National College of Ireland

Project Submission Sheet – 2019/2020

School of Computing

Student Name: Abdul Azadh Abdul Saleem

Student ID: x18203621@student.ncirl.ie

Programme: Master of Science in Data Analytics **Year:** 2020/2021

Module: Domain Application of Predictive Analytics

Lecturer: Vikas Sahni

Submission Due Date: 28/06/2020

Project Title: Prediction of Defaulters in Credit Card Payment

Word Count: 3170

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Abdul Azadh Abdul Saleem

Date: 28/06/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Defaulters in Credit Card Payment

Abdul Azadh Abdul Saleem
M.Sc. in Data Analytics
National College of Ireland
Dublin, Ireland
x18203621@student.ncirl.ie

***Abstract*—This research is focused on the predicting the credit card customers in Taiwan, who are likely to have less potential to pay their credit card bill in the forthcoming month, which could help the credit card lenders to take further steps to reduce such default payments from their customers and to enhance their customer relationship. The motive of this analysis is to build different prediction models and to compare and identify one, which performs better in predicting the customers who will not pay their bill on up-coming month.**

Keywords— Prediction, Customer, Credit Card, Default, RapidMiner.

I. Background and Scope

Every industry in this digital era has to pay more importance in collecting the details of the customer, which may help mainly for research on the Customer Behavior, which plays a most significant role in businesses, especially in Banking Sector [1]. With the help of conducting predictions on the customer behavior depending the past records, the company classify the customers and can undertake suitable steps to maintain their relationship with customers, because customers has the major role in deciding the success of the business. When it comes to lending loan for the customers by a banking company, it is important to check their past transaction records, net worth, and the current debts. This analysis on the customer of a Credit Card provider helps the business to take further steps like as follows,

- Provide alerts by message or any other means to the customer, who has the probability to miss their upcoming month bill payment.
- Take suitable steps to inform the customer about the benefits of payment of bills without any due, and about the credit score, which may help them in future if they need loan.
- Communicating the customers by call or mail to collect the information about any concerns that may act as an obstacle for their bill payment.
- Providing rewards, offers, coupons or discount to the customer which may increase their interest towards paying the credit card bill on time.

Additional to the above this analysis helps in maintaining the company's reputation towards the customers, which helps them to successful among their competitors.

The dataset [2] which is considered for this analysis is collected in Taiwan about the credit card clients from the April 2005 to September 2005. Information about the customer, like default payments, their credit balance, available balance, and the past payment records which includes the pattern of payment and the debt the debt balance for every month, and the factor to be predicted is whether the particular customer would pay their bill for the following month or may dodge the payment. In the year 2005, the credit card issuers in Taiwan faced crisis because they had issued credit card and debts to customers irrespective to their financial condition and other factors which has to be checked before lending the amount, in order to increase their share in the raising market, which ultimately had given them financial blow. The industry must face a hard time since the customers started to overuse of the credit card and started accumulating heavy credits on them which they couldn't afford. This situation affected the customers financial confidence and the reputation of the credit card industry among the people. In order to avoid such mishappens further this dataset has been collected for conducting research on the repayment ability of the customers in future which had helped them to gain an idea customer behavior and the lenders to limit the loan so that the customers has the ability to pay back the amount without affecting the agreement.

II. Goal of the Project

The goal of this analysis is to identify the customers with their repayment pattern and their past credits. The Banking sectors commonly have a record of the customers about their history of transactions, financial status and their credit scoring which may help the lenders to have an idea about their new customers, so that they can provide their service according to the customers status, this may help the industry to maintain a good value with customers and also can avoid customer retention. And, they project also concentrates on identifying the factor or the information of customers which help the analysis to precisely predict the customer behavior.

III. Ethical Concerns

It is important for an analysis that it must conducted with ethical concerns. Digitalization of things around us has been a boon to our lifestyle, and it must be kept in mind that that possibility of activities being tracked and recorded, with or without our consent or knowledge[3]. We have seen major industries have failed miserably and been sued since they did not follow proper ethics in their business. Ethics of utilizing a data analysis includes the methods that is followed to collect the data, the sensitivity of the data, purpose for which the data is collected and the protection level of the data storage.

As a researcher, it is important for the person to be conscious on maintaining the ethics in the research. The data that has been recorded should have no data that violates the rule of ethics. If the data involves containing the personal information of the sample people, then the it is important for the researcher to gain their consent.

The dataset [2] which has been chosen for this research is extracted from the publicly available open repository, where they multiple datasets which can be used for academic, research or any other purpose. This dataset has been collected for the purpose of building up the model that can predict the future payment probability of the customer based on their history of records, in Taiwan. The dataset here has no sensitive records of the customers that may affect the privacy of the subjects, the customers here are identified as the unique IDs.

IV. Strategy Applied

This project is based on building an effective prediction technique that can forecast the behavior of the customers before, and to take necessary actions to prevent it. The methodology that has been involved in this research is Cross-Industry Standard Process for Data Mining. This methodology has been chosen for this research because it involves the business importance of the analysis for which it has been conducted. This research is conducted is conducted on a focus to address the issue which has been occurring on the business and to find the measure to avoid them further. The dataset has the collection of factors about the customers, which has been recorded with the ideology of the purpose of utilization.

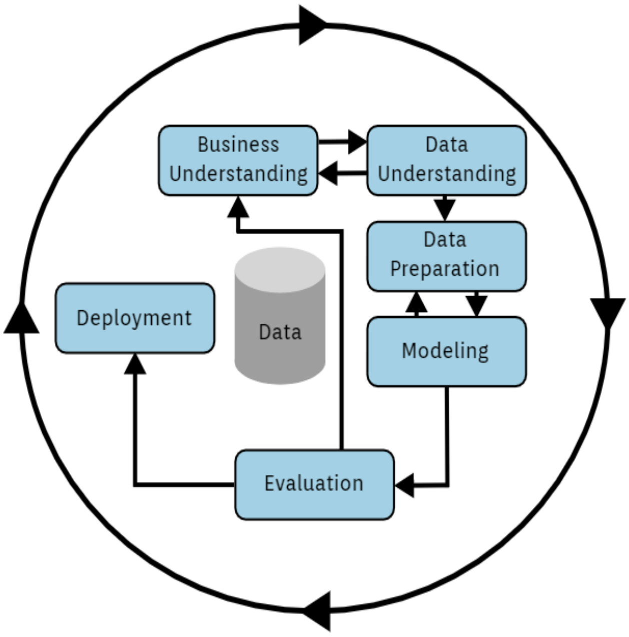


Fig. 1 Steps involved in CRISP-DM Methodology.

This CRISP-DM Methodology[4] involves six different stages that could help to successfully imply this research for the business purpose as shown in Fig.1.

1. *Business Understanding*

   As explained before in this report this research has been conducted to overcome the financial crisis that has occurred in the credit card company in Taiwan, which has excessively issued the card to the people, irrespective of their financial condition and other financial factors. The customers have also used the credit cards beyond the limit which they cannot afford to pay back. This has led to a condition to record the past payment records of the customers, for prediction analysis.

   This research may help the company to communicate the customers, who are probably become a defaulter for the future payment, and to help them pay the bill without any hassles. Along with this purpose, this research can help the credit card lenders to assign a credit limit for the customers depending on their records, which may help them from customer relationship being affected.

2. *Data Understanding.*

   The data which has considered for this research has been analyzed completely. The dataset does not contain any sensitive records of the customers that may affect their privacy. Dataset has information such as the payment made by the customer for the past six months, their level of education, marital status, gender and the current available balance and credit limit.

3. *Data Preparation.*

   The Dataset has been initially explored and by visualizing certain factors that may help to gain insights about the data. The data exploration has been explained on this report on the upcoming sections.

4. *Modeling.*

   This project focuses on developing a model that can analyze and predict the future occurrence depending the factors of prediction. This project involves comparison of multiple models has to be developed for analyzing the data and to select the one which produces high prediction accuracy and efficient.

5. *Evaluation.*

   The model developed will be initially analyzed with the test dataset will be provided to the model without the output value, which will be further used to compare with the prediction of the model and to evaluate the efficiency of the prediction model.

6. *Deployment.*

   This part of the methodology involves presenting the insights of the results to the business to prove the effectiveness of the research. This stage also includes documenting or preparing report for the analysis which may help the business in future.

In addition to this methodology, this research will be consciously conducted that it follows the following four key aspects, they are mentioned in the Fig.2

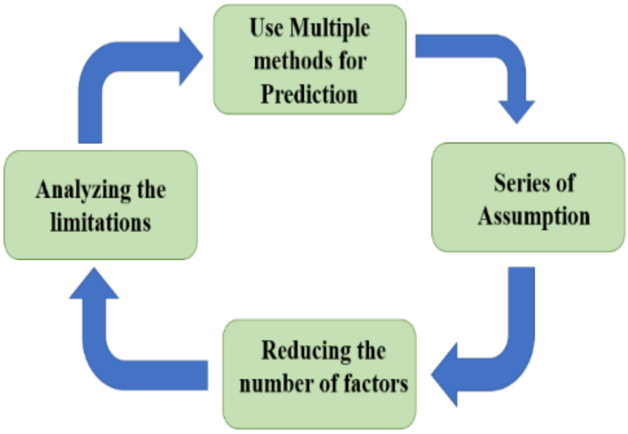

Use Multiple methods for Prediction
Series of Assumption
Reducing the number of factors
Analyzing the limitations

Fig. 2 Key terms of the research.

i. *Use Multiple Methods for Predictions.*

Since the accuracy and efficiency of the models on the dataset cannot be identified without being conducted, this project involves building multiple model following different data mining methods and the compared to get the best method that can provide the best prediction.

ii. *Series of assumptions.*

Conducting a research with more than one assumption has always proved to be the best in bring the outcome the project to the fullest. The research will be conducted considering different factors as the independent factors which affects the dependent factor. There has be a best-case scenario and a worst-case scenario to be considered in the prediction which may help the research to address any further issue that may occur after deployment.

iii. *Reducing the number of factors.*

This project also involves in identifying and neglecting the factors which do not significantly contribute for the prediction. This further helps in collecting of data in future where less or no efforts where implied on collecting the insignificant data of the customer.

iv. *Analyzing the Limitation.*

It is equally important for the researcher to identify the limitations of the research. This project involves identifying the scenarios where the model has less ability to produce the desired result. Here, the behavior of the customer cannot completely identify only with the information provided in the dataset. There is always a human factor involved, which cannot be predicted with this model.

V. Preliminary visualization.

The dataset [2] which has been considered for this research has information which has been analyzed by visualizing them before conducting the research which may help us to get an better idea about the content of the data. The preliminary visualization of the dataset has been done with the help of Tableau tool, which is a popular visualization software used to explore the data. The dataset has been imported on the tableau tool and it is found that the data has no missing values which could affect out prediction. The dataset has many factors which are considered for this analysis, starting with the 'ID' column, which has the unique identification codes for the individual customers.

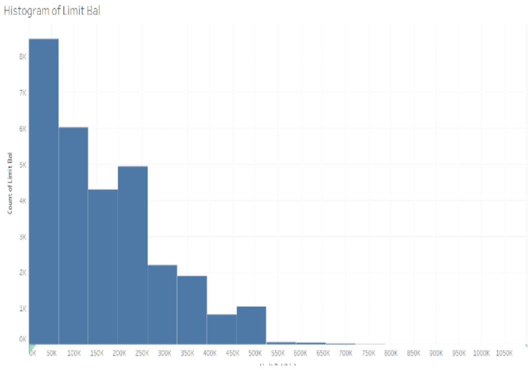


Fig.3 Histogram of Limit Balance of the Customers

The diagram Fig.3 has been plotted in histogram in order understand the spread Balance Limit for the population considered for this analysis and it has been observed that the histogram is skewed in one direction and not normally distributed. And from sample population, the highest number of people has their Limit Balance up to 50,000 New Taiwan Dollar (NT$) and the count gradually decreases as the Limit Balance increases, leading to least set of people who has limit balance between 650,000 NT$ to 725,000 NT$.

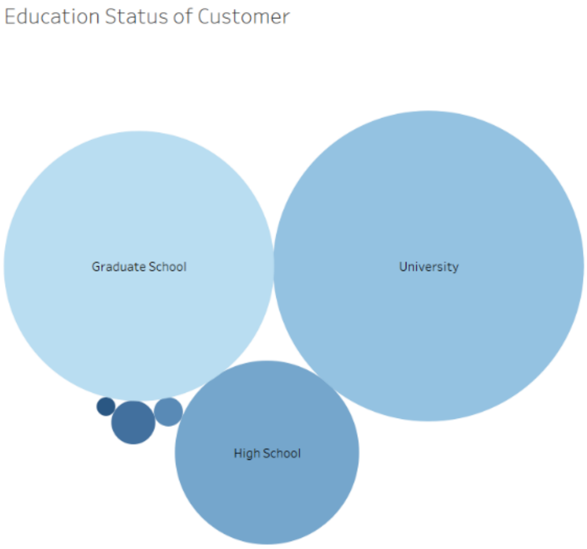


Fig. 4 Educational status of the Customers.

The level of education of the customers has been plotted with Bubble plot as shown in Fig.4, where the size of the bubble for every category infers the count of population in each category. Here I have been observed that Highest among the population are the graduates who are educated in university. There is also a constraint in this factor because there are some customers whose educational status is unknown, which may affect the prediction.

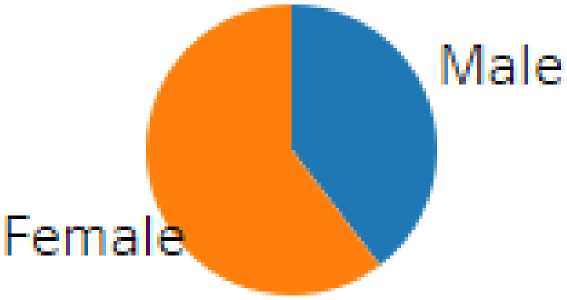
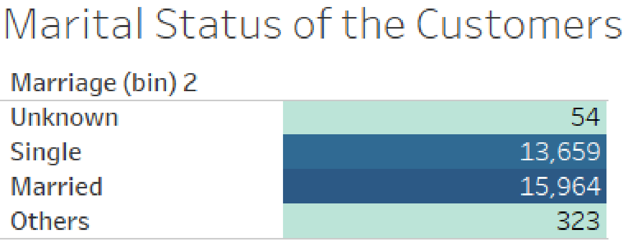


Fig.5 Gender of the customer

Amount of population in the sample is differentiated based on their gender and plotted with pie chart as shown in Fig. 5. Observation of this plot provides the insight that majority of the population considered for this analysis belong to the Female gender.



Marital Status of the Customers

Marriage (bin) 2	
Unknown	54
Single	13,659
Married	15,964
Others	323

Fig.6 Marital Status of the Customer.

This visualization provided in Fig.6, could provide an overall view of the Marital status of the customers in the analysis. The population has the highest crowd of people who are married. And here we have few customers whose marital status is unknown, but since the count is very low and they cannot make a significant different in the prediction, they can be neglected.

Further this section will cover the past credit record of the customer, which is visualized with the help of histogram which will help us to get insight about the spread of each value among the population. Bill amount, which is the amount the customer supposed to pay to the credit card lender for the six-month period which are considered for this analysis, has been shown in the following figures.

The histograms in Fig.7 below are plotted with the bin size kept for 50,000 NT$, in the Fig.7. On closer observation of each of the histogram it has been observed that most of the customers have their Bill amount up to 50,000 NT$, the highest bar near 0K in all the plots proves that.

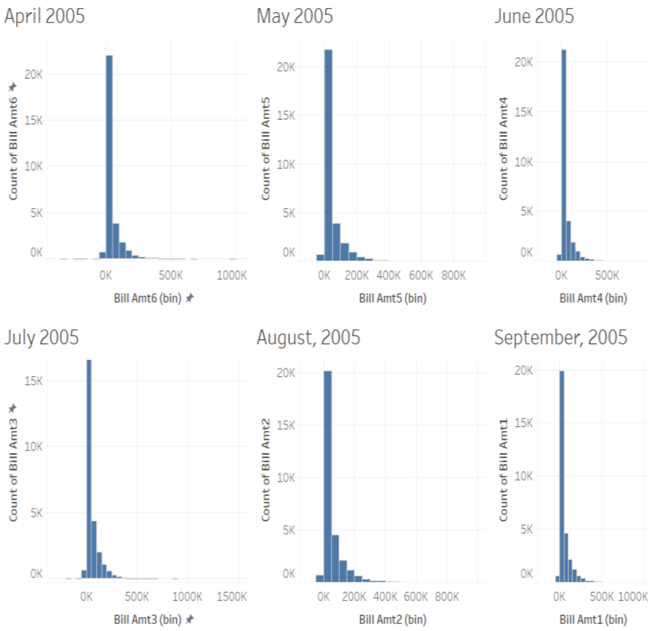


Fig.7 Histogram of the Bill amount of the customer for six-month period.

The payment pattern of the credit card bill has been categorized into multiple factors for the period of six months. '-1' represents 'pay duly', '1' represents 'payment delay for one month', likewise '2' represents 'payment delay for 2 months', following the same increment pattern the factors representing the delay for months increase till '9', where '9' represents the 'payment delay for nine months and above'. This repayment pattern is also represented with the help of histogram as shown in the Fig.8 and it infers that most of the customers pay their bills on-time or payment delayed by one month on all the months considered here in the analysis. And statistically the values are slightly distributed in normal way.

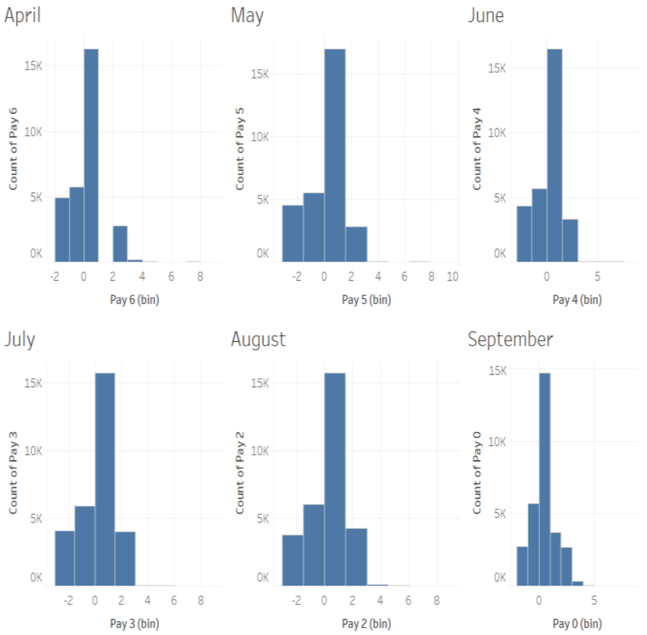


Fig.8 Histogram of Repayment pattern.

The dataset has been examined with the RapidMiner software, which has provided a key issue in the dataset that there is an imbalance factor to be predicted in the dataset, which has been shown in the Fig.10, where the factor '0' represents the customer will not be a defaulter in the payment of bill in the following month, and '1' represents the probability of the customer becoming the defaulter of the credit card bill payment. For and effective model to make a prediction, the dataset must be provided with equal amount of both the factors to be predicted. This issue will be solved at the data pre-processing stage of the developing the prediction model.

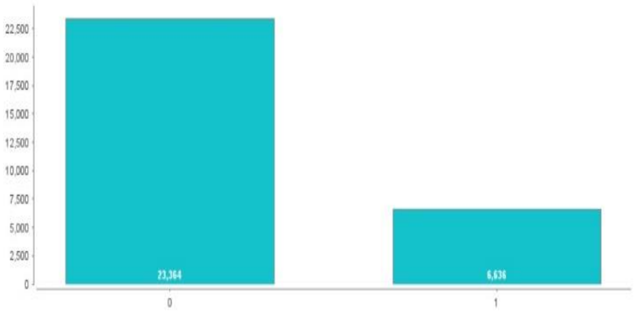


Fig.10 Visualization showing the data imbalance in the factor to be predicted.

VI. APPLICABLE TECHNIQUES.

This research includes building multiple models for the prediction and comparing the performance of the model for the data. The initial analysis of techniques that can be applied to this research has been done with the help of RapidMiner Tool, which has provided the following result as shown in the below figures Fig.11 and Fig.12

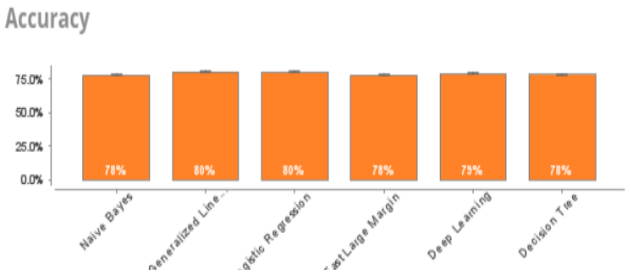
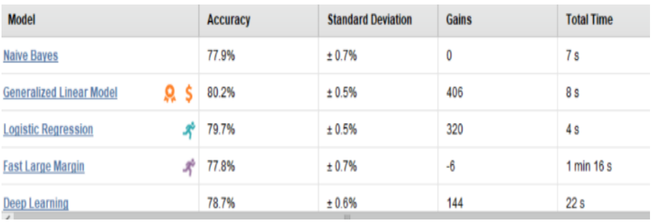


Fig.11 Bar plot showing the accuracy of the applicable method on the dataset.



Model	Accuracy	Standard Deviation	Gains	Total Time
Naive Bayes	77.9%	± 0.7%	0	7 s
Generalized Linear Model	80.2%	± 0.5%	406	8 s
Logistic Regression	79.7%	± 0.5%	320	4 s
Fast Large Margin	77.8%	± 0.7%	-6	1 min 16 s
Deep Learning	78.7%	± 0.6%	144	22 s

Fig.12 Comparison to applicable methods for the research.

From the above figure it has been observed that the Generalized Linear Model perform better in the dataset considered for this analysis, which is followed by the performance of Logistic Regression. The prediction in this research can be carried-out with the following methods and their performance is compared with the help Evaluation metrics and the model which performs better is utilized. From analyzing the data with RapidMiner software the following methods of Linear Regression and Logistic Regression were chosen for the analysis. In addition to this evaluation the additional methods like Naïve Bayes classification and K-Nearest Neighbors classifier has been chosen for this analysis as referred from the article [5].

- **Regression Model:** This model is applicable for this analysis because this method allows the flexibility in analyzing the response variables which are not normally distributed in our case. This model works based on the strength of impact of the input variables on the prediction values. In order to run this model, and to make successful prediction the A correlation plot has been plotted as a part of pre-processing and the independent variables which do not contribute significantly for the analysis will not be considered for the analysis.

- **Logistic Regression Model:** The reason for considering this method is that since the variable to be predicted is categorical and binomial, which makes this dataset suitable to be utilized with this model for analysis. The variables of the data must be converted into integer before the running the model, and further the variables were normalized such that standard deviations of each values will replace the values of each customers in the data. This model works in the way that a S-Shaped sigmoid curve will be considered from the values given which acts as a separator for predicted values of the model.

- **Naïve Bayes Classification:** This classifier works on the principle of the Bayes Theorem. With the help of this theorem, the probability of occurrence of events to be predicted will be found, depending on which the classification happens. Bernoulli Naïve Bayes Classification is the model which will be carried out for this research in the dataset, because this helps to predict the two parts of the occurrence, either yes or no, here since the prediction is identify whether the customer will pay the bill next month or not, which makes this algorithm to be suitable for the data.

- **K-Nearest Neighbor Algorithm:** KNN algorithm works on the basic principle that values with closer proximity will exist together, and predicts the value based on this region. When a sample of unknown has been given, this classifier tries to identify the pattern for the closest values from the sample. K here represents the number of neighboring factors that helps in decision making. KNN algorithm has the advantage of faster computation time than the other algorithm.

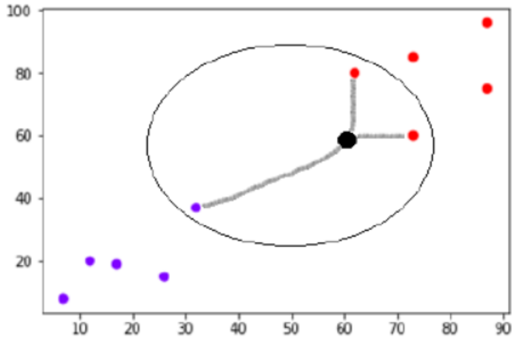


Fig.13 Working of KNN Classifier

This algorithm works on the method of plotting the values of the data on a plot as shown in Fig.13, and considering the neighbor values for the prediction depending on the K values and the prediction will be done with the category which has the majority of the votes. Normalizing the data will improve the performance of the algorithm.

VII. REFRENCES

[1] Y. Peng, G. Kou, Y. Shi, and Z. Chen, “Improving clustering analysis for credit card accounts classification,” in *Lecture Notes in Computer Science*, 2005, vol. 3516, no. III, pp. 548–553, doi: 10.1007/11428862_75.

[2] “UCI Machine Learning Repository: default of credit card clients Data Set.” https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (accessed Jun. 22, 2020).

[3] L. Van Wel and L. Royakkers, “Ethical issues in web data mining,” *Ethics Inf. Technol.*, vol. 6, no. 2, pp. 129–140, 2004, doi: 10.1023/B:ETIN.0000047476.05912.3d.

[4] “Cross-industry standard process for data mining - Wikipedia.” https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining (accessed Jun. 26, 2020).

[5] I. C. Yeh and C. hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 2473–2480, Mar. 2009, doi: 10.1016/j.eswa.2007.12.020.