Predicting German Credit Risk

SEP 767 Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Prathamesh Joshi – joship14@mcmaster.ca

Master of Engineering – Systems and Technology

Student ID – 400485705

Under the guidance of Professor Dr Brandon Corbett – corbeb@mcmaster.ca and McMaster University

1. Abstract

Credit Risk for bank customers has gained tremendous attention in recent years. Typically, most banks throughout the world earn money by giving credit loans. By analysing past lending data, the risk of banks providing loans to customers can be minimized. Modern data mining and machine learning approaches have been discovered to be extremely beneficial and accurate in predicting credit risk and making the right decisions. This report provides univariate, bivariate and multivariate analysis of the dataset, an implementation of principal component analysis (PCA) on the dataset and implementation of various machine learning classification algorithms to predict the credit risk. The outcome turns out to be that the XGBoost classifier works the best on the given training dataset of 75% and testing dataset of 25% with an accuracy of 78.4% and is compared to other models in terms of accuracy, area under the curve (AUC), training and testing times respectively, resulting in effectively improving the decision-making process of lenders.

2. Introduction

Background and Literature Review

Nowadays, assessing credit risk has become an essential task in the banking sector. Financial institutions have started considering customer's personal data, credit history, living status,

employment status etc to make decisions on allowing credit to new customers, increasing credit limits or granting a loan to avoid financial loss to the bank. Traditionally this risk evaluation was done solely by human judgment to recognize whether the applicant is a good credit risk or bad to the bank. Today, machine learning and data mining techniques have been developed to assist in financial decision-making based on big data, in order to improve predictive accuracy. Credit scoring models use information about applicants to determine whether they have good credit or bad credit. Good credit means a person is likely to pay back a debt, while bad credit means a person is more likely to default on a debt [5]. Efficient and accurate credit scoring models are vital for banks and financial institutions for two main reasons. First, credit scoring models can help banks identify high-risk borrowers, which can help them avoid financial disasters. Second, accurate credit scoring can help banks make more informed decisions about the loans they offer and the products they sell, which can lead to increased profits. One way to reduce credit risk is to use data mining techniques to find useful and meaningful information from the datasets [7] and classification models for the predictive part of data mining. Classification is the process of grouping records together into a set of similar categories.

A. Approval of Credit from Bank

Credit is usually given in stages, starting with the application process and ending with the monitoring of the loan. Throughout this process, the bank must be careful not to grant too much credit or too little, as this could have a negative impact on the bank's overall wealth. The bank's goal is to help the customer by granting them the credit they need, while at the same time protecting the bank's own interests. Credit comes from the Italian language, namely Credere, which means to trust. In order to trust someone, we must first have a good understanding of them and what they can and cannot do. This is where credit comes in. Credit is a way of trusting someone, or more specifically, a loan that is given to someone in order to allow them to

purchase something or pay for something that they may not have been able to afford otherwise[6].

B. Credit Risk Prediction Classification

Classification models are often used in credit risk assessment, helping to find relationships between attributes in grants/loans such as history, number of jobs, yearly income, home ownership status and various factors [8]. Various machine learning models and classification algorithms have been used to predict credit risk which includes likes of Naïve Bayes, Decision trees, Logistic regression, K-NN, K-means clustering and Neural Networks [9,10]. Yoga Religia et al [3] propose a Random Forest algorithm on the South German Dataset to classify into good and bad credit where they get around 72.5% accuracy on the train-test split of 75%: 25%. Malekipirbazari M et al [11] too compare the performance of various algorithms and conclude that Random Forest Algorithm works the best in helping lenders make better investment decisions. Pandey, T.N. *et al* [10] survey proposes that ELM- a single layer feed forward network works the best for classification of good and bad credit risk.

C. Algorithm Performance Testing

When it comes to classification, it is very important to how one evaluates the performance of a machine learning model. Various assessment indicators that are commonly used are Accuracy, precision, recall, the area under the curve (AUC) etc. Hossin M et al [12] suggest confusion matrix, accuracy, sensitivity, precision and recall to be the best way to evaluate a binary classification problem. The confusion matrix is a matrix consisting of predicted and actual values in the form of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions. Accuracy is the ratio of correctly predicted labels to the sum of all the correct and falsely predicted labels. F-measure gives the harmonic mean between recall and precision values[12].

3. Database Description

The dataset was taken from the website www.kaggle.com in the form of a CSV file - German Credit Risk [1]. The dataset consists of 10 attributes namely 'Id column', 'Age', 'Sex', 'Job', 'Housing', 'Savings account', 'Checking account', 'Credit amount', 'Duration', 'Purpose' and 1 target variable – 'Risk' which we want to predict. The attribute Age(quantitative) contains information of the customers, Sex(categorical) describes the gender whether male or female, Job(categorical) contains occupational information in the form 0/1/2/3 number of jobs, Housing(categorical) includes the housing status whether they own/rent/free, Saving Accounts(categorical) contains savings classified as little, moderate, quite rich, rich, Checking Accounts(categorical) is same as previous one, Credit Amount(quantitative) contains how much credit the bank customers got, Duration(quantitative) is the specified time to pay the credit, Purpose(categorical) shows the reason for loan and is classified as car, furniture / equipment, radio / TV, domestic appliances, repairs, education, business, vacation / others. Risk is the target which we want to predict Good - 1/Bad – 0 risk for bank. The currency is in Deutsche Mark (DM) and there are 1000 records and only missing values we have are in Savings and Checking account.

4. Objectives

The 3 main objectives of this project are data analysis, exploratory data analysis, principal component analysis (PCA) of the dataset and applying machine learning classification algorithms to predict the credit risk. Data analysis is done to answer a few questions about the general trend from the dataset and estimate central tendency i.e., the mean and median or quartile ranges for numerical attributes. Exploratory Data Analysis (EDA) is done to see what the data can reveal about itself and provides a better understanding of the relationship between the dataset attributes/variables using visualization methods. PCA is performed on the dataset

to understand the correlated variables which might result in poor accuracy of the ML model but without losing the essence of these variables. Lastly, Classification algorithms will be applied to the dataset to see which performs the best in terms of various evaluation factors.

5. Methodology

A. Data Pre-processing and EDA

The dataset consisted of the 'id' column which was removed as it doesn't provide any significant information rather than just naming the sample. There were null values in the Savings Account and Checking Account attributes of the dataset. Those were handled by adding 'no-info' and making it another category as dropping them would result in a significant loss of data also both the attributes contain categories as mentioned in section 3. Univariate, Bivariate and Multivariate analysis was then done by plotting graphs to find outliers, trends and patterns in the dataset. For PCA and Classification, all the categorical attributes were label encoded. Label encoder was used compared to the one-hot encoder as label encoded keeps the ranking between class labels example little as 0, rich as 2. Also, the disadvantage is that for high cardinality, the feature space can really blow up quickly if there are too many categories in each attribute and you start fighting with the curse of dimensionality [20].

B. PCA

The dataset is then mean-centred and scaled using sklearn's built-in library called Standard Scaler [Figure 18 shows scaled data]. It is carried out using the equation 1.

$$x_k = \frac{x - mean}{standard\ deviation} \quad (1)$$

After that PCA is applied and visualization of principal components is done.

PCA helps in dimensionality reduction, as it converts a set of correlated variables to non-correlated variables. PCA with 7 components is applied to get the maximum variance explained by the principals. Interpretation of loadings for 1st principal component is done using visualizations and loadings as well as score plots. Interesting trends are found in the dataset which will be discussed in section 6.

C. Data splitting

Afterwards, dataset is then split into training set and testing set using the train-test split method for training the model. The data was divided into 75% for training and 25% for testing. Other parameters which were passed while splitting the dataset were random_state and stratify. Random_State is the object that controls randomization during splitting and stratify helps to homogenously split the data as much as possible based on our target variable. In other words, it will help in dividing equal proportions of good and bad credit samples in train and test sets.

D. Classification

For better and more reliable credit risk analysis, a comparison between different machine learning algorithms has been done. In this project the credit risk was predicted using five models which are as follows.

I. Logistic Regression

Logistic regression is a statistical analytic method that uses prior observations of a dataset to predict a binary outcome, such as yes or no. A logistic regression model forecasts a dependent variable by examining the connection between one or more existing independent variables. For example it will tell us whether the customer has a good credit or bad credit. The logistic function is given by equation 2.

$$f(x) = \frac{1}{1+e^{-(a+bx)}} \quad (2)$$

II. K-Nearest Neighbour (KNN)

KNN is the non-parametric method used for classification and regression. It includes a training set that contains both positive and negative cases. For real valued input variable, the Euclidian Distance is used. KNN is a popular algorithm used for classification. It can be calculated as the class with the highest frequency for the k-most similar instances. Each instance effectively votes for their class, and the class with the most votes is taken as the prediction [16].

III. Decision Tree Classifier

A decision tree is a prediction model that connects observations about an item represented by branches to conclusions about a target value represented by leaves. Each leaf node in the tree is labelled with a class or a probabilistic distribution across the classes, and the terminal nodes provide the ultimate value of the dependent variables [17].

IV. Random Forest Classifier (RFC)

RF is a classification and regression ensemble learning system. Bagging is used to pick samples from the original dataset. Then, from a pool of actual features, x features are chosen at random. One of these x features is chosen as a candidate split node. At each split, a new set of qualities is chosen. Splitting is continued until a decision tree is finished at depth d. A random forest is created by constructing a huge number of such decision trees. Each decision tree selects the class label for each new instance (vote). The votes cast by the multiple decision trees are then tallied, and the class with the greatest number of votes is believed to be the predicted class [18].

V. XGBoost Classifier

XGBoost is an implementation of Gradient Boosted decision trees. Here, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights

are assigned to all the independent variables which are then fed into the decision tree which predicts results. Boosting is an ensemble modelling, technique that attempts to build a strong classifier from the number of weak classifiers. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model [19].

F. Performance Metrics

A variety of performance indicators, such as predicted accuracy, confusion matrix, and AUC metrics, could be used to report the performance of credit scoring classifiers [12]. The accuracy can be measured by the following equation.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Confusion matrix is also used for binary classification problems. It is determined by using the built-in scikit learn library [12]. It is represented by the following matrix table

Predicted classification

	Positive Class	Negative Class	
Positive Class	True Positive	False Negative	Actual Classification
Negative Class	False Positive	True Negative	

AUC, Area under the Receiver Operating Characteristic (ROC) curve is the most promising metric for binary classification & in the field of credit scoring. This curve plots two parameters:

$$True\ Positive\ Rate\ (TPR)\ or\ Sensitivity\ or\ Recall = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate\ (FPR) = 1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$$

While precision is also used to determine how correctly the labels are classified. It is given by

$$Precision = \frac{TP}{TP + FP}$$

Although Accuracy does play a role in telling which algorithm works the best. F1 score is the more used technique to compared various classification algorithms [12].

F1 Score = 2(Recall * Precision) / (Recall + Precision)*

6. Results and Discussion

The project evaluation was done in three main sections EDA, PCA and Classification. EDA was done to investigate the data and get critical insights using graphical and non-graphical means. PCA was carried out on the dataset to emphasize variations and find patterns in the dataset. Lastly, various machine learning classification algorithms were used to predict the credit risk and a comparison was made between them based on different performance metrics.

A. EDA

The first step was to drop the id column in the dataset as it will not provide significant information. The next was to see the data type of each feature and how many missing/null values are present in each variable. This is shown in [Fig 1]. Following this was to analyze the data using the pandas describe method. However, only numerical columns are described by this method and not the categorical columns as you can see in [Fig 2]. So, we can see that the age column has a mean of 35 years meaning, the majority of the people who previously took credit from the bank were of this age. While the minimum and maximum aged for customers who took a loan are 19 and 75 years old respectively. Similarly, from Jobs [Fig 2] we can see that the mean was 1.9 ~ 2 which means the people in the data were skilled having at least 2 jobs, while the minimum and a maximum number of jobs were 0 and 3 respectively. The mean credit amount which customers took was 3271 DM while the maximum that someone took was 18424 DM and the minimum was 250 DM. This makes it kind of an outlier but just one observation. The mean Duration [Fig 2] for which bank customers took loan were 20 months while the maximum was 75 months. The 25th and 75th percentile are 12 months and 24 months

meaning much of the duration data is saturated in a very short range and the same goes for the credit amount. The number of unique values per feature is shown in [Fig 3] with Age having 53 different values, the housing, savings and checking accounts, Purpose for loan, all have unique values which means they are categorical features. The count per each unique value is shown in [Fig 4]. From this, we can see that males at 690 of 1000 samples took loans more than females. The majority of the people who took loans own the house at 713 counts. One major insight that we get from this is that many of the people were having little money in their savings accounts that's the critical point of purchasing a loan in the first place. The Checking account has significantly large no-info values which is fine as it will become a category label while doing classification. [Fig 4] One can also see that many people who took loan was for everyday utilities like car, radio or television. Post some data analysis we move on to graphical methods to visualize and understand the data better. We can see from [Fig 5] that majority of the credit amount is being saturated in the 1000 to 5000 DM. But as we can see it follows a right-skewness distribution meaning as we go to the positive side the values are increasing but the count of them is decreasing. Meaning there are very less people who took a higher credit amount than 6000DM. This figure also makes our conclusion right for the duration of months people took loans is clustered around 5 to 30 months. Most of the people taking loans are between the ages 20 to 50 which makes sense, as we can hypothesize that below this age range many will not be having jobs so they did not take a loan and above that many people have far fewer requirements to take credit from banks. From [Fig 6] histogram and boxplot we can conclude far more males take loans than females, and more young females take credit than young males. Males taking a loan are concentrated between the late 20s and mid-40s being the upper quartile range while females have a range of early 20s to mid-30s. We can also see some females taking more loans at a later age than the maximum value of the boxplot. [Fig 7] shows a count plot of our target in a 70:30 ratio of good and bad risks. A bivariate analysis of how

our risk is associated with sex can be seen in [Fig 8]. Males account for more good credit than females which makes it obvious as 69% of the total samples account for males. [Fig 9] shows a boxplot of Risk vs Credit amount and Risk vs Duration. From this, we can conclude that as the credit amount increases above 3000, the chance of bad risk increases meaning the person may not be able to repay the loan. Duration follows a similar trend as we can see, as the duration for which the credit is taken increases the bad risk increases. As we can see the median of bad risk is equal to that of the upper quartile range of good which means duration must be having a very high negative correlation with risk. [Fig 10] gives job distribution with sex. It shows that almost 600 skilled professionals among the 1000 people asked for a loan. This accounts for 63% of all credit applications received. More males have skilled jobs compared to females. [Fig 11] shows housing and its associated risk plot. It shows that 527 customers own the house, hence took loans and had good credit which is actually true. However, there are people who might not be able to pay off the loan so has a high count of bad risk too. More than half of the people who rent their place had a bad risk. Afterwards, we plot the Savings and Checking account with Risk [Fig 12]. Customers with little savings and little money in checking accounts request loans at a higher rate than customers who have more or better savings. One more unique observation we see in the Checking account is the no-info category. Many people who took credit from the bank but had no info are surprisingly a good risk. This can be noted from [Fig 13] with many good risks – no info customers taking less amount of credit. When we analyze the purpose feature of the dataset, we understand that the majority of loan applications are for automobiles. This accounts for 33.7% of all applications. Surprisingly, radio/TV is the second most prevalent reason for loans at 28%. A trend of males purchasing more loans for cars with good credit can be seen while for females more than half of those who take loans for automobiles are bad credit [Fig 14,15,16]. Moreover, after label encoding the values, we create a heatmap of correlation [Fig 17] to see how our features are correlated to each other. We notice

that Duration and Credit amount have a major impact on our target (Risk). Duration is highly negatively correlated, meaning as the duration increases, the chance of good risk decreases in our case we go close to 0 which is a bad risk. Similarly, as credit amount increases, chances of bad risks are high too. On other hand, Checking and Savings account have positive correlation.

B. Principal Component Analysis

PCA also called Principal Component analysis was performed on the dataset as it helps us to understand more about the data and as the first few components retain the maximum variation it helps us to visualize and summarise the feature of original high-dimensional datasets in low-dimensional space. After mean centring and scaling the data, we see that all the values are now in a similar range which is helpful so that one feature does not dominate the other if its values are large [Fig 18]. The First 2 and 3 components were added to see how much variance of the data is captured and it is easier to visualize the 2d and 3d plots [Fig 19]. By adding 1st and 2nd component a cumulative 37% of the variation was only explained by the components. While adding a 3rd saw it jump to 50.23%. To get a maximum variance on the dataset I added 7 components which explain a variance of around 88.8% [Fig 20,21,22]. It means we notice that it takes 7 PCs to explain this amount of variation in the data. To understand which features are contributing more we take the loadings and plot them [Fig 23]. PCA Loadings range from -1 to 1. Loadings close to -1 or 1 indicate that the variable strongly influences the component. Loadings close to 0 indicate that the variable has a weak influence on the component. In our case variables, 6 and 7 mean, credit and duration have a strong influence on risk, as concluded by the loadings plot for 1st PC. A PCA Biplot is a combination of a PCA score plot and PCA loadings plot. The length of PCs in the biplot refers to the amount of variance contributed by the PCs. The longer the length of the PC, the higher the variance contributed and well represented in space. [Fig 24, 25] shows the biplot. The interpretation from it is Housing has a large negative value on the PC1-y axis which means it might not affect the target but if it

influences it will be inversely proportional. Similarly, Credit amount, duration and job have a large positive value on the PC0-x axis which means it tells us the component focuses on negative risk offered by them if the value of any of it gets too large. We interpreted the same in EDA. Checking, Savings Account, Age and Sex have high values on PC1 which means all are positively correlated to risk, we also saw this in [Fig 17]. This biplot also gives us a few outliers, which I investigated but kept in the model-building stage to give a bit of variation for the algorithm otherwise it will start to overfit on either of the class due to the removal of some of the significant important samples.

C. Classification Results

Based on the train and test split of 75%:25% respectively five machine learning classification algorithms were tested. [Table 2] shows the model comparison based on Test Accuracy, Random Forest and XGBoost Classifiers works the best. The training and testing times of all five algorithms were also calculated with the Decision Tree Classifier being the fastest and Random Forest Classifier the slowest [Fig 26,31,36,41,46]. The classification report gives the Training and testing accuracy of all the five algorithms, where the lowest training accuracy on the training set was achieved by Logistic Regression at 73.2% and the highest by Random Forest Classifier at 100% [Fig 27, 32, 37, 42, 47]. This report also gives the precision, recall and f1-score for all algorithms where the highest precision was achieved by Random Forest and the lowest by Decision Tree. The highest Sensitivity/ Recall value which is also called the true positive rate was achieved by KNN and lowest by the Decision Tree. The F1 score is a weighted average of precision and recall. It is usually more useful than accuracy, especially if you have an uneven class distribution which we have right now. The highest f1 score is for XGBoost Classifier at 86% and hence it works the best to manage to detect all TP, FP, TN and FN perfectly. The confusion matrix [Fig 29, 34, 39, 44, 49] tells us how the algorithm has performed on detecting actual and predicted values. As random forest and XGBoost are the

best performing classifiers we investigate them and conclude that XGBoost works best in getting TP and less of FN meaning as a bank we don't want to lose a customer by detecting the person as a bad risk but actually, it was good. At the end of the day, the financial institution will have more loss when it does not give out loans to anybody. While False Positives mean that model detected it as a good risk but was actually a bad risk, yes, the bank might lose money according to the classifier but it can be avoided by placing a human decision factor. Area Under the ROC curve is the highest for XGBoost Classifier at 68.2% and lowest for the Decision Tree classifier. [Table 1] gives which features played an important role in determining the credit risk by the two of the best performing algorithms. The XGBoost Classifier tells the important features exactly as concluded in all the stages of EDA, PCA and Evaluation metrics so this project says XGBoost Classifier works the best when it comes to detecting Credit Risk.

7. Conclusions and Future Work

A credit scoring model is a method of measuring the risk associated with a potential client by assessing his data to determine the likelihood of him repaying his debts to the financial institution. This report summarizes the important trends, patterns and generates insights from the data using exploratory data analysis, followed by principal component analysis to visualize the variance in the data and give the important features which influences the credit risk factor. Lastly, it combines many classification approaches to find the best algorithm to determine a customer being a potential good or a bad credit risk to the bank. The test carried out showed the following important results:

1. People between the ages of 20 and 40 are more likely than others to seek for credit.
2. Customers with their own properties/houses are more likely to apply for a loan.
3. When a loan is taken for a duration greater than 25 months then the probability of being a Bad Risk is high as compared to being Good.

4. Savings, Checking accounts, Duration and Credit amount play the major role in determining the Credit Risk associated to the customer from PCA and EDA.
5. The highest accuracy as well as the f1 score was achieved using XGBoost algorithm, while the lowest was from Decision Tree Classifier.

Although, this study provided good results, I do think more exceptional results can be found out if the dataset is huge, also additional features like employment duration of customer, property value of the customer, instalment rate, etc which could play a huge role in determing good results for classifying good or bad credit. Lastly, it might be a possibility that an artificial neural networks-based model might perform better for this task as it can handle huge data as well as extract more features from it. Findings are going on for the same currently, as it is a very hot topic for almost all financial institutions throughout the world.

8. References

1. https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk
2. Sivasankar, E., Selvi, C. and Mahalakshmi, S. (2020) 'Rough set-based feature selection for credit risk prediction using weight-adjusted boosting ensemble method', *Soft Computing - A Fusion of Foundations, Methodologies & Applications*, 24(6), pp. 3975–3988. doi:10.1007/s00500-019-04167-0
3. Yoga Religia, Gatot Tri Pranoto and Egar Dika Santosa (2020) 'South German Credit Data Classification Using Random Forest Algorithm to Predict Bank Credit Receipts', *JISA (Jurnal Informatika dan Sains)*, 3(2), pp. 62–66. doi:10.31326/jisa.v3i2.837
4. Arora, N. and Kaur, P.D. (2020) 'A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment', *Applied Soft Computing Journal*, 86. doi:10.1016/j.asoc.2019.105936.
5. Hamdy Abeer and Hussein Walid B. (2016) 'Credit Risk Assessment Model Based Using Principal component Analysis And Artificial Neural Network', *MATEC Web of Conferences*, 76, p. 02039. doi:10.1051/matecconf/20167602039.
6. Jimenez, Gabriel & Saurina, Jesus. (2003). Loan Characteristics and Credit Risk. Proceedings. 170-183.

7. W. Gan, J. C.-W. C. H.-C. Lin dan J. Zhan, "Data mining in Distributed Environment: A Survey," Wiley Interdiscriplinary Reviews: *Data Mining and Knowledge Discovery* , vol. 7, no. 6, pp. 1-19, 2017. DOI:10.1002/widm.1216

8. Huang, X., Liu, X., Ren, Y., Enterprise Credit Risk Evaluation Based on Neural Network Algorithm, *Cognitive Systems Research (2018)*, doi: https://doi.org/10.1016/j.cogsys.2018.07.023

9. Lessmann, S. et al. (2015) 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research*, 247(1), pp. 124–136. doi:10.1016/j.ejor.2015.05.030.

10. Pandey, T.N. *et al.* (2017) 'Credit risk analysis using machine learning classifiers', *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* [Preprint]. doi:10.1109/icecds.2017.8389769

11. Malekipirbazari, M. and Aksakalli, V. (2015) 'Risk assessment in social lending via random forests', *Expert Systems With Applications*, 42(10), p. 4621. doi:10.1016/j.eswa.2015.02.001.

12. M, H. and M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, [online] 5(2), pp.01-11. doi:10.5121/ijdkp.2015.5201.

13. Qasem, M. and Nemer, L. (2020) Extreme Learning Machine for Credit Risk Analysis. *Journal of Intelligent Systems, Vol*. 29 (Issue 1), pp. 640-652. https://doi-org.libaccess.lib.mcmaster.ca/10.1515/jisys-2018-0058

14. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

15. https://www.kaggle.com/code/parulpandey/analysing-machine-learning-models-with-yellowbrick/notebook

16. M. J. Islam, Q. M. J. Wu, M. Ahmadi and M. A. Sid-Ahmed, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers," 2007 *International Conference on Convergence Information Technology (ICCIT 2007)*, 2007, pp. 1541-1546, doi: 10.1109/ICCIT.2007.148.

17. Rokach, L. and Maimon, O. (n.d.). Decision Trees. In: *Data Mining and Knowledge Discovery Handbook*. [online] pp.165–192. doi:10.1007/0-387-25465-x_9.

18. Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:10.1023/a:1010933404324

19. Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. doi:10.1145/2939672.2939785.

20. https://stackoverflow.com/