**Design and Application of a Machine**

**Learning System for a Practical Problem**

**Report on the Investigation**

**By**

**Smruti Das**

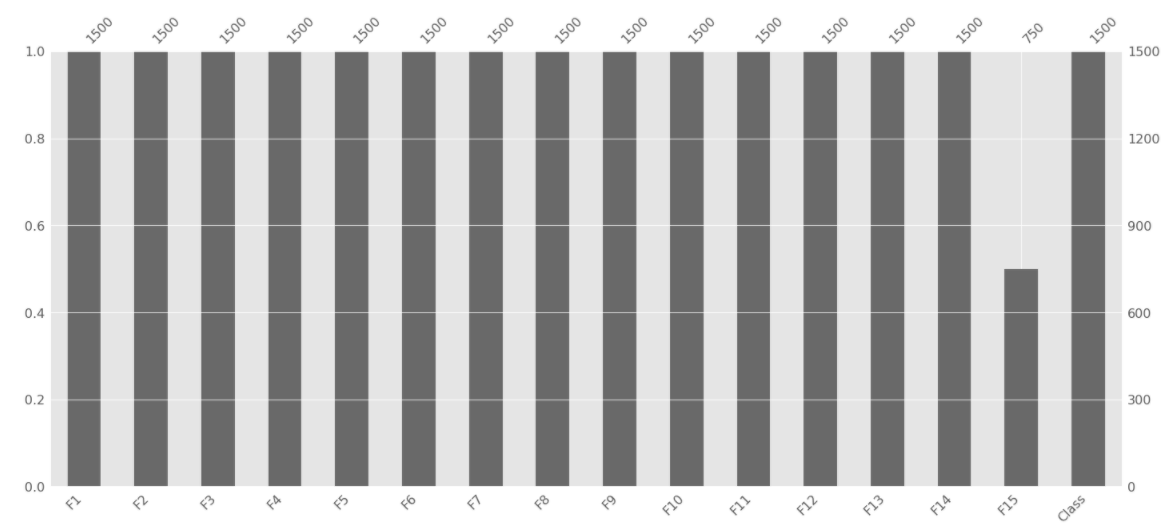
**Overview:**

This is a comparative study of different machine learning techniques and algorithms to predict if a customer will claim a travel insurance in future or not. To perform this task, 1500 historic data of customers with a combination of customers who claimed and those who did not was given.

Based on this dataset, below study has been performed.

1. **Pre-Processing:**
   1. **Comparison of Data Pre-processing Methods**

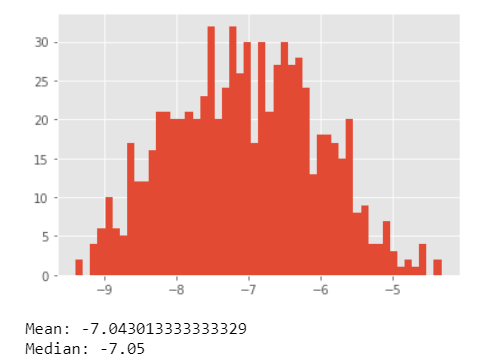
Before Appling any machine learning algorithms, it is important to explore and treat the data if there are any anomalies like null values or duplicate records. As there were no duplicate records, null values were checked



*Fig 1.1.a: Distribution of F15 feature*

There are 750 nulls in ‘F15’ Feature. Below methods are implemented,

1. Removing F15 feature
2. Replace all null with 0
3. Replace all null with Mean or Median (Mean or Median Imputer)



*Fig 1.1.b: Distribution of F15 feature*

Above methods were implemented and data was fed to Pruned decision tree. Below is the model accuracy. From figure 1.1.b if it clear that the mean and median are nearly same for F15. Hence, we can implement only mean imputer (There are no outliers)

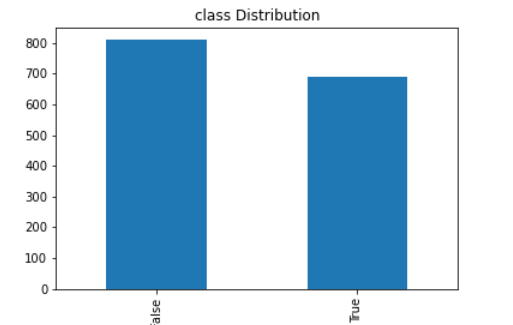
|  |  |
| --- | --- |
| **Methods** | **Accuracy** |
| Removing F15 feature | 0.7899 |
| Replace all null with 0 | 0.79083 |
| Replace all null with Mean (Mean Imputer) | 0.8108 |

*Table 1.1.a Accuracy of Data Pre-processing Methods*

Hence, using Mean Imputer gives the best result.

* 1. **Checking Class Label counts**

The Target Class should be balanced in order to get best result from the algorithms. Otherwise, we would need to implement techniques like over sampling or under sampling to make the data balanced.

****

*Fig 1.2.a: Distribution of F15 feature*

The Class distribution seems to be almost evenly distributed.

**1.3 Scale the Dataset**

Some Algorithms which uses distance as a measure, performs better when we use scaled data.

We have used MinMax scaler for SKlearn library. It transforms the data between 0 and 1 with mean as 0 and standard deviation as 1. K-NN, xgboost works best with scaled data.

**2. Comparative Study**

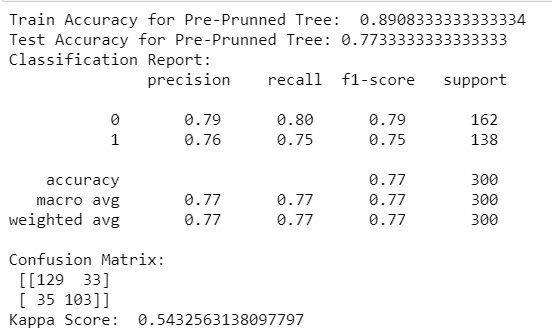
The Performance of Model can be determined by many Measures. Few are listed below

* Accuracy
* Confusion score
* confusion matrix
* kappa scores

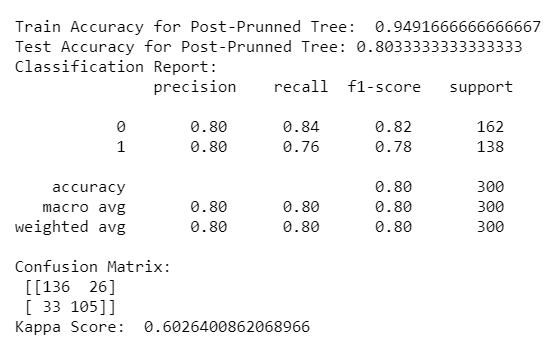
**2.1 Decision Tree**

Decision Tree is widely used algorithm which is easy to implement and visualize. It works well with categorical and/or numerical data. However, it doesn’t perform well with higher dimension data. We can easily overfit a decision tree of not pruned. Decision Tree is implemented with the help of Scikit learn library. *DecisonTreeClassifier* has many parameters which we can tuned to get best accuracy.

**2.1.1 Pre-Pruned Decision Tree Accuracy**

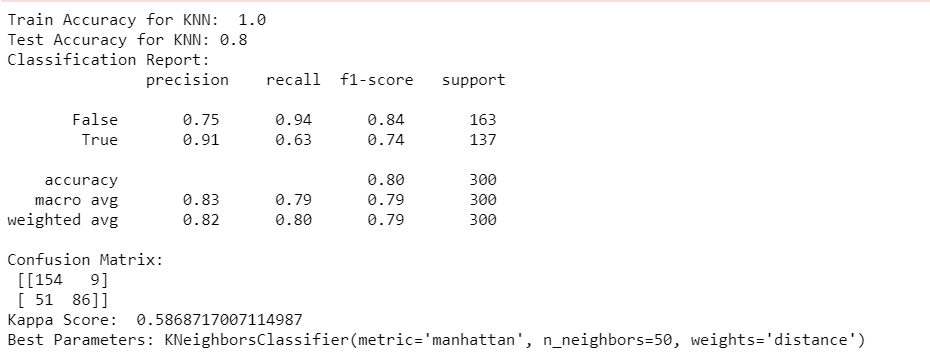


**2.1.2 Post Pruned Decision Tree Accuracy**



**2.2 K-NN**

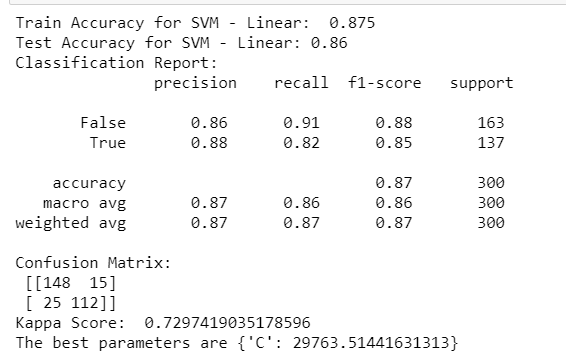
K-NN is an instance-based algorithm. Using Grid search cross validation Based hyperparameter k is tuned to avoid over or under fitting. It uses different distance measures like Manhattan, Minkowski or Euclidian distance to find the nearest neighbours. *KNeighborsClassifier* in Sciket Learn library is used.



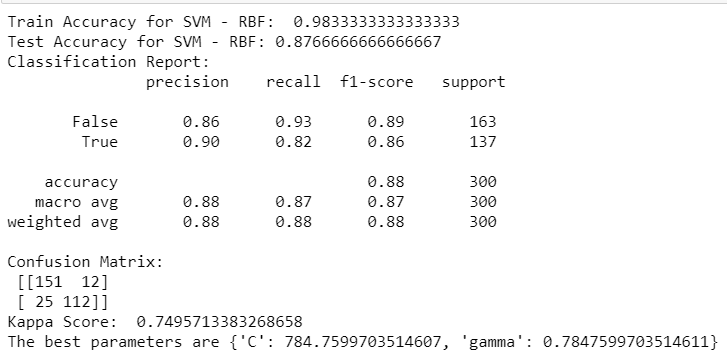
**2.3 Support Vector Machine**

SVM works well it linear as well as non-linear dataset. We can use different types of kernels for non-linear dataset. RBF(Radial basic function) is generally used non-linear kernel.

**2.3.1 Linear Kernel SVM Performance measure**

****

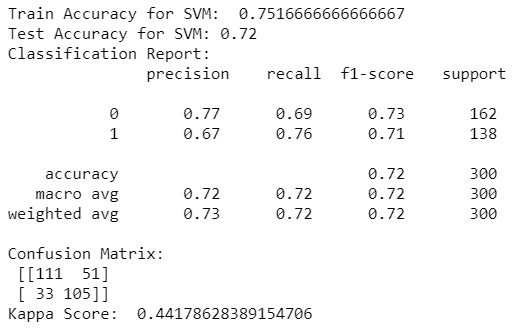
**2.3.2 RBF kernel Performance measure**

****

RBF kernel performs a bit better than linear Kernel SVM.

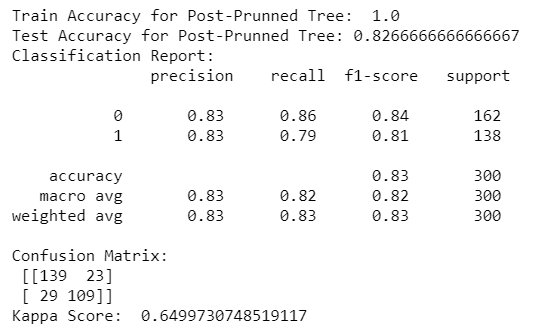
**2.4. Logistic Regression**

Logistic regression is relatively fast compared to other supervised classification techniques such as kernel SVM or ensemble methods. Logistic regression tends to underperform when the decision boundary is nonlinear.



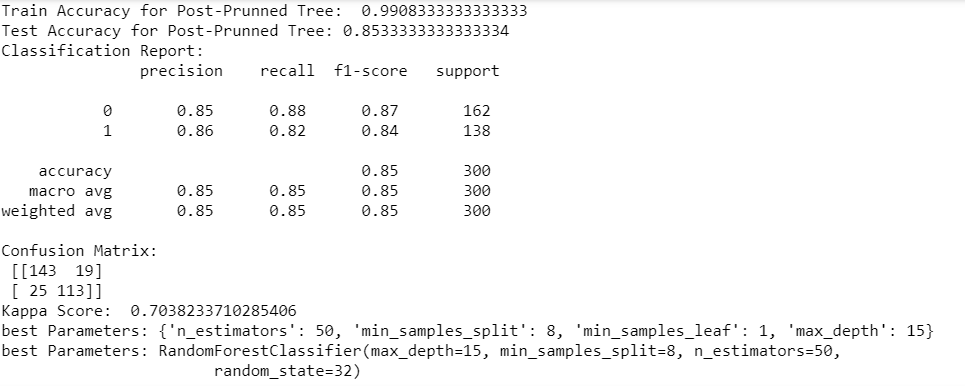
**2.5 Random Forest**

Random Forest uses bagging ensemble technique. It uses combination of decision tree, bootstrap row sampling, column sampling to produce different models and make prediction based on majority rule for classification problem. Decision trees are allowed to overfit in random forest. As it uses more than one model the overall variance of model is low.

****

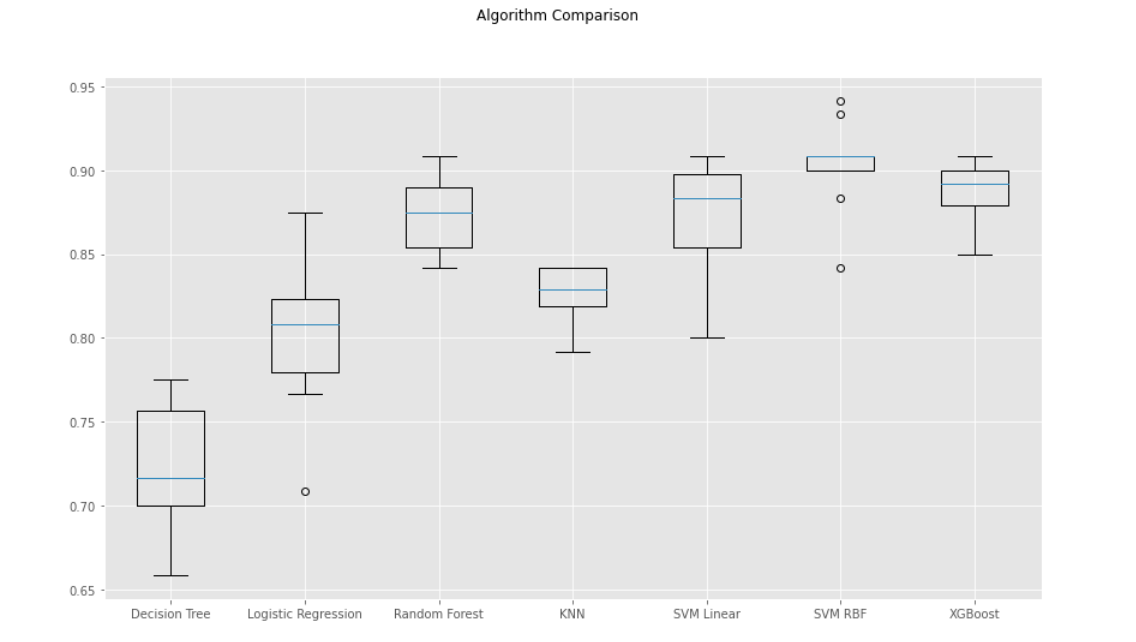
**2.6 Xgboost**

Xgboost uses boosting ensemble technique. It uses combination of decision tree, bootstrap row sampling, column sampling and aggregation of models. We start with under sampled models and gradually achieve best model by reducing the residue. As there is aggregation of models the overall bias of the model is less.

****

**Conclusion:**

Xgboost and SVM with RBF kernel gives best results of all other models. We are going to use SVM with RBF kernel as our final model. We will use the trained model to make prediction for the given test dataset.



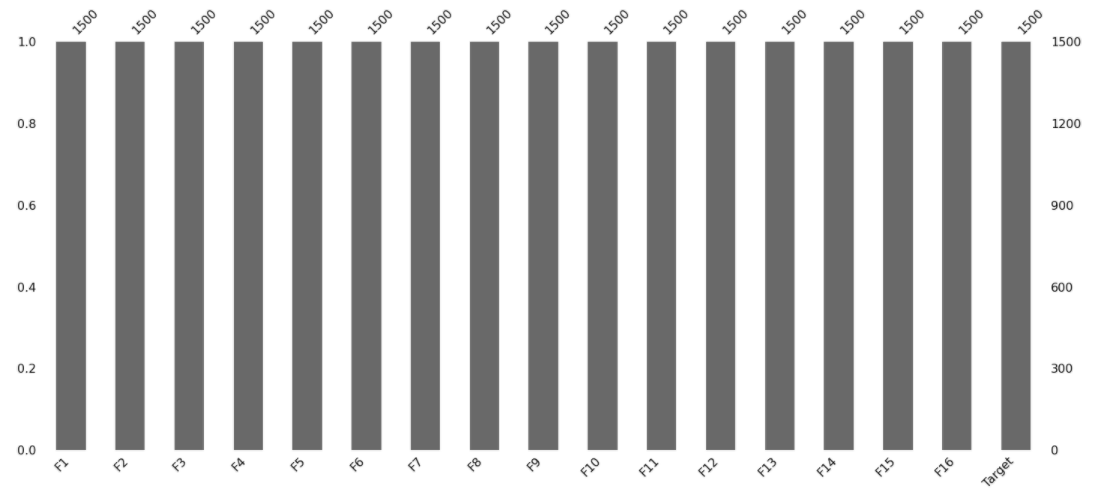
**3. Additional Comparative study**

Afterachieving the model to classify if the insured would claim or not, we need to predict the amount of claim. To perform this task, 1500 records with the value of claim made by the customers was shared which was used to train the model and make the prediction on the test dataset.

**Pre-Processing:**

Before Appling any machine learning algorithms, it is important to explore and treat the data if there are any anomalies like null values or duplicate records.

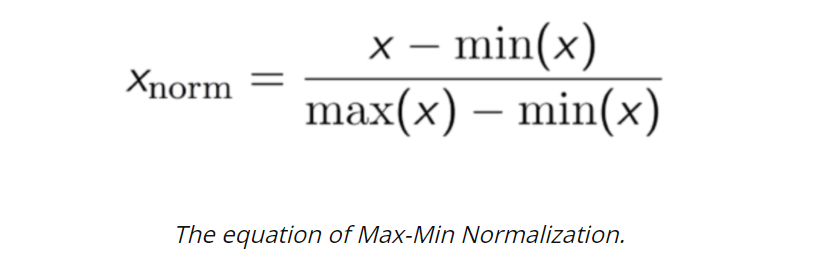
There are no null values as shown in below figure.



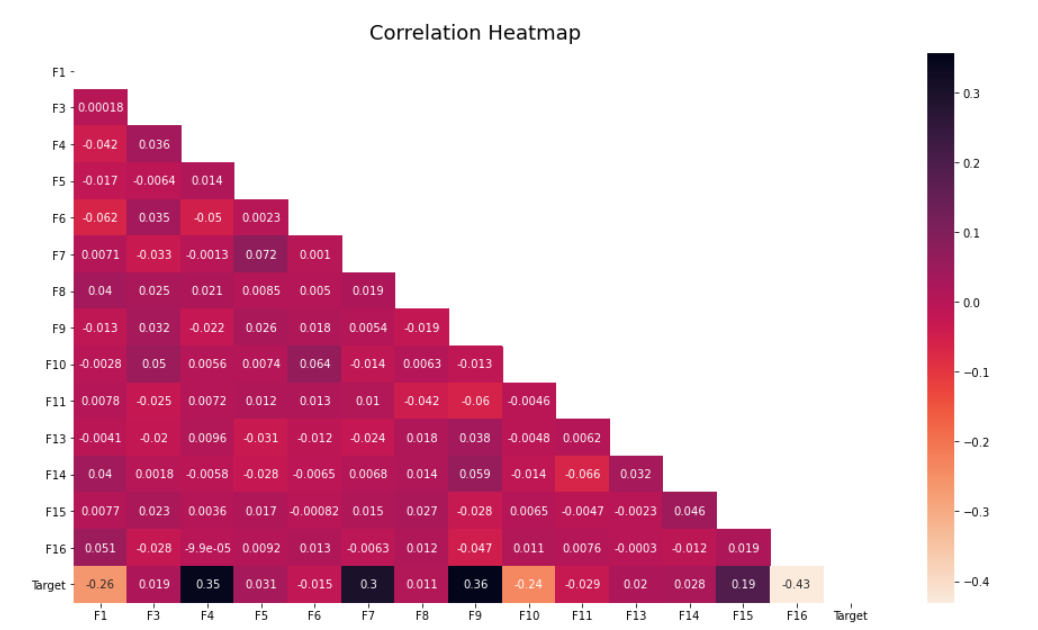
*Fig 3.1.a Count of records*

**Scale the Dataset**

For a regression problem, it is advisable to scale the input as well as target data. This helps to bet optimised model quicker. We have used MinMax scaler for SKlearn library. It transforms the data between 0 and 1 with mean as 0 and standard deviation as 1.



**Co-relation Matrix**

****

*Fig 3.1.b Co-relation Matrix*

This shows that few features like F1, F4, F7, F9, F10, F15 and F16 are more co-related to the target than other features.

**3.1 Linear Regression**

In Linear regression the models depend linearly on their unknown parameters. Hence, it’s easier to fit than other models which are non-linearly related to their parameters. If the goal is prediction or forecasting, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, the fitted model can be used to make a prediction.

Below are the Accuracy results for Linear regression



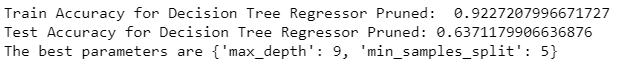
**3.2 Decision Tree Regressor**

A **decision tree** is a supervised machine learning model used to predict a target by learning decision rules from features. To use a decision tree for regression, we need an impurity metric that is suitable for continuous variables, so we define the impurity measure using the weighted mean squared error (MSE) of the children nodes. One major disadvantage of Decision Trees is that they are prone to overfitting and it looses information when it categorizes variables in different categories.

**Fully grown decision tree Accuracy:**



**Pruned Decision Tree Accuracy:**

****

Decision tree doesn’t work well in this case.

**3.3 Polynomial Regression**

Polynomial regression is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) in which the relationship between the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) x and the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) y is modelled as an nth degree [polynomial](https://en.wikipedia.org/wiki/Polynomial) in x. Polynomial regression is considered to be a special case of [multiple linear regression](https://en.wikipedia.org/wiki/Multiple_linear_regression). It provides the best approximation of the relationship between the dependent and independent variable. However, these are too sensitive to the outliers.

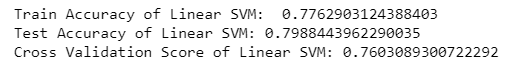
Below are the Accuracy results for Polynomial regression with degree 2.



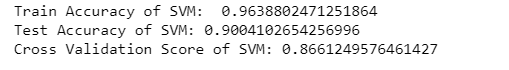
**3.4 Support Vector Machine**

SVM is widely used for regression problem along with classification. We have used non-linear Kernels and RBF gives the best results. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. It uses support vectors for its calculation.

**Linear SVM Accuracy:**

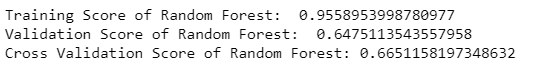


**Non-Linear Kernel ‘RBF’ SVM Accuracy:**



**3.5 Random Forest**

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.



**Conclusion:**

SVM with RBF kernel gives the best prediction amount all the models. The test Accuracy of is 0.9 and cross validation score is 0.87. This shows that the features have non-linear relationship with target.

**References:**

* Scikit-learn.org. 2021. *Sklearn.Metrics.Accuracy\_Score — Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html>
* Scikit-learn.org. 2021. *Sklearn.Metrics.confusion\_matrix — Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>
* Scikit-learn.org. 2021. *Sklearn.Metrics.DecisionTreeClassifier— Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
* Scikit-learn.org. 2021. *Sklearn.Metrics.SVM — Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/svm.html>
* Mitchell, T. M., (1997),“Machine Learning”,McGraw-Hill. Chapter 3. Tan, Steinbach & Kumar (2006) “Introduction to Data Mining”.
* Exsilio Blog. 2021. *Accuracy, Precision, Recall & F1 Score: Interpretation Of Performance Measures - Exsilio Blog*. [online] Available at: <https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>
* Medium. 2021. *Do You Know How To Choose The Right Machine Learning Algorithm Among 7 Different Types?*. [online] Available at: <https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60>