Machine Learning in Banking Churn Prediction

I. Introduction

In today's modern world, there is a high competitiveness in businesses for every industries as the market is massively dynamic and exponentially growth. Particularly, for the banking industry mostly has been disrupted by the financial technology (FinTech). Due to the fact that there is a rise of using innovative information and automation technology for interacting with customers to improve the level of customer satisfaction. Despite one of the main reasons in which could cause an unpleasant situation in an organisation is 'customer attrition' (customer churn). It is described as the phenomenon where by a customer leaves a service provider or the loss of customer by a business [1].

Customers have the right to change their decisions to cease a relationship with a business for several reasons depending on various factors. Furthermore, without the in-dept understanding and utilising by the information technology (IT) and big data in business, it could cause a difficulty to maintain the competitive edge in this era. Due to there is a very large impact of digital technology in terms of highly competitiveness and contestability of the banking markets [2].

In terms of measurement, one of the metrics to keep track of customer churn is improving retention rate so-called 'Churn Prediction'. As the result can provide a reasonable outcome to analyse the tendency of customer churn. The importance of predicting result can indicate the key performance indicator (KPI) to consider a firm's performance whether a good strategic decision have been applied or not. For instance, if a performance shows that there is a high rate in a customer attrition. Then, companies have to find a new pathway to adapt their strategies.

As a core business, customer retention is one of the main considerations. Oliver [3] stated that customer retention refers to "A deeply held commitment to re-buy or re-patronize a preferred product or service consistently in the future, despite situational influences and marketing efforts having the potential to cause switching behavior". Moreover, one of the best way to understand the customer behaviour is analysing through the analysis of historical and new customer data.

Companies that make extensive use of customer analytics see a 126 percent profit improvement over competitors [4]. Therefore, by applying machine learning is one of the powerful tools which assists a firm to understand the insight of customer behaviour and improve a performance in a business strategic decision. Machine learning (ML) is defined as a category of artificial intelligence (AI) by utilising the capability of computers to think and learn on their own [5]. A computer will initially digest the input data afterwards generate the performance for the output result. Hence, having adequate data and appropriate techniques can provide many advantages to maintain a good direction of the business [6]. Nonetheless, the aim of this paper is to represent the concept of big data and machine learning. Also, using some Machine learning models and techniques to investigate and predict client churning who tends to discontinue using a credit card in the banking sector [7].

II. Overview and Concepts

The advent of big data is a valuable component in every organisations for the predictive analysis. It is defined as a term for the massive scale of datasets with a complex and varied structure [7]. These data are mostly collected from various sources in form of surveys, health records, videos, images and online banking transaction. Moreover, big data and machine learning have a synergic relationship in which can play a significant role in finding the insight of information [8]. As an illustration, big data is more like an engine before putting into a machine in order to leverage capabilities.

Theoretically, the framework of machine learning on big data is displayed in Fig.1. Initially, by starting from the input of big data and process through machine learning segment by pre-processing, learning and evaluation. Pre-processing concept is to converting tidy formats of raw data into appropriate formats. The raw data can come as a form of inconsistent, unstructured and noisy aspect. These incomplete cases can be managed by data cleaning, extraction and transformation. Next, learning phase is learning from the data which are already cleaned and processed in pre-processing step through chosen learning algorithms to predict the output. Afterwards, the result of evaluation can follow to determine the performance of learned models such as feature selection, performance accuracy and error-estimation [9]. Furthermore, users can interact with machine learning by providing domain knowledge such as preparing training examples to improve in decision making.

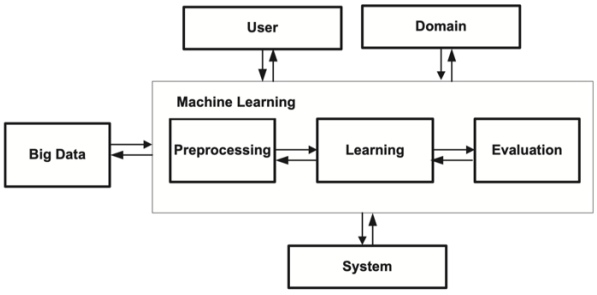

Fig. 1. The framework of machine learning on big data. [9]

In terms of learning system, machine learning can be classified into three main types: supervised learning, unsupervised learning and reinforcement learning (See Fig.2). On the one hand, in supervised learning focuses on the processing in between each pair of input and output when historical input and output data are generated. Then, the main result will learn a function that maps the inputs into bringing out the performance of the output. Furthermore, the supervised learning are classified into classification and regression algorithms. For example, Bayesian Networks, Support Vector Machines, Decision Trees and Neural networks are common algorithms to utilise. Additionally, supervised learning is the most appropriated and widely used for business purposes such as sales forecasting, fraud detection including a customer churn prediction.

On the other hand, unsupervised learning is not considering the targeted output. All of the attributes are used as inputs, the techniques are suitable for clustering and association. Hofmann [10] claimed that unsupervised learning algorithms are suitable for creating the labels in the data that are subsequently used to implement supervised learning tasks. Nonetheless, clustering is a technique for grouping unlabeled data based on similarities and differences mostly used for image processing and market segmentation. For learning algorithms are using K-means and Spectral Clustering.

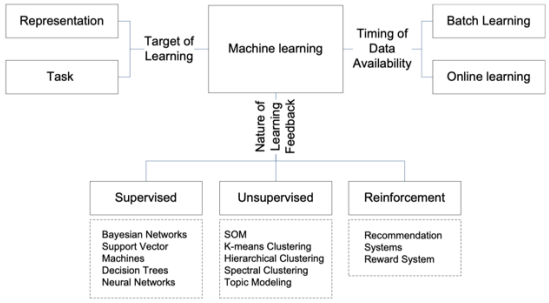

Fig. 2. A multi-dimensional taxonomy of machine learning. [9]

III. METHODOLOGY

This experiment aims to analyse the customer churn prediction who tend to drop off from the credit card service. Therefore, a bank manager can come up with new marketing strategies to convince and engage in a customer satisfaction by using machine learning techniques and algorithms to find out the best-fit model and predict the result. However, from this experiment would be engaged with supervised learning algorithms as making an assumption on predicted parameter as a binary classification output for the outcome to the prediction of customer churn. A diagrammatic processing system is shown in Fig. 3.

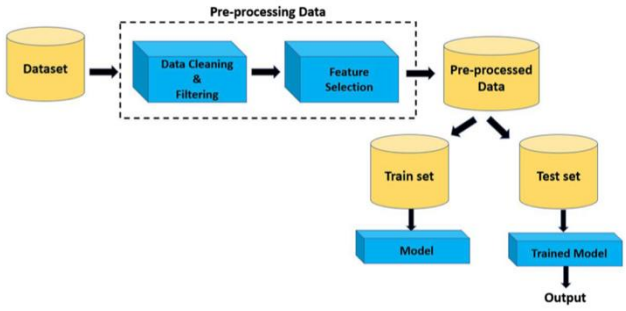

Fig. 3. The System Architecture.

A. Understanding Dataset

The consumer credit card dataset includes 10,127 rows which represents the amount of bank client and 23 columns are characterised the demographic and general details of each individual bank clients. Furthermore, there are some demographic variables such as age, gender, education level, marital status and account holder's annual income. Additionally, the rest variables are mostly about numeric client's details.

As in this dataset, the data type are in different forms such as in numeric type (int, float) and object. By separating the data into two different groups can be a good practice for visualizing the data. Firstly, the dataset would be divided into two groups; categorical and numeric group. There are six features in the categorical table and fourteen features are in numerical table as shown in Table I. and Table II.

TABLE I. CATEGORICAL FEATURES

Feature Name	***Feature Description***
1.Attrition Flag	Exiting customer (0) or Attrited customer (1) – output predicted variable
2.Education Level	High school, College graduate, etc.
3. Marital Status	Married, Single, Divorced, Unknown
4. Card Category	Type of Card (Blue, Silver, Gold, Platinum)
5. Gender	Male or Female
6. Income Category	Less than $40K, $40K - 60K, $60K - $80K, $80K-$120K, Greater than $120K, Unknown).

TABLE II. NUMERIC FEATURES

Feature Name	***Feature Description***
1. Customer Age	Customer's Age in Years
2. Dependent count	Number of dependents
3. Months on book	Period of relationship with bank
4. Total Relationship Count	Total number of products held by the customer
5. Months Inactive 12 mon	Number of months inactive in the last 12 months
6. Contacts Count 12 mon	No. of Contacts in the last 12 months.
7. Credit Limit	Credit Limit on the Credit Card
8. Total Revolving Bal	Total Revolving Balance on the Credit Card.
9. Avg Open To Buy	Open to Buy Credit Line (Average of last 12 months).
10. Total Amt Chng Q4 Q1	Change in Transaction Amount (Q4 over Q1).
11. Total Trans Amt	Total Transaction Amount (Last 12 months).
12. Total Trans Ct	Total Transaction Count (Last 12 months).
13. Total Ct Chng Q4 Q1	Change in Transaction Count (Q4 over Q1)
14. Avg Utilization Ratio	Average Card Utilization Ratio.

B. Data Pre-processing

Data preprocessing is a significant process of data mining in which transforms raw data into an understandable format. The techniques are included data integration, aggregation, transformation, dimension deduction and feature selection by dealing with unnecessarily data, irrelevance, noisiness and unreliability [11]. Which of these can improve an overall quality of the data. The importance of pre-processing is improving in terms of accuracy, completeness, consistency, timeliness and interpretability [12].

1) Irrelevancy

By summarizing the data information, for a feature that has nothing relevant to the prediction might have a negative impact on the performance of classifiers [13]. Therefore, the 'Client No' column must be dropped out from this experiment. Moreover, the dataset does not contain any missing values. Nevertheless, there are some 'Unknown' strings in the three columns (Educational level, Martial Status and Income category respectively).

In the Fig. 4, the bar charts illustrate the demographic information about the clients that there are approximately 83.93 percent of existing customer. Whereas, there are only 16.07 percent are attrited clients from the total of 10,127 clients. It can be noticed that data obviously shows the imbalance of the targeted parameter. Also, the overall of other bar charts are imbalanced.

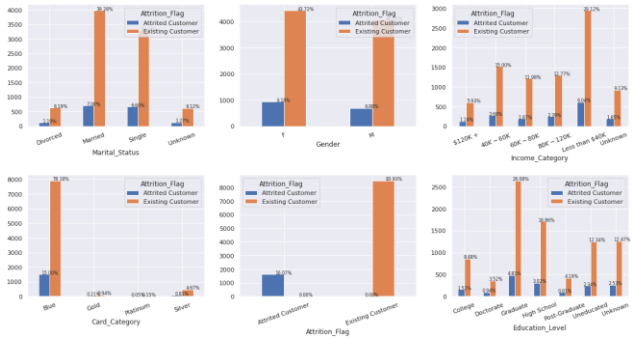

Fig. 4. Imbalanced Data (*Attrition_Flag & Others*)

Furthermore, in Fig. 4, the bar charts also demonstrate more information about the probabilities of the attrited clients. Firstly, 46.28 percent of churned customers are married. Followed by 52.9 percent of churned coustomers are female, 30.88 percent have graduate educational level, 35.15 percent of churned clients obtain less than 40K dollars for the income and 93.17 percent of those churners hold a blue card membership. Furthermore, for the 'Unknown' strings, there are 30 percent of the data. These strings need to be retained in the dataset. Otherwise, there will be lossing scale of data which could cause a bias intrepretation in a prediction.

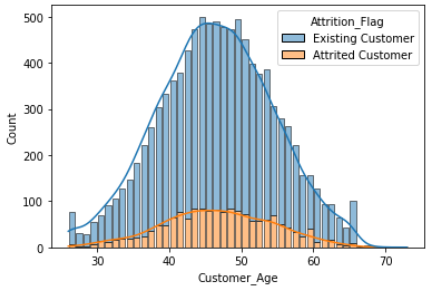

Fig. 5. Nomal Distribution (*Customer Age*)

In the Fig. 5, the histogram demonstrates a normal distribution of the customer age from minimum age of 26 year-old to the maximum of 73 year-old. However, the mean of customer age is 46 year-old. Nonetheless, the amount of attrited clients are at a very low scale.

2) Transformation:

Data transformation is the practice of converting data format or the structure of the data into a usable format [14]. By converting the data type, dealing with missing values and enriching the dataset.

For the string 'Unknown' values cannot be deleted because there are about 30 % of the data. Hence, by using Simple Imputer is an appropriated method to deal with missing vales in dataset. It is a class that can be used to impute with most frequent value by setting option as "most_frequent". It would replace the 'Unknown' strings with the most common value in the column. Afterwards, as the Attrition Flag column is a predicted parameter which is a binary variable as it can indicate whether customers are still retaining or leaving. Numeric '1' would be replaced by the attrited customers and '0' would be represented the existing customers.

3) Data Imbalance

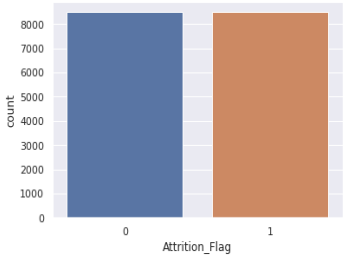

Fig. 6. Synthetic Minority Over-Sampling Technique (SMOTE)

It is very normal to face the barrier of imbalanced data when dealing with the binary classification. However, there is a common technique that has widely used is 'Resampling data'. Previously, as the target parameter in Attrition Flag has shown that there are the proportion of churned clients (16.07 %) and existing clients (83.93 %). The aim of oversampling technique is attempting the synthetic samples to generate the amount of the minority class [15]. In Fig 6, by applying SMOTE technique to deal with imbalanced data for overcoming with the overfitting when applying with algorithms in prediction.

4) Feature Selection

In machine learning, feature selection is the method of selecting the important parameters or highly dependent features. The purpose of selection phase is very optimal as it can assist with shorten the time for training and testing model as well as avoid the issue of high-dimensionality [16].

Correlation matrix is useful to measure and analyse the independency and dependency which displays the correlation coefficients between each feature [17]. In Fig 7., the feature that shows the highest correlation with target parameter is 'Total Trans Ct' (0.37). If the value is closer to 1 which means there is a possibility of a high confidence in correlation.

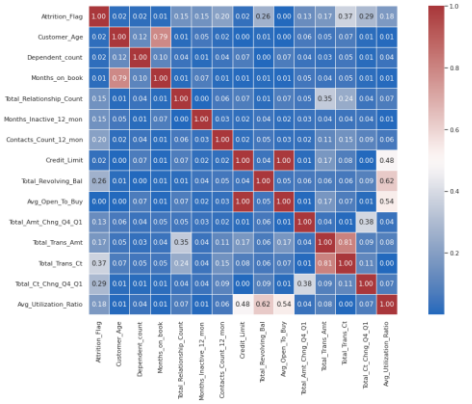

Fig. 7. Correlation Matrix

Importantly, before applying to machine leaning models. All type of input and output data are required to be in a numeric form. Therefore, all data value have to be converted. The encoder technique is a solution. As, there are some categorical data as in 'Income Category' column and 'Education Level' column. Thus, by applying with ordinal encoder can assist from categorical data to numerical value. For instance, the system operating as '1' for 'Uneducated', '2' for 'High School' and '3' for 'College' orderly. Furthermore, one-hot encoder is another method to assist with binary variables. The feature that is being one-hot encoded, it would be automated as numeric '1' and the others are turned to '0'. For instance, in Marital Status column, 'single' would be '1' and 'divorced' would be replaced by 0 as well as 'married' would be '0'. Hence, all of the selected features are now prepared for using in the next process.

C. Classification (Learning Algorithms)

After the step of pre-processing. The fifteen features are selected and prepared for the classification. The classification methods would be applied. For the supervised learning classifiers are Logistic Regression, Support Vector Machine, Decision Trees, Random Forrest and AdaBoost. In this experiment, the comparison for each model would be evaluated and discussed. Among this, 80 % of data would be processed for training and the remaining 20 % would be operated by testing as random for every models. Also, some learning algorithm techniques would be included in some models.

1) Logistic Regression (LR): Logistic regression is a predictive analysis algorithm which is used to calculate or predict the probability of binary event occurring [18]. It is a supervised learning. The observation is to discrete a set of classes. Also, it can be called as the sigmoid function. When using grid search parameter tuning in LR model. This function would find the best set of the machine learning model and parametrized by the grid of the hyperparameters. However, this experiment was also attempted to set the hyper-parameter space for the best cross-validation score which could improve in the overall performance of the model.

Nevertheless, the second attempt for this model experiment, by executing the model without using the hyper-parameter tunning. The result obviously showed the decreasing of the training and testing accuracy and the overall performance was slightly dropped (approximately 3%).

2) Support Vector Machine (SVM): Support Vector Machine is a supervised machine learning algorithm for two-group classification [19]. SVM is attempting to find an efficiency of hyperplane that precisely categorizes data points which means after inputting the data training set, the SVM would search for the best hyperplane and have a decision boundary to separate the features. The best position of hyperplane would have the maximum distance as the best margin size, a high size of margin is producing a good result in a robustness. Furthermore, there is a technique called kernel. The SVM kernel is a function to uplifting a low dimensional input space to a higher dimensional space [20]. The tunning parameter would effectively improve the overall performance of the model.

From the experiment, by initially testing kernel for finding the highest score, the result has shown that 'linear' scored 0.74, 'poly' scored 0.50 and 'rbf' scored 0.71. Thus, the linear would be used as a function for training and testing. However, in terms of time efficiency of this model is not pleasant as it took quite a while to progress the result. However, the result of accuracy training and testing were very suitable.

3) Decision Tree (DT): A decision tree is a supervised learning algorithm for both classification and regression tasks. The learning model is transforming a branch-like segment as a top-down scale [21]. However, if the model receiving an overwhelming weight of data, it could cause data fragmentation which leads to overfitting. From the experiment, by executing the training and testing the performance was acceptable and score for both training and testing accuracy was also normal.

4) Random Forrest (RF): A random forrest is an ensemble classifier for tree learners. The learning algorithm is a supervised learning which is constructed from decision tree algorithms. The aim is to combine several classifiers to provide outcomes of the prediction. Furthermore, bagging is an ensemble meta-algorithm method that is commonly used in RF as a technuiqe to transform and combine learning algorithms togeher which could reduce variance within noisy dataset [22]. The result of the Random Forrest seems to be the highest and the most efficient performance from the experiment. In Fig 8., the graph displays a difference between the accuracy of a decision tree (84%) and a random forrest (97%).

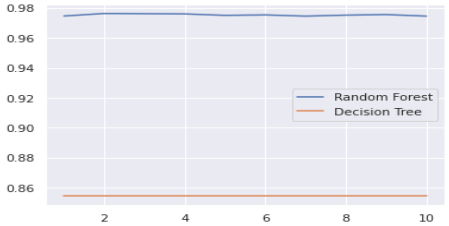

Fig. 8. Accuracy Comparing (Decision Tree & RandomForrest)

5) AdaBoost (AdB): AdaBoost algorithm or adaptive boosting is an ensemble classifier. The learning algorithm works by distributing and providing equal weights to all the data point. From the experiment, by setting the number of estimators to 'one-hundred', learning rate as 'one' and random state at 'forty two'. The model was generating a good result of performance as well as training and testing accuracy are very high.

IV. RESULT AND DISSCUSSIONS

After the performance of all supervised learning algorithms. Firstly, SVM presented 79.5 percent of the training accuracy and 79.6 percent of testing accuracy. The overall accuracy of the SVM model is 80 percent. The result is not pleasurable. Secondly, in LR model, the performance shows that with the training accuracy of 85 percent and testing accuracy of 84 percent. The LR model was finalized the accuracy for the prediction of 81 percent which is not high enough. However, the second attempt for LR was being applied by Grid Search and there is a slightly increasing in the number of performance (3% higher). Afterwards, the training and testing accuracy of a decision tree model is 85.5 percent and 84 percent respectively. And, the overall performance accuracy is 84 percent.

Next, the AdB model performed 96.3 percent of the training accuracy and 95.1 percent of testing accuracy. The accuracy result is 95 percent which is satisfied. At the end of the experiment with RF model, the score of the training accuracy is 99 percent and 97 percent for the testing accuracy. And, the model accuracy performed 97 percent for the prediction. So far, RF is the highest result in accuracy and the outcome is very satisfied. In the Table III. and Table. IV show the score for each algorithm models in accuracy, precision, recall and F1-Score for the existing customer and the attrited customer. The result of training and testing accuracy exposed that there are no overfitting for every models as applying SMOTE technique to deal with imbalance in the preprocessing step.

TABLE III. MODEL COMPARISON (EXISTING CUSTOMER)

Classifier	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
SVM	80	83	75	79
LR	81	81	81	81
DT	84	85	83	84
AdB	95	96	94	95
RF	97	98	96	97

TABLE IV. MODEL COMPARISON (ATTRITED CUSTOMER)

Classifier	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
SVM	80	77	84	80
LR	81	81	80	81
DT	84	83	85	84
AdB	95	94	96	95
RF	97	97	98	97

Due to the RF model has the highest overall score of performance accuracy among the other models. Hence, it would be taken as a consideration in terms of analysing on precision, recall and f-1 score to measure the customer churn prediction. The precision indicates how quality of the positive prediction is performed by the model. In the mathematical term, the precision equals to the value of true positive divided by the values of true positive add with the false positive. In the RF model, the precision shows the value of precision rate which is 98 percent for predicting existing customer and 97 percent for correctly predicting in attrited customers. Thus, the results are clearly robust and precise.

Additionally, the result of recall refers to the ability of measurement of how accurately the model is able to identify. Mathematically, the recall is equal to true positive value divided by true positive plus false negative value. F-1 score is the indicator or the mean for proving the balance between precision and recall. In short, the RF model has the highest accuracy based on the confident value of precision, recall and F-1 score in predicting to type of clients weather they are still retaining or leaving the service. Lastly, the feature importance in Fig 9. It can be used for analysing the insight reasons why attrited clients are leaving the service. As the numeric '1' is stated as attrited customers which could interpreted that the total transaction amount could be a factor for client churning.

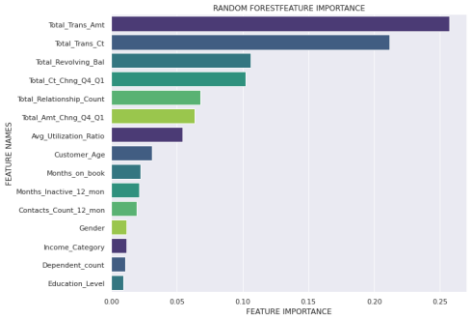

Fig. 9. Feature Importance (Random Forrest)

V. CONCLUSION

In conclusion, the importance of big data and machine learning are a core efficiency for a business in terms of bringing out the best accuracy for prediction rate. In terms of measuring a customer segmentation in churned client behavior, from the graph visualization in the pre-processing process found that women tend to have the possibility of churning than men. And, for the people who are at the age of 45 to 50 year-old tend to stop using credit card. And, the majority of both clients are holding the blue card membership. The concept and techniques of pre-processing can have an impact in the overall result of the model performance. The supervised algorithms are used to deal with binary classification. The random forest algorithm has the highest accuracy comparing with other learning models for the churn rate prediction. However, the banking manager can analyse the insight reasons by considering at the feature importance to improve the service for retaining customers.

REFERENCES

[1] Au T, Ma G, Li S. Applying and evaluating models to predict customer attrition using data mining techniques. Journal of Comparative International Management. 2003 Jun;6(1):10-22.

[2] Vives X. Digital disruption in banking and its impact on competition. Organisation for Economic Co-operation and Development (OECD). 2020.

[3] Singh H. The importance of customer satisfaction in relation to customer loyalty and retention. Academy of Marketing Science. 2006 May;60(193-225):46.

[4] Desmet D, Duncan E, Scanlan J, Singer M. Six building blocks for creating a high-performing digital enterprise.

[5] El Naqa I, Murphy MJ. What is machine learning?. Springer International Publishing; 2015.

[6] Wang H, Ma C, Zhou L. A brief review of machine learning and its application. In2009 international conference on information engineering and computer science 2009 Dec 19 (pp. 1-4). IEEE.

[7] Sagiroglu S, Sinanc D. Big data: A review. In2013 international conference on collaboration technologies and systems (CTS) 2013 May 20 (pp. 42-47). IEEE.

[8] Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, Gehrke J, Haas L, Halevy A, Han J, Jagadish HV. Challenges and opportunities with Big Data 2011-1.

[9] Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. Neurocomputing. 2017 May 10;237:350-61.

[10] Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences. 2017 Sep;12(16):4102-7.

[11] Berry MW, Mohamed A, Yap BW, editors. Supervised and unsupervised learning for data science. Springer Nature; 2019 Sep 4.

[12] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and information systems. 2012 Oct;33(1):1-33.

[13] Vijayarani S, Ilamathi MJ, Nithya M. Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks. 2015 Feb;5(1):7-16.

[14] Heer J, Hellerstein JM, Kandel S. Predictive Interaction for Data Transformation. InCIDR 2015 Jan 4.

[15] Torgo L, Ribeiro RP, Pfahringer B, Branco P. Smote for regression. InProgress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo, Azores, Portugal, September 9-12, 2013. Proceedings 16 2013 (pp. 378-389). Springer Berlin Heidelberg.

[16] Kumar V, Minz S. Feature selection: a literature review. SmartCR. 2014 Jun;4(3):211-29.

[17] Sisodia DS, Vishwakarma S, Pujahari A. Evaluation of machine learning models for employee churn prediction. In2017 international conference on inventive computing and informatics (icici) 2017 Nov 23 (pp. 1016-1020). IEEE.

[18] Thabtah F, Abdelhamid N, Peebles D. A machine learning autism classification based on logistic regression analysis. Health information science and systems. 2019 Dec;7:1-1.

[19] Noble WS. What is a support vector machine?. Nature biotechnology. 2006 Dec 1;24(12):1565-7.

[20] Jijo BT, Abdulazeez AM. Classification based on decision tree algorithm for machine learning. evaluation. 2021;6:7.

[21] Suthaharan S, Suthaharan S. Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning. 2016:207-35.

[22] Hussain MA, Bhuiyan A, D. Luu C, Theodore Smith R, H. Guymer R, Ishikawa H, S. Schuman J, Ramamohanarao K. Classification of healthy and diseased retina using SD-OCT imaging and Random Forest algorithm. PloS one. 2018 Jun 4;13(6):e0198281.