Data Mining and Machine Learning I

Raj R. Kupekar
National College of Ireland
(Of Affiliation)
Dublin, Ireland
Email Id: kupekarraj@gmail.com

***Abstract*—The research includes the performance study of different Machine Learning algorithms on a variety of large datasets. For analysis, these datasets are gathered from different domains including gaming, marketing and educational & learning. The primary objective is to implement and overcome the limitations associated with Machine Learning methods; to evaluate an efficient predictive model. A Decision Tree Classifier is applied on Chess dataset from gaming domain, which consists of records of different chess board positions. Thus, an optimal model is evaluated by pruning the tree parameters. Furthermore, a Random Forest is implemented, after which its accuracy is compared with Decision Tree model and an efficient model is carried forward to predict the class of the dependent variable.**

Similarly, effective predictive model is been designed for Letter Recognition dataset holding an educational and learning background. Initially, K Nearest Neighbor model is been trained on this dataset and a scalable model is obtained by setting the appropriate k value. This model is then compared with a Support Vector Machine Classification model which is tuned to its best fit by modifying the cost parameter for the model. Significantly, SVM model outperforms the KNN model acquiring a model accuracy of 97%.

Finally, Binomial Logistic Regression is implemented on Tesco dataset from marketing domain. The dataset consists of entries of content card and the expenditure history for each of its customers. Logistics Regression algorithm is implemented to classify the class of their customers to come up with the best suited content card. An accuracy of 65 % is obtained by tuning the regression model.

Keywords—Decision Tree, Random Forest, K Nearest Neighbor, Support Vector Machine, Logistics Regression.

I. INTRODUCTION

In games like Chess, it is always difficult to anticipate the opponent move and accordingly decide our own strategical attacking move. There are almost infinite chess board positions which can be successfully performed to checkmate the opponent. Accordingly, it is practically tricky for a human being to train himself and summarize the wining move. Thus, a Machine Learning model can be built and trained based on available data and past experiences; so that an individual can compete with it. To do so, a variety of 37 large number of chess board instances are gathered along with the status of end game. Each attribute represents a chess end game position between the white pieces left only with king and rook and black pieces with king and pawn. This immense data is been used to train the model using the Machine Learning technique. As the status of end game is either win or lose which is binomial in nature; a Decision Tree Classifier and Random Forest algorithms are implemented to obtain the desired results.

Another popular application of Machine Learning technique is text extraction and recognition. Also known as Text mining or Character recognition, is widely used in real time application in a variety of domain such as reading license plates, captchas detection, extracting the appropriate character from a handwritten character. Similarly, K Nearest Neighbor and Support Vector Machine algorithms are used to predict the letter which are the image representation of a combination of different alphabets in which each alphabet is possibly written in different manner. In our case, the image pixels are scaled down to numeric values ranging from 0-15. Class distribution of the dataset is uniform thus, eliminating the issue of over sampling or under sampling. Both KNN and SVM model are implemented and an efficient model with highest accuracy judged on different parameter is choose as the final model.

In recent years, modern businesses are adopting technology to solve crucial problem for their business. E-commerce websites like Amazon, eBay, Flipkart, Pin Interest are using Machine Learning and Deep learning techniques in applications like product recommendation, searches, anomaly detection and so on. With primary focus of processing and extracting insights from the immense number of facts generating at each sec, Machine Learning models can be implemented for pattern discovery. Tesco, which is one of the leading Supermarket in Ireland uses the same Machine Learning principle. In this dataset, customer expenditure behavior and history of the marketing content card is recorded. This information is then used to categorize the customer so that appropriate content card can be displayed to them and business opportunities can be diversified. Logistics Regression methodology is performed to achieve a profitable business.

II. RELATED WORK

In [1], the focus is on various Decision Tree algorithms with their compared characteristics and limitations. In this paper, ID3 algorithm is studied which constructs a Decision Tree using the Information Gain value of attributes. Similarly, a more powerful algorithm C 4.5 is discussed which works on the same principle of Information Gain as ID3 but, it also accepts continuous data and set threshold to convert it in categorical values. Furthermore, a non-parametric CART technique, which automatically does variable selection is reviewed in this research paper.

In this paper [2], classification techniques like Decision Tree, KNN, Neural Networks and SVM are been studied and compared using different parameters such as accuracy in general, speed of learning of model, tolerance to missing values. After evaluation Decision Tree and KNN algorithms are carried forward for further application. During performance analysis of these two models, Decision Tree yields a higher accuracy than KNN. Similarly, in comparison using cost complexity and prediction Decision Tree outperforms KNN.

Economic Events are predicted in [3] using KNN Machine Learning methodology. This paper first discusses the different types of mining techniques such as regression, classification and clustering and accordingly uses the KNN classification technique for its application. It also examines the KNN regression and classification method and its application in different sectors.

In [4] , accuracy of KNN model is improved by local mean based and distance weight technique. A KNN model is implemented and the class of the data point is determined using two different class deciding techniques. Initially, Euclidean distance is calculated which is then sorted in ascending order. Thereafter, local mean vector and weights of distances are calculated for each class. Finally, the proposed method model attains the highest accuracy as compared to the default KNN model.

In this paper [5], two classification techniques i.e. Random Forest and Decision Tree are implemented on Breast Cancer dataset. The models are evaluated using accuracy, precision and F-score statistics. The classification results outline that Random Forest performs better for large datasets while decision tree performs better with small dataset. The accuracy of Random Forest increases from 70% to 96% when the number of observations is increased from 300 to 700.

This paper [6], deals with background history of Random Forest in the past 15 years. The developments and improvements in the algorithms are examined since Brie man 2001 approach. This study also discusses the current research work on Random Forest algorithm. Initially, a Meta Random Forest was been designed in 2006 which used the Meta learning techniques with concepts of base classifier. Subsequently, several advancements in Random Forest are studied in this paper.

In this research paper [7], Support Vector Machine algorithm is used for detection and classification of Acoustic Breathing Cycles. Digital signal processing technique is used for collection of acoustic signals of respiration. These signals are processed and differentiated between voiced and unvoiced period. Finally, SVM is implemented which yields an accuracy of 95% when tested on the subjects.

This paper [8], introduces a new Support Vector Machine algorithm applied on variety of datasets like Diabetes, Heart, Satellite and Shuttle datasets having multi-class levels. SVM is evaluated by tuning kernel parameters from linear to sigmoid and rbf. In this manner an efficient model is built which make use of grid search with 5-fold cross validation to select the best parameters.

This paper [9], surveys about Naïve Bayes algorithm and interprets the Augmented Naïve Bayes text classification, Spam filtration and Sentiment Analysis applications. The issue of Zero conditional probability is resolved using a kernel density estimation.

In this paper [10], Naïve Bayes classifier is used to evaluate a model for Diabetes Mellitus. The data is gathered from WHO site. Naïve Bayes classifier is implemented on the data of patients with two and three levels. The model accuracy acquired for these levels are 78.88% and 68.50%. The model accuracy is then improved by inserting an additional attribute particle swarm optimization which acquires an accuracy of 82.58% and 71% respectively.

This research paper [11], makes use of Machine Learning Logistics Regression in predicting the results of RNA molecular hybridization. The problem of deciding the exponential complexity of RNA in a limited time, Logistics Regression is implemented. Prior to this, several Machine Learning algorithms like Decision Tree, Random Forest and Boosting tree are implemented. But the optimal model is evaluated using Logistics Regression by comparing the ROC curve for each of the methodology.

III. DATA MINING METHODOLOGY

The methodology implemented is the CRISP- DM which stands for Cross Industry Process for Data Mining. It is a robust and well demonstrated methodology which provides a structural procedure in extracting insights from the given data. Conceptually, it is a broad iterative process through which a business case can be solved by using underline application of data mining methods. The CRISP-DM technique involves a six-step process from selection of dataset to knowledge discovery. The following are the steps involved in CRISP-DM process.

- Understanding of Business
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

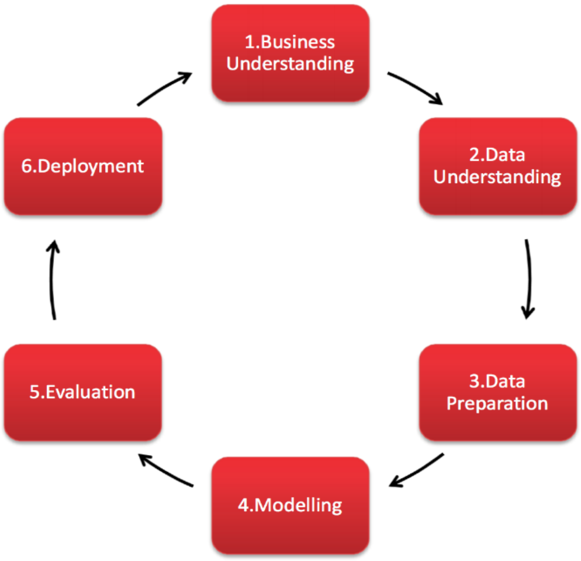


Fig.1 Cross Industry Process for Data Mining (CRISP-DM)

1. **Understanding of Business**

   The first stage of CRISP-DM process is to determine the business objective. This means to outline the primary objectives from business prospective and defining the associated clause of business case. This is followed by describing the intended action plan in achieving the desired goal. For chess dataset, the elementary objective is to predict whether the white is going to win the end game or not, given that the pawn is only 1 square away from queening. Similarly, for Letter Recognition the objective is to recognize the alphabetical letters which are encoded in from its image representation to its equivalent integer values using various attributes to represent its image representation precisely. Finally, displaying appropriate marketing content cards to its customers based on their expenditure and historical events of content cards.

   In order to accomplish our objective, an effective and executable plan is essential. This includes selection of data for analytics and using appropriate data mining tools and methodology. In this study, for chess data the relevant attributes are selected for analytics and Decision Tree along with Random Forest is implemented to achieve the desired result. K Nearest Neighbor and Support Vector Machine is used for evaluation for Letter Recognition. Accordingly, normalizing and converting the data type of the variables is carried out before implementation of KNN model. Finally, for Tesco marketing dataset, Logistics Regression model is executed.

2. **Data Understanding**

   This stage requires initial understanding of data, sources of data, loading of data into data mining tools and exploring the data for analysis. Data description report is designed which acquires the information of data, its format, quantity as well as quality of data. This is followed by exploring the data which includes deciding the target variable for analysis and distribution in key attributes and relationship between the predictors. The next important sub-stage is substantiating the quality of data. Examining whether the data is complete with absence of null entries in data and checking if any error in data entries.

   2.1 Chess Dataset

   In this study, Chess end-game dataset is a collection from UCI Machine Learning Repository which is having an immense data resources for Machine Learning application. The collected dataset is readily available in Comma Separated Values (.csv) format. There are in total 3196 instances with 36 independent variables and one dependent variable. The data type of all the variables is categorical for Chess dataset with response variable with two levels “win” or “no win”. Furthermore, there is an absence of missing values and no impact of data entries error.

Status (Response variable)	‘Win’, ‘No win’
Feature_1	‘f’, ‘l’
Feature_2	‘f’, ‘l’
Feature_3	‘f’, ‘l’
Feature_4	‘f’, ‘l’
Feature_5	‘f’, ‘l’
Feature_6	‘f’, ‘l’
Feature_7	‘f’, ‘l’
Feature_8	‘f’, ‘l’
Feature_9	‘g’, ‘t’
Feature_10	‘f’, ‘l’
Feature_11	‘f’, ‘l’
Feature_12	‘f’, ‘l’
Feature_13	‘f’, ‘l’
Feature_14	‘f’, ‘l’
Feature_15	‘b’, ‘n’, ‘w’
Feature_16	‘f’, ‘l’
Feature_17	‘f’, ‘l’
Feature_18	‘f’, ‘l’
Feature_19	‘f’, ‘l’
Feature_20	‘f’, ‘l’
Feature_21	‘f’, ‘l’
Feature_22	‘f’, ‘l’
Feature_23	‘f’, ‘l’
Feature_24	‘f’, ‘l’
Feature_25	‘f’, ‘l’
Feature_26	‘f’, ‘l’
Feature_27	‘f’, ‘l’
Feature_28	‘f’, ‘l’
Feature_29	‘f’, ‘l’
Feature_30	‘f’, ‘l’
Feature_31	‘f’, ‘l’
Feature_32	‘f’, ‘l’
Feature_33	‘f’, ‘l’
Feature_34	‘f’, ‘l’
Feature_35	‘f’, ‘l’
Feature_36	‘f’, ‘l’

Table 1. Chess Dataset

2.2 Letter Recognition

Like Chess end game dataset, this dataset is also collected from UCI Repository which is in Tab-separated values (.tsv) format. It is then converted into .csv format which is the required form for data mining application. There are in total 20000 observations with about 17 variables. In this dataset, ‘letter’ variable is the target variable for prediction, which is of factor data type with 26 levels. These levels represent the 26 English alphabets with uniform class distribution thus, eliminating the effect of biasness in data.

Letter (Response Variable)	26 values in factor form (A-Z)
x-box	Horizontal position of box integer values
y-box	Horizontal position of box integer values
width	Width of box integer values
high	Height of box integer values
onpix	Total number on pixels in integer values
x-bar	mean x of on pixels in box in integer values
y-bar	mean y of on pixels in box in integer values
x2bar	Mean x variance
y2bar	Mean y variance
xybar	Mean x y correlation in integer
x2ybr	Mean of x*x*y in integer values
xy2br	Mean of x*y*y
x-ege	Mean edge count left to right in integers
xegvy	Correlation of x-ege with y
y-ege	Mean edge count bottom to top in integers
yegvx	Correlation of y-ege with x

Table 2. Letter Recognition Dataset

2.3 Tesco Marketing

Finally, the third dataset is the Tesco Marketing Content Dataset. This dataset is collected from Kaggle which is a storehouse, that host thousands of datasets such that data mining and machine learning techniques can be implemented. Like Letter Recognition dataset, the file format of Tesco dataset is in .tsv format which is converted to .csv format in Excel. The dataset consists of 10000 records with 27 columns and ‘content_1’ as the dependent variable. There are 9 different marketing content history associated with additional expenditure behavior of individual customers. The dependent and the content card variables are having 3 levels with following entries, where ‘1’ means the customer clicked on the content card, ‘0’ means the customer haven’t clicked on the content card and ‘NA’ means the card was never shown to the customer. The dataset is having zero null values but, the ‘NA’ levels in content card columns is coded to different values to avoid fetching it as null value.

Content_1 (Response Variable)	‘0’ – Customer clicked on the content card ‘1’- Customer haven’t clicked on the content card ‘NA’-The card was never shown to the customer
Customer.id	The unique Customer Id
Content_2 to Content_9	‘0’ – Customer clicked on the content card ‘1’- Customer haven’t clicked on the content card ‘NA’-The card was never shown to the customer
express.no.transactions	Transaction in Express store
express.total.spend	Expenditure in Express
metro.no.transactions	Transaction in Metro store
metro.total.spend	Expenditure in Metro
superstore.no. transactions	Transaction in Super store
superstore.total.spend	Expenditure in Superstore
direct.total.spend	Expenditure in Online store
gender	‘Male’, ‘Female’
affluency	Affluence of customer
county	The county where the customer lives

Table 3. Tesco Marketing Dataset

3. Data Preparation

This is the third stage of CRISP-DM methodology. This stage requires selection of relevant data for analysis followed by cleaning and transformation of data if required. In Chess dataset, all the variables seem to be relevant for modelling thus, none of the variable is dropped at the initial stage. Similarly, the data type of the dependent variable i.e. status is converted to categorical type from string data type. It is then encoded to Boolean value (i.e. win= ‘0’, no win= ‘1’) to interpret it easily while modelling. In our second dataset, column names are given to the data frame which represents the various parameter in predicting the outcome variable. The data type of dependent variable is converted to factor type from character type as required for analysis. Furthermore, the integer and numeric data type variables are normalized in order to scale them down in the range of 0 to 1. This is done to achieve the data form requirement before applying the Machine Learning models. Finally, in Tesco Marketing, the variables with qualitative data is transformed to categorical data type. At initial stage, the county variable is dropped from the data frame as because it is not much significant for the model. Neither of the dataset is having missing values in it thus, missing values imputation is excluded. Additionally, all the three datasets are divided into two data frames i.e. training and testing data such that a model is implemented on training data and evaluation of model is carried out on testing data.

4. Data Modelling

This stage requires, selecting the appropriate modelling technique and satisfying the associated model assumptions. Implementing the selected Machine Learning technique on training data and testing the model accuracy on testing data. This is followed by building and running a model with alteration in the default parameters to evaluate an efficient model. In our case, for Chess end game dataset the best suited algorithms are Decision Tree and Random Forest Classifier. As both the algorithm have a similar working principle of building trees for prediction thus, it is accessible to compare and estimate the best fit model.

Similarly, K Nearest Neighbor and Support Vector Machine Classification are implemented on Letter Recognition dataset. KNN classifies the response variable by calculating the Euclidean distance between the response variable and the specified number of K value. Thus, aggregating the most frequent label as the final class for that response variable. In case of SVM, a hyperplane is built in between the different class of the outcome variables having similar features. As, both the algorithms have a common working principle of building decision boundaries between different data points, so with this similarity KNN and SVM is implemented on Letter Recognition dataset.

Subsequently, Logistics Regression is carried on Tesco Marketing dataset. It is a classification technique that makes use of Sigmoid function and takes probability into account and sets a decision boundary to decide the class of the data point. It works on the principle of Maximum Likelihood which estimates the values for coefficient that minimizes the error in predicting the probabilities of the model.

5. Evaluation

This is the important stage of CRISP-DM methodology. In this stage the actual accomplishment of the business objective is evaluated. Also, insights can be drawn for the deficient model and accordingly business plan can be made. Likewise, this stage helps to unveil other related information for future directions.

In our first dataset, as discussed Decision Tree and Radom Forest models are implement. Decision Tree in implemented using tree() library in R framework. The model is first trained on training dataset and its efficiency is tested on testing dataset. Caret library is used for partitioning the data into training and testing. Initially, Decision Tree is modelled on training data of Chess dataset without any alteration in the default parameters of tree() function. This model is then used to predict the data points of testing data. Here predict() function is used for prediction in R. Similarly, model accuracy is calculated using the confusion matrix which is a function from e1071 library. The attained model accuracy with default parameters of decision tree is 97% with Specificity at approximately 99%, which is an indication of best fit model. This accuracy is then validated using a cross validation technique with misclassification as basis for pruning the tree model. The results of cross validation give clear evident that a tree with best parameter set to 14 generates a productive predictive model with accuracy approximately to 98%.

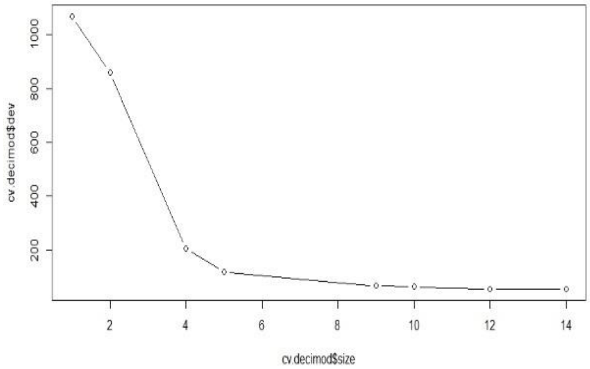
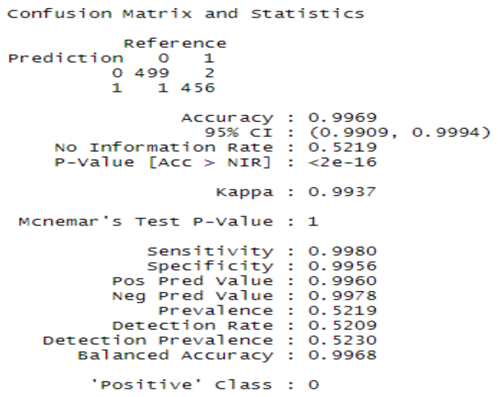


Fig. 1 Plot of cv size and dev parameter

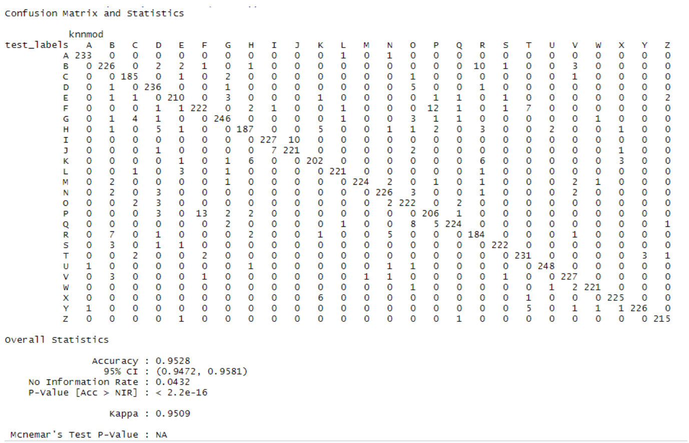
On the other hand, Random Forest is implemented using randomforest() library in R. Like Decision Tree, Random Forest is implemented without any tuning parameters. Thus, Random Forest achieve a model accuracy of 99% which is approximately 1% higher than Decision Tree. As, the maximum possible is accuracy is obtained from Random Forest with out any tuning parameters, thus Random Forest is chosen as the final predictive model for Chess dataset.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 499   2
         1   1 456

               Accuracy : 0.9969
                 95% CI : (0.9909, 0.9994)
    No Information Rate : 0.5219
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9937

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9980
            Specificity : 0.9956
         Pos Pred Value : 0.9960
         Neg Pred Value : 0.9978
             Prevalence : 0.5219
         Detection Rate : 0.5209
   Detection Prevalence : 0.5230
      Balanced Accuracy : 0.9968

       'Positive' Class : 0
```

Fig.2 Confusion Matrix of Random Forest

Another classification technique known as KNN is implemented on Letter Recognition dataset. This dataset comprises of 26 multiclass levels as its response variable having categorical datatype and rest of its variable Integer data type. Thus, in order to perform KNN algorithm the integer data types are normalized using normalization function in R. Thereafter, KNN model is implemented using knn() function from class library. Initially, the k-value is decided as the square root of the number of observations in training dataset. Thus, in our case the k value is decided as 118 which indicates that the KNN model will calculate the Euclidean distance of the new data point with 118 nearest neighbor and likewise aggregate the final class as the value with the most frequent class. The accuracy of the KNN model with k value set to 118 is 80.57%.This model accuracy is increased by selecting the optimum value for k which maximizes the accuracy of the model. This is achieved by coding a for loop which returns the accuracy of model with corresponding k-value. Accordingly, model with k-value 1, attains a maximum accuracy of 95% which is distinctly higher than a KNN model with k-value 118.

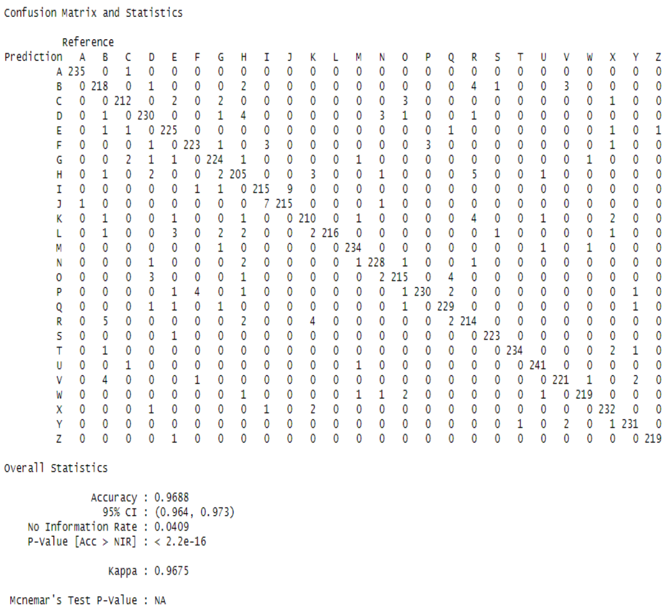


```
Confusion Matrix and Statistics

           knnmod
test_labels   A   B   C   D   E   F   G   H   I   J   K   L   M   N   O   P   Q   R   S   T   U   V   W   X   Y   Z
          A 233   0   0   0   0   0   0   0   0   0   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0
          B   0 226   0   2   2   1   0   1   0   0   0   0   0   0   0   0   0  10   1   0   0   3   0   0   0   0
          C   0   0 185   0   1   0   2   0   0   0   0   0   0   0   1   0   0   0   0   0   0   1   0   0   0   0
          D   0   1   0 236   0   0   1   0   0   0   0   0   0   0   5   0   0   1   0   0   0   0   0   0   0   0
          E   0   1   1   0 210   0   3   0   0   0   1   0   0   0   0   1   1   0   1   0   0   0   0   0   0   2
          F   0   0   0   1   1 222   0   2   1   0   0   1   0   0   0  12   1   0   1   7   0   0   0   0   0   0
          G   0   1   4   1   0   0 246   0   0   0   0   1   0   0   3   1   1   0   0   0   0   0   1   0   0   0
          H   0   1   0   5   1   0   0 187   0   0   5   0   0   1   1   2   0   3   0   0   2   0   0   1   0   0
          I   0   0   0   0   0   0   0   0 227  10   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
          J   0   0   0   1   0   0   0   0   7 221   0   0   0   0   2   0   0   0   0   0   0   0   0   1   0   0
          K   0   0   0   0   1   0   1   6   0   0 202   0   0   0   0   0   0   6   0   0   0   0   0   3   0   0
          L   0   0   1   0   3   0   1   0   0   0   0 221   0   0   0   0   0   1   0   0   0   0   0   0   0   0
          M   0   2   0   0   0   0   1   0   0   0   0   0 224   2   0   1   0   1   0   0   0   2   1   0   0   0
          N   0   2   0   3   0   0   0   0   0   0   0   0   0 226   3   0   0   1   0   0   0   2   0   0   0   0
          O   0   0   2   3   0   0   0   0   0   0   0   0   0   2 222   0   2   0   0   0   0   0   0   0   0   0
          P   0   0   0   3   0  13   2   2   0   0   0   0   0   0   0 206   1   0   0   0   0   0   0   0   0   0
          Q   0   0   0   0   0   0   2   0   0   0   0   1   0   0   8   5 224   0   0   0   0   0   0   0   0   1
          R   0   7   0   1   0   0   0   2   0   0   1   0   0   5   0   0   0 184   0   0   0   1   0   0   0   0
          S   0   3   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0 222   0   0   0   0   0   0   0
          T   0   0   2   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0 231   0   0   0   0   3   1
          U   1   0   0   0   0   0   0   1   0   0   0   0   0   1   1   0   0   0   0   0 248   0   0   0   0   0
          V   0   3   0   0   0   1   0   0   0   0   0   0   1   1   0   0   0   0   1   0   0 227   0   0   0   0
          W   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   1   2 221   0   0   0
          X   0   0   0   0   0   0   0   0   0   0   6   0   0   0   0   0   0   0   0   1   0   0   0 225   0   0
          Y   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   5   0   1   1   1 226   0
          Z   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0 215

Overall Statistics

               Accuracy : 0.9528
                 95% CI : (0.9472, 0.9581)
    No Information Rate : 0.0432
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9509

 Mcnemar's Test P-Value : NA
```

Fig.3 Confusion Matrix for KNN model with k=1

A cross table is from 'gmodels' library is executed to interpret the percent of correct ani incorrect classification the predicted data class.

KNN model is then challenged using SVM algorithm to check whether the SVM performs better yielding higher accuracy than it. SVM is implemented using svm() function from 'e1071' library. The model attains an accuracy of 94%. Thus, after tuning the SVM model using cross validation and train control parameter with cost parameter set to 10, the model accuracy is increased by 2% .i.e. 96.88% . This obtained accuracy from tuned SVM model is greater than the KNN final model. Hence, the tuned SVM model is taken as the final predictive model for Letter Recognition dataset.



```
Confusion Matrix and Statistics

          Reference
Prediction   A   B   C   D   E   F   G   H   I   J   K   L   M   N   O   P   Q   R   S   T   U   V   W   X   Y   Z
         A 235   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
         B   0 218   0   1   0   0   0   2   0   0   0   0   0   0   0   0   0   4   1   0   0   3   0   0   0   0
         C   0   0 212   0   2   0   2   0   0   0   0   0   0   0   3   0   0   0   0   0   0   0   0   1   0   0
         D   0   1   0 230   0   0   1   4   0   0   0   0   0   3   1   0   0   1   0   0   0   0   0   0   0   0
         E   0   1   1   0 225   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   1   0   1
         F   0   0   0   1   0 223   1   0   3   0   0   0   0   0   0   3   0   0   0   0   0   0   0   1   0   0
         G   0   0   2   1   1   0 224   1   0   0   0   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0
         H   0   1   0   2   0   0   2 205   0   0   3   0   0   1   0   0   0   5   0   0   1   0   0   0   0   0
         I   0   0   0   0   0   1   1   0 215   9   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
         J   1   0   0   0   0   0   0   0   7 215   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
         K   0   1   0   0   1   0   0   1   0   0 210   0   1   0   0   0   0   4   0   0   1   0   0   2   0   0
         L   0   1   0   0   3   0   2   2   0   0   2 216   0   0   0   0   0   0   1   0   0   0   0   1   0   0
         M   0   0   0   0   0   0   1   0   0   0   0   0 234   0   0   0   0   0   0   0   1   0   1   0   0   0
         N   0   0   0   1   0   0   0   2   0   0   0   0   1 228   1   0   0   1   0   0   0   0   0   0   0   0
         O   0   0   0   3   0   0   0   1   0   0   0   0   0   2 215   0   4   0   0   0   0   0   0   0   0   0
         P   0   0   0   0   1   4   0   1   0   0   0   0   0   0   1 230   2   0   0   0   0   0   0   0   1   0
         Q   0   0   0   1   1   0   1   0   0   0   0   0   0   0   1   0 229   0   0   0   0   0   0   0   1   0
         R   0   5   0   0   0   0   0   2   0   0   4   0   0   0   0   0   2 214   0   0   0   0   0   0   0   0
         S   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0 223   0   0   0   0   0   0   0
         T   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 234   0   0   0   2   1   0
         U   0   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0 241   0   0   0   0   0
         V   0   4   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 221   1   0   2   0
         W   0   0   0   0   0   0   0   1   0   0   0   0   1   1   2   0   0   0   0   0   1   0 219   0   0   0
         X   0   0   0   1   0   0   0   0   1   0   2   0   0   0   0   0   0   0   0   0   0   0   0 232   0   0
         Y   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   2   0   1 231   0
         Z   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 219

Overall Statistics

               Accuracy : 0.9688
                 95% CI : (0.964, 0.973)
    No Information Rate : 0.0409
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9675

 Mcnemar's Test P-Value : NA
```

Fig.4 Confusion Matrix of SVM tuned model

Finally, Logistics Regression is implemented for Tesco dataset. In this dataset, the values for the content card which are not shown to the customers are represented using 'NA'. Thus, while dealing with the missing values, these entries of content card are detected as missing value in data. Also, the class distribution is highly bias towards the 'NA' level in the content card attributes. Practically, the columns with more than 30% of missing values are dropped from the dataset. But in our case, dropping such columns resulted in only 100 observations in the final dataset. Hence, to avoid the above-mentioned issues the 'NA' levels are coded to '1' which means the customer never clicked on the content card. Accordingly, Logistics Regression is implemented on the transformed data. Initially, the class of testing data are predicted by setting the threshold probability at 0.70%. The accuracy obtained with this threshold value is less than 50% which is indication of underfitting model. Thus, to overcome this the regression model is tuned by setting the threshold value at 0.58%. Hence, a model accuracy of 65% is obtained from the tuned Logistics model. Similarly, ROC curve is used to determine the suitable threshold value for prediction. In our case, the ROC curve is somewhat approaching the diagonal of the image grid as shown in the figure below.

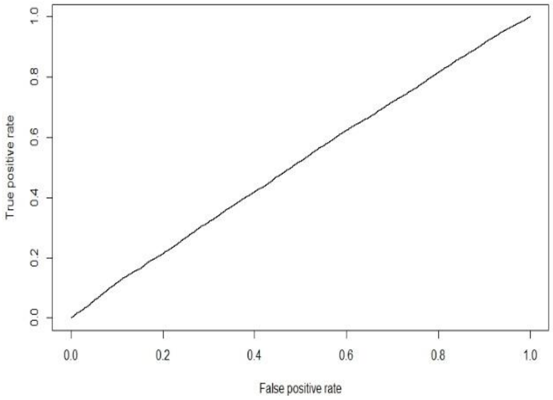


Fig.5 ROC Curve

IV. Counclusion and Future work

Thus, for Chess dataset the final predictive model is evaluated using Random Forest whereas for Letter Recognition, Support Vector Machine is elected as the final data mining model while Logistics Regression been the final model for Tesco Marketing dataset. In Chess dataset, Ensemble modelling technique can be performed to generate a more powerful predictive model. Also, the accuracy of Random Forest is in close approximation to 100%. Thus, this can be a warning indication of an overfitting model.

Moreover, the SVM model can be tuned using a list of different cost parameter passed as a vector in the SVM model to evaluate the optimum cost parameter value. But sometimes the kernel goes in an infinite mode while performing the above code.

The model accuracy for Tesco dataset can be improved by collecting a greater number of records for the class with lowest number as compared to the ‘NA’ class in content attribute.

V. References

[1] S. K. Himani Sharma, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research*, p. 5, 2015.

[2] S. G. P. D. K. I. L. Bhaskar N. Patel, "Efficient Classification of Data Using Decision," *Bonfring International Journal of Data Mining*, vol. 2, no. 1, p. 7, 2012.

[3] M. B. Sadegh Bafandeh Imandoust, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *S B Imandoust et al. Int. Journal of Engineering Research and Applications* , vol. 3, no. 5, p. 6, 2013.

[4] E. B. N. O. S. S. K U Syaliman, "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," *Journal of Physics*, vol. 10, p. 7, 2017.

[5] R. K. N. A. Jehad Ali, "Random Forests and Decision Trees," *IJCSI International Journal of Computer Science*, vol. 9, no. 5, p. 8, 2012.

[6] E. A. Eesha Goel, "Random Forest: A Review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 1, p. 7, 2017.

[7] M. F. Omar Yahya, "Automatic Detection and Classification of Acoustic Breathing Cycles," *Conference of the American Society for Engineering Education*, vol. 1, p. 5, 2014.

[8] S. L. B. DURGESH K, "DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE," *Journal of Theoretical and Applied Information Technology* , vol. 1, no. 1, p. 7, 2009.

[9] M. S. D. Pouria Kaviani, "Short Survey on Naive Bayes Algorithm," *International Journal of Advance Engineering and Research Development*, vol. 4, no. 11, p. 8, 2017.

[10] D. R. Desi Susilawati, "Optimization The Naive Bayes Classifier Method to diagnose diabetes Mellitus," *IAIC Transactions on Sustainable Digital Innovation* , vol. 1, no. 1, p. 9, 2019.

[11] X. L. M. X. a. H. W. Weijun Zhu, "Predicting the Results of RNA Molecular Specific Hybridization Using Machine Learning," *IEEE/CAA JOURNAL OF AUTOMATICA SINICA*, vol. 6, no. 6, p. 9, 2019.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.