the ‘linear’ type of kernel has proved to provide better results with an accuracy percentage of 0.994%, as presented in **Figure 17**. A thorough evaluation of the algorithm with the use of “**linear**” kernel is shown in **Figure 18.**

****

**Figure 17: SVC model building with linear kernel**



**Figure 18: SVM model evaluation**

In summary, it seems that data pre-processing was a significant step in the creation of a precise classification algorithm. Additional tuning of the classifier by manipulating the SVC model’s parameters such as the kernel type and the gamma provides better results. Specifically kernel type equal to “**rbf**” provides accuracy of 0.99%, while the “**linear**” kernel provides accuracy of 0.994%. The specific type of classifier has provided substantially accurate results. However, it is important to mention, should the specific model be implemented within in a small dataset it may have created a false maximum margin decision boundary. Hence, it is significant, Auditors of India concentrate on the frequency they carry out audit checks on the firms.

# **References**

*Dataset: https://archive.ics.uci.edu/ml/datasets/Audit+Data*

*Ravisankar, P. et al. (2011) ‘Detection of financial statement fraud and feature selection using data mining techniques’, Decision Support Systems. North-Holland, 50(2), pp. 491–500. doi: 10.1016/j.dss.2010.11.006.*

*Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992) ‘Training algorithm for optimal margin classifiers’, in Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory. New York, New York, USA: Publ by ACM, pp. 144–152. doi: 10.1145/130385.130401.*

*Shankar, R. S. et al. (2018) ‘PREDICTION of EMPLOYEE ATTRITION USING DATAMINING’, in 2018 IEEE International Conference on System, Computation, Automation and Networking, ICSCA 2018. Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ICSCAN.2018.8541242.*

*Vona, Leonard. The Fraud Audit: Responding to the Risk of Fraud in Core Business Systems. 1. Aufl. Vol. 16. Hoboken, N.J: Wiley, 2011. Print.*