Data Mining and Machine Learning Project

1st Dhruv Vimal Shah
Dept. of Computing
National College of Ireland
Dublin, Ireland
x21121087@student.ncirl.ie

***Abstract*—This is research on a machine learning project that used 5 different machine learning models across 3 different big datasets. The first dataset was from a credit card domain, and the other two were from the HR Analytics domain. The information was found on Kaggle. All the datasets will be used for classification problems. The first set of data was about people who didn't pay their credit card bill and were defaulted. Models KNN, Random Forest Classifier, and Logistics Regression were used. Logistics Regression worked best, with an accuracy of 81.46%. In the second set of data, Logistics Regression and Gaussian Naive Bayes were used, and Logistics Regression again worked best, with an accuracy of 78.01%. The third dataset is predicting if an employee who may or may not be looking for a new job. Random Forest Classifier and Decision Tree are used to analyze this data, with Random Forest giving the best results of 80.91%.**

***Index Terms*—Machine Learning, HR Analytics, KNN, Random Forest Classifier, Logistics Regression, Naive Bayes, Decision Tree**

I. Introduction

As a result of the many technological developments that have taken place over the years, internet banking is becoming more popular since it allows customers to make transfers, payments, and monitor their accounts while they are on the move, removing the need to visit a bank in person. As a result, banks now have a large amount of information about their clients' behavioural habits as well as their financial patterns. Aside from that, the credit card is one of the most important elements linked to the consumer. If a consumer fails to make his or her payments on time, there are significant hazards involved with using credit cards. A credit limit is issued to each client after his or her account information, financial situation, and past loan payback history are taken into consideration. As there is a significant likelihood of commercial problems associated with credit card failure, many financial institutions and banks are paying more and more attention to the problem of credit card default. Furthermore, because of the huge amount of house loans and credit card defaults in the United States, the subprime mortgage crisis [1] was declared in 2008, and it has since expanded to worldwide credit markets and financial systems, as well as had a significant impact on the essential economy in the previous few years. This demonstrates that even after taking all of the necessary precautions, there is still a possibility that the consumer will fail on his credit card payment. If the patterns of each client were closely watched, it should be possible to determine whether or not the customer would default on his or her payment. Machine learning algorithms may help detect and anticipate similar situations. I utilized K-Nearest Neighbor, Random Forest, and Logistics Regression to analyze data in this study. For credit card default, the research question is:

"What is the probability that a person's future credit card default can be forecasted based on their account information and historical bill payments?"

Human resource analytics for promotion has gained popularity in recent years. When an employee performs well, the individual is often promoted. This significantly assists in determining an employee's performance and efficiency; also, promotion adds to an employee's work happiness.Most firms and organizations have a systematic performance evaluation system where employees are evaluated periodically, generally once or twice a year. A very well performance evaluation system may help a company. Personnel decisions such as salary increases, bonuses, or firing of employees are made easier by knowing what is expected of them. Constructing an efficient system is not a simple process [2]. Employees may enhance their performance by tracking their performance over time. It takes time to train an employee for a certain function, particularly a manager or team lead since these roles require them to manage and lead a team. If the promotion is done incorrectly, this would result in a large loss of earnings for the firm. Thus, one of the most significant challenges in workforce analytics is selecting the best candidates for promotion. Final promotions are not disclosed until after the evaluation, resulting in a delay in transferring to the new designation. As a result, the organization needs help to find potential candidates at a certain step to speed up the overall promotion procedure. Machine learning algorithms may be utilized to detect the critical elements of an organization's future promotion forecasts. I utilized machine learning techniques such as logistics regression and Gaussian Naive Bayes to do this. For HR Analytics: Employee promotion data the research questions is:

"What are the factors that help determine whether an employee will be promoted or not?"

Almost every company's most important asset is its human resources department. Companies that deal with big data and data science are increasingly eager to hire data scientists who have completed the organization's training programs. Investing in learning and development throughout training is beneficial to both organizations and learners. Businesses have spent years selecting and developing these employees to compete in a growing market. A significant amount of time and money is invested in the recruiting process as well as in training the employee so that he or she understands how the firm operates and what he or she is responsible for. As part of their on-boarding process, most multinational corporations provide 3–6 months of training at no expense to the employee. However, some individuals choose to quit after completing the training program for a variety of reasons. As a result, if any of these employees leave the company, the company will suffer a significant loss. When an employee is dissatisfied with his or her employment or with the firm, he or she will resign from their position. The study found that most people leave their jobs if their work and home life are not balanced, which impacts their job performance. So they want to know whether they want to continue with the firm or search elsewhere. Thus, HR analytics was created. Businesses are increasingly using this area to study past employee behaviour, establish trends, and plan future strategies. A highly useful tool, this forecast developed using the Random Forest and Decision Tree Classifier machine learning models, would save the cost and time involved with teaching or developing courses and classifying candidates. For HR Analytics: Job Change of a Data Scientists, the research question is:

"What factors are utilized to determine whether or not an employee is presently searching a new work?

The following is a breakdown of the document's structure. A brief description of previous research in the subject is provided in Section 2 of this report. Specifically, in Section 3, we go through our data mining methodology. In the next part, Section 4, we used machine learning algorithms to the data and conducted an analysis to determine which methods provided greater accuracy, F1 score, precision, and recall value. We came to a conclusion on the results and observations in Section 5, and we spoke about what might be added in the future to make the performance even better in the long run.

II. Related Work

Research in the domain of credit risk prevention via credit default predictions has taken many different forms over time. In this field, Ruilin Liu [3] conducted a study in which several machine learning algorithms, including SVM, KNN, Random Forest, Decision Tree, etc., were examined to predict credit default. KNN and decision trees were found to be more accurate at 79.8%, while neural networks were found to be slightly more accurate at 82%. Another study by Admel et al [4], used credit card data to predict if a credit card holder would default, in which they tried multiple machine learning models to obtain the most accurate findings attainable. They used a variety of models, including KNN, Naive Bayes, SVM, and logistic regression, to analyze the data. They arrived at a conclusion in which they discovered that logistic regression and Nave Bayes provided the most accurate findings, with accuracy rates of nearly 80%. A further investigation by Yingying et al [5], attempted to determine the contributing factors to credit card defaults by taking into account both financial and social customer data. They concluded that the stability of funds, instead of the income of the customer, is more significant in determining credit card default. To get this result, they employed the Cox proportional hazards model. The authors, Yashna et al. [6], The variable's capacity to predict credit default was tested using machine learning approaches including logistic regression, random forest, and rpart decision trees. They divided the data into two groups: training (70%) and test(30%). The random forest technique has the best accuracy and AUC of all the algorithms examined. AUC for evaluating credit risk of credit card users was found to be 81.81% using Random Forest. [7], the paper's author, developed several machine learning models to test his idea, including the decision tree, logistic regression, random forest, and AdaBoost. As a result of the imbalanced data, he constructed appropriate weighted models in each model. The results showed that random forest had the highest accuracy, with an accuracy of 82.12%. It also met the objectives of having the quickest training rate, the most data volume, and a high degree of parallel processing. Amongst all the models, logistic regression had the lowest accuracy. In addition, the weighted machine learning model outperformed the unweighted models. Another research conducted by Jason et al [8]. found a link between credit-card defaults and flu infections in around 80 metro regions throughout the United States and discovered that when illness outbreaks occur in a given place, the defaults on credit cards rise. When they looked into the link between influenza-related Google searches and 30-day, 60-day, and 90-day credit card and loan default rates, they did ordinary least squares and fixed effects, as well as 2-stage least squares instrumental factor regression. Another research by Leow et al [9]. They employed the Mixture model to solve their problem. According to the findings, they found that taking into account the balance in the customer's account right from the beginning of the loan application procedure, rather than just looking at the amount after the default, increased the accuracy of projecting credit defaults. The author Oded et al [10] discovered that loan request applications that contain specific keywords are more likely to default on loan payments than those that do not contain those words. A variety of text mining and machine learning methods were utilized to analyse more than 12,000 loan applications, including word clouds, binary regression, and Naive Bayes.

A study from Abreham et al [11] was to investigate the relationship between work satisfaction and promotion practices. This case study made use of both primary and secondary data sources. The data from the employees was subjected to multi-stage sampling as well as simple random sample techniques.

TABLE I
SUMMARY TABLE OF RELATED WORK FOR DATASET-1

Authors & Year	Methodologies Used
Ruilin Liu (2018)	SVM, KNN, Random Forest, Decision Tree
Admel et al (2018)	KNN, Naïve Bayes, SVM, Logistics Regression
Yingying et al (2019)	Cox Proportional Model
Yashna et al (2018)	Logistics Regression, Random Forest, rpart Decision Tree
Yue Yu (2020)	Decision Tree, Logistics Regression, Random Forest, Adaboost
Jason et al (2015)	Ordinary Least Squares
Mindy et al (2016)	Mixture Model
Oded et al (2019)	Word Clouds, Binary Regression, Naïve Bayes

SPSS, as well as correlation and regression models and other methods of analysis, were used to examine the data. It was possible to forecast both positive and negative associations between independent and dependent variables. The regression model result in this case study revealed that independent factors explained 44.5 percent of the variation in the target variable. a result of these findings, the author reached the conclusion that employees perceived the bank's promotion practices as irregular and dissatisfactory in general and those employees were dissatisfied with the current promotion opportunities available at the bank in particular. Additionally, the findings of this case study reveal that effective promotion methods are an extremely important aspect in keeping an employee pleased. Yuxi et al [12] did a similar study, but this time with an emphasis on human resource management. They were successfull in obtaining data from a Chinese corporation, and they developed several characteristics and applied a machine learning model to the data as a result. According to the findings, the random forest model performed better and was validated based on its validity characteristics. In addition, the Gini relevance for each characteristic was determined, which demonstrated how each element affects the advancement of employees. Ultimately, research revealed that job-related characteristics had a greater influence on promotion than personal characteristics, but personal characteristics had a lower impact. Other factors, such as the length of the working year, the variety of positions available, and the higher level of the department, had a significant influence on employee advancement. The author of this research project, Liyuan et al. [13], provides assistance to the HR team in developing methods for managing people and making management choices. The purpose of this paperwork is to establish a framework for a corporation that will assist them in forecasting numerous aspects, such as staff turnover and employee satisfaction, among others. In order to construct this framework, the author used a variety of analytical approaches, including descriptive analysis, predictive analysis, and entity sentiment analysis. Employee turnover and employee satisfaction were predicted using a variety of machine learning techniques, including K-nearest Neighbors, Logistics Regression, Random Forest, and Gradient Boosting, which were employed in this research by the authors. They even employed SMOTE and ADASYN for data rebalancing in order to achieve this. Ultimately, the conclusion was reached that by using these models, the HR staff may gain greater knowledge about the company's workers, such as pay expectations, employee promotions, and other risks associated with the individuals in question. A clustering technique that falls under the domain of unsupervised learning was used by the authors, Sateesh et al [14]. The algorithm adopted was medium partition (PAM). The authors may examine whether or not these dimensions can lead to the selection of the "perfect fit" candidate by comparing multiple P-E fits. They achieve this by studying PAM algorithm output. They used this strategy in their study to choose relevant and suitable workers for employment, group, and company fit. A discussion on how the performance of workers determines whether a firm will succeed or fail is provided by the authors, Ananya et al [15]. Determining an employee's performance for the current year is aided by the decision tree approach. The author's method employed K-means clustering to classify employees based on their performance. The author's conclusion predicts the number of employees who will be considered for promotion as well as their performance.

TABLE II
SUMMARY TABLE OF RELATED WORK FOR DATASET-2

Authors & Year	Methodologies Used
Abreham et al (2022)	Regression Model
Yuxi et al (2018)	Random Forest Model
Liyuan et al (2020)	KNN, Logistics Regression, Random Forest, Gradient Boosting, SMOTE, ADASYN
Sateesh et al (2022)	PAM (Partitioning Around Medoids)
Ananya et al (2018)	K-means Clustering, Decision Tree

Through this article, the author, Tanya [16], assists the human resources staff in predicting why workers tend to quit the organization and the various causes for their departure. This study investigation was carried out utilizing a variety of machine learning approaches, including SVM, Linear Regression, and SMOTE. The accuracy of the model was determined by the author by concentrating on true positives. The author completed the research by providing exact data and advising that the management team should choose key elements that would aid in the development of various employee traits in different departments. It was discovered from this article that the author used SA-SVM with Bayesian Optimization, which provided the highest accuracy percentage when compared to the other models tested. An alternate approach to identifying employee resignation has been implemented in the work by Ana et al. [17], which differs from the standard approaches available. They have compiled a dataset from LinkedIn in which they have identified characteristics that cause an employee to quit their position. In addition, they employed several machine learning approaches such as back propagation, decision trees, and self-organizing maps, where supervised learning algorithms such as cross-validation score were applied with 10 folds in a single experiment. The decision tree performed the best out of the other options, with an accuracy of 88.4 percent, whereas the other options achieved an accuracy of roughly 50 percent and 56 percent, correspondingly. The mean of the Kappa scores for the Decision Tree was 0.34, SOM was 0.02 and Back Propagation was 0.03. Francesca et al. [18] used the Random Forest classifier, K-nearest neighbours, Support Vector Machines, Gaussian Naive Bayes, Naive Bayes classifier for multivariate Bernoulli models, Logistic Regression classifier, Decision tree classifier, and Linear Support Vector Machines to determine the elements that may lead to an employee leaving the firm and, most importantly, to forecast the possibility of specific workers leaving the organization. They analysed the data statistically before categorizing it. The data was handled by separating it into training and testing phases to ensure that the target variable had the same distribution. The projected outcomes were gathered and entered into the corresponding confusion matrices to assess the algorithm's performance. The fundamental parameters required for an overall assessment (precision, recall, accuracy, f1 score, ROC curve, AUC, and so on) could be calculated from these, and the best classifier to predict whether an employee was likely to quit the organization could be identified. The Gaussian Nave Bayes classifier achieved the best result for the given dataset: it showed a good recall rate (0.54) and attained an overall false-negative rate of 4.5% observations. The suggested automated predictor's results show that the primary attrition factors are age, monthly income, distance from home and overtime. Dilip et al. [19] used 5 different machine learning algorithms, including the linear support vector machine, Random Forest, C 5.0 Decision tree, k-nearest neighbour, Tree classifier, and Naive Bayes classifier. This article discusses the factors that influence employee attrition in every firm. It accurately forecasts workers who are likely to leave a company based on their job specifics and work surroundings. Random Forest dominates all other classifiers in terms of performance standards. This paper presented by Oanh et al. [20] predicts the chance of applicants intending to quit or remain in the organization after training sessions. This work was done to interpret the primary factors influencing applicant judgment and then to build a prediction model to predict whether an applicant will search for a job or stay with the company based on the candidate's existing qualifications, location, and experience. They did a lot of research with single classifiers (SVM, KNN, Logistics Regression, MLP Classifier, and Decision Tree) and ensemble classifiers (Soft Voting Classifier, Hard Voting Classifier, Random Forest, XGBoost, and LGBM classifier). Experimental outcomes on data proved that ensemble classifiers performed much better than single classifiers in general. The best classifier was LGBM, which produced up to an 80% F1 score. SMOTE was also utilized to deal with data that was not balanced. The SMOTE approach enhanced performance somewhat across all assessment measures. Additionally, they assessed each class's performance using SMOTE and discovered that predicting class 1-(person leaving the organization) is more difficult than predicting class 0-(person not leaving the company). In better detail, it eventually increased the F1 score by 61.44% and 81% for classes 1 and 0, respectively.

TABLE III
SUMMARY TABLE OF RELATED WORK FOR DATASET-3

Authors & Year	**Methodologies Used**
Tanya (2018)	SVM, Linear Regression, SMOTE
Ana et al (2018)	Back Propagation, Decision Tree, Self-organizing maps, Cross-validation
Francesca et al (2020)	Random Forest, KNN, SVM, Gaussian Naïve Bayes, Multivariate Bernoulli model
Dilip et al(2017)	SVM, Random Forest, C5.0 Decision Tree, KNN, Naïve Bayes
Oanh et al	SVM, KNN, Logistics Regression, MLP, Decision Tree, Soft Voting, Hard Voting, Random Forest, XGBoost, LGBM, SMOTE

III. METHODOLOGY

The KDD technique was used for these datasets to extract relevant information from the complete data set. Knowledge Discovery from Databases (KDD) is a term that refers to the exploration of data knowledge and underlines the high degree of complexity of the specific data mining technique. It is the technique of extracting information from databases which has previously been hidden. This approach has five stages: collecting data, preparing and transforming it to remove nulls and other incorrect values, and finally data mining and result. The final step is the application of data mining models, followed by an evaluation of the findings. The five steps of the KDD approach, as seen in Figure 1 [21], are described in further detail below.

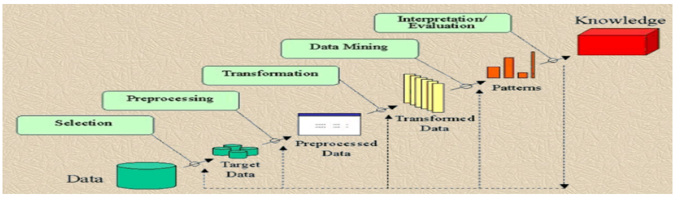


Fig. 1. Flow of KDD Methodolgy

A. Selection

The initial stage of KDD is the selection of acceptable data, which is then used to run data mining methods on it. It is critical at this stage to understand and create the relevant knowledge required for the project plan. This will therefore aid in the right selection of data since the fundamental understanding of the domain will make it easier to grasp the dataset columns and ultimately finalize the desired dataset. The desired dataset is chosen from among the data sources that are accessible and then utilized in the following processes. For all three datasets, I used Kaggle, which is the largest data data science forum, enabling users to create models online data science platform, engage with other researchers and machine learning specialists, and enter data science tournaments.

1) Dataset 1: Default of Credit Card Clients : I obtained this dataset using Kaggle, which was initially scraped from the UCI Machine Learning Repository. On Kaggle, the data was presented in CSV format with the filename (UCI_Credit_Card.csv). It covers the period from April to September 2005. The dataset has 25 columns and is based on 30,000 Taiwanese credit card users. For this dataset I have chosen the dependent variable as default.payment.next.month where (0 means 'No') and (1 means 'Yes') and independent factors as id, limit_bal, sex, education, marriage, age, pay_0,pay_2,pay_3, pay_4, pay_5, pay_6, bill_amt1, bill_amt2, bill_amt3,bill_amt4, bill_amt5, bill_amt6,pay_amt1,pay_amt2, pay_amt3, pay_amt4,pay_amt5 and pay_amt6. Based on an individual's account data and prior bill repayments, we will calculate the possibility that a future credit card default will be predicted in advance, and we will provide our findings.

2) Dataset 2: HR Analytics : Employee promotion data: I have fetched this dataset from Kaggle website. On Kaggle the data it divided into 2 csv files (train and test). For this project I have chosen the (train.csv) file. Later I have renamed the csv file to (employee_promotion.csv). This dataset contains employee characteristics, and it may be used to identify the elements that influence an employee's promotion. The characteristics of the employees are included in the dataset. It has 54809 rows and 13 columns. It enables for a significant volume of data for machine learning training and testing. Independent factors: employee_id; department; region; education; gender; recruiting_channel; no_of_trainings; age; previous year rating; length_of_service; awards_won; avg_training_score. The dependent variable for this dataset is is_promoted where '0' indicates not promoted and '1' indicates is promoted. We will try to figure out what qualities help an employee to get a promotion based on factors that are not related to each other.

3) Dataset 3: HR Analytics: Job change of Data Scientist: This dataset was obtained from the Kaggle website as well. On Kaggle, the data was separated into three csv files (aug_test, aug_train, and sample_submission). I've decided to use the (aug_train.csv) file for this particular project.I have renamed the csv file to (job_change.csv). This dataset contains employee characteristics, and based on the characteristics, it may be used to identify the factors that cause an employee to quit or remain with the organization. The dataset has 19159 rows and 14 columns, and it is structured as follows: enrollee_id, city, city_development_index, gender, relevant_experience, enrolled_university, education_level, major_discipline, experience, company_size, company_type, last_new_job and training_hours were all independent factors. The dependent variables target where '1' means looking for a new job and '0' means not looking for a job. Based on independent variables, we will attempt to determine what criteria assist to determine whether or not an employee is seeking a new job or not.

B. Preprocessing and Transformation

The next step in the KDD process to get the greatest results from machine learning models is preprocessing the data and it's critical to clean up and preprocess data before applying the models so that the data can be better understood by the machines. To develop a model, data preprocessing is an extremely important stage. Many factors go into data preprocessing, such as determining if the data is categorical or numeric, missing values and whether there are any outliers. Once these factors are determined, the data will be cleaned and transformed properly.

1) Dataset 1: Default of Credit Card Clients : The dataset was tested for Null values before preprocessing and modification. A column 'ID' was removed after that since it included nothing but a typical serial number of observations, which was considered not needed. After that, we generated a heatmap, as shown in Figure 2 , and concluded that no factors are significantly connected with the target attribute (default). Each of the 'PAY_' variables is highly correlated with the others, although only a moderate positive correlation exists between them and the target variable (default). There is a significant positive link between all of the 'BILL_AMT' variables in the data set. Furthermore, the variables 'LIMIT_BAL' and 'BILL_AMT' have a strong positive association with one another. Using sklearn.model selection train test split, we have divided the data into train and test sets of 75% and 25%, respectively, to alter the data for transformation. It was necessary to construct the variables x_train, x_test, y_train, and y_ test. Also, we transformed the data with the help of sklearn.preprocessing MinMaxScaler.

2) Dataset 2: HR Analytics : Employee promotion data: As this stage is all about the preprocessing and transformation of the data, the dataset was verified for null values and it was discovered that just two columns, 'education' and 'previous_ year_rating', had missing values. A total of 2409 missing data were found in the category of 'education,' while 4124 missing values were found in the category of 'previous_year_rating,' representing 4.395% and 7.52% missing values, respectively.

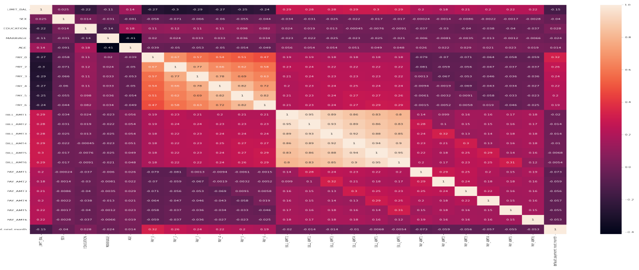


Fig. 2. Heatmap

But, since the dataset had 54809 entries and was big enough to support the construction of a model, it was chosen to eliminate null values. After removing all of the null values from the dataset, the dataset still had 48660 rows and 13 columns. Afterwards, it was examined to see how many categorical features the data included. We came to know there were five categorical features in total, with the columns ’department, ’education’, ’gender’, ’recruitment_channel’, and ’region’. Accordingly, for feature engineering purposes, the education column was handled using ordinal encoding, which means that a ranking was provided depending on the education with the low being ”Below Secondary,” followed by ”Bachelor’s,” and the top is ”Master’s above.” Certain columns, such as ’department’, ’region’, and ’recruitment_channel’, cannot be ranked; thus, dummy variables were generated and a dummy trap was implemented to prevent the computer from being confused. As soon as the data was converted to a numeric format, it was attempted to determine which columns would be useful in building the model, and it was discovered that the column identifying an ’employee id’ was not useful in determining whether or not that individual was eligible for promotion, and thus that column was removed from the data. There were outliers detected in several variables such as the no_of trainings, age and length_of_service that were not damaging to the model development, therefore feature selection was performed. Figure 3 depicts outliers, whereas Figure 4 depicts the best attributes for model construction by using the ANOVA test. Before moving on to the next data mining phase, it was determined if the dependent variable ’is_promoted’ was balanced or not. It was discovered that the variable was not balanced, thus the SMOTE resampling approach has been employed to balance the data. Figure 5 and 6 shows the results of sampling before and after of the dependent variable. Finally, I separated the data into testing and training, with 80% training and 20% testing.

3) Dataset 3: HR Analytics: Job change of Data Scientist: Initially, the dataset was verified for null values and it was discovered that this dataset had a lot of Null values. There were 8 columns out of 14 columns having null values. The columns ‘gender’ had 4508, ’enrolled_university’ had 386, ’education_level’ had 460, ’major_discipline’ had 2813, ’experience’ had 65, ’company_size’ had 5938, ’company_type’ had 6140 and ’last_new_job’ had 423 missing values. Fig.7 shows that the block with dark shades exhibited a correlation between columns with missing data, as shown by the heatmap. Dropping missing value was not a suitable choice since there is significant association between the data points. Additionally, if null values were omitted, the dataset’s structure would have been severely reduced, which would have made it more difficult to develop a model. As a result, data imputation was performed statistically using the mode of each column. Columns with the lowest percentage of null values were removed, for example, the ’experience’ column had 0.34 percent missing data, hence null values were removed from that column. The columns ”enrollee_id” and ”city” were deleted from the dataset since they were not helpful

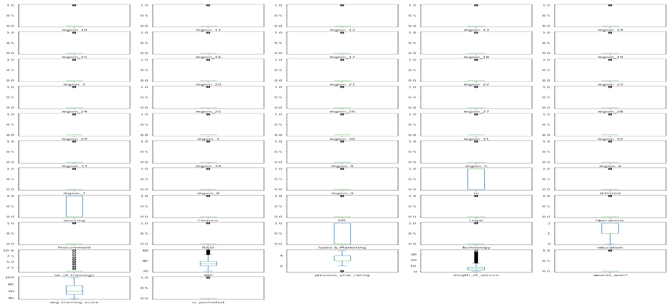


Fig. 3. Outliers in the Data

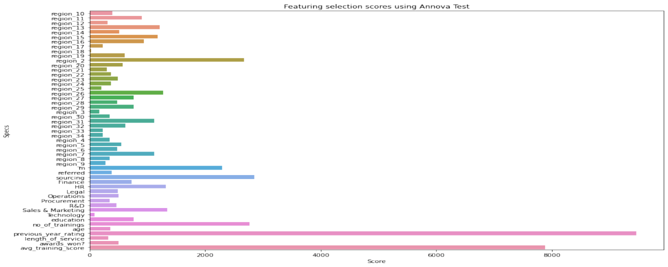


Fig. 4. Best Features extracted by ANOVA Test used in Final Model

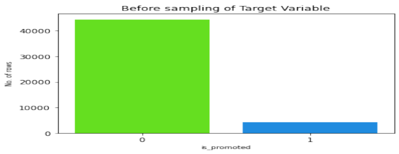


Fig. 5. Before Sampling of Data

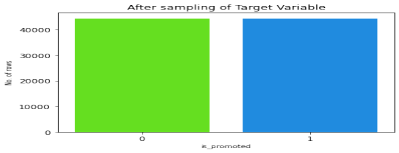


Fig. 6. After Sampling of Data

in determining whether an employee will remain or quit the organization. To aid in the development and analysis of the model, ordinal encoding was performed on columns such as ’relevant_experience’, ’enrolled_university’, ’education_level’, and ’company_size’, among others. By ranking these columns, the model may be built and analysed more effectively. Because ranking cannot be provided for columns such as ”gender,” ”major_discipline,” and ”company_type,” dummy variables were developed, and a dummy trap was also set up to prevent this from happening. OneHotEncoding was also applied to the variables listed above. When everything was done with the preprocessing and transformation, there were 19093 rows and 22 columns in the dataset. When it came to outliers, there was just one column, ’training_hours,’ and they were not considered to be harmful to the model’s construction, thus they were not deleted. Outliers in this dataset is shown in Fig. 8. A assessment was made as to whether or not the dependent variable ’target’ was balanced before going on to the next data mining step. As a result of discovering that the variable was not balanced, the SMOTE resampling technique was used to ensure that the data was properly balanced. In Figures 9 and 10, you can see the outcomes of sampling before and after the dependent variable is included. Finally, I separated the data into testing and training, with 80% training and 20% testing.

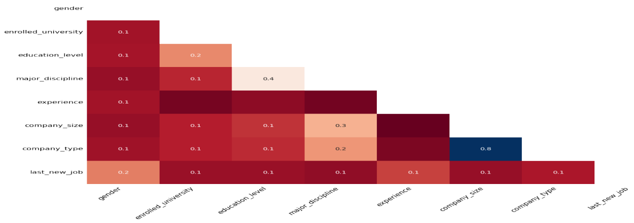


Fig. 7. Correlation between Columns by Heatmap

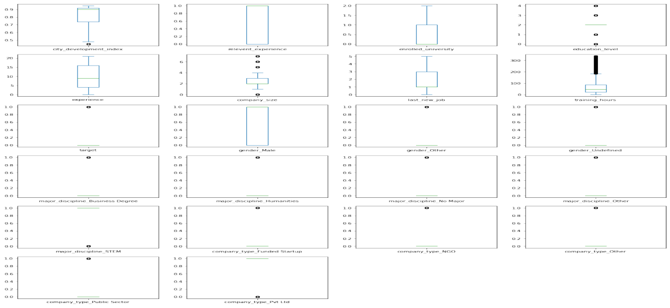


Fig. 8. Outliers for Dataset 3

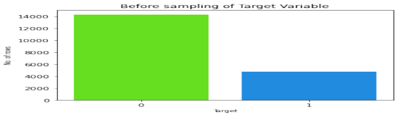


Fig. 9. Before Sampling of Data

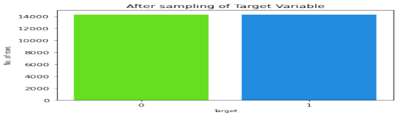


Fig. 10. After Sampling of Data

C. Data Mining

The data mining step of the KDD approach is the fourth phase. This stage comprises of selecting the appropriate data mining algorithms for the data collection that has been selected. This data mining technique searches for patterns that may be present in the data set being mined. It would be necessary to find the appropriate model parameters in order to assure optimal algorithm execution. All three data sets in this research are classified using machine learning methods, and only classification machine learning techniques are employed on them since they have characteristics that allow classification to be performed. I utilized a total of five machine learning classification approaches and compared them to see which produced the best results.

1) Dataset 1: Default of Credit Card Clients : The K-Nearest Neighbors (KNN) technique, the Random Forest Classifier algorithm, and the Logistics Regression algorithm were all employed in this dataset to classify the data using machine learning. I’ve also worked with hyperparameter tuning on algorithms such as the Random Forest classifier and the Logistics Regression method. A rationale for using KNN is as follows: KNN is a slow learning method that is based on supervised machine learning. This means that when new data is received, it is classified into a group that is quite similar to the previously stored information. So our dataset can simply predict whether or not the consumer would default. A rationale for using the Random Forest Classifier: It is the most versatile and simple approach for classification and regression. A forest is made up of trees that grow together. The number of trees in a forest is thought to increase the durability of the forest. However, since there are a large number of categorical variables in this dataset, it aids in the production of the best outcomes when trees are formed. When compared to a decision tree, this model is more difficult to interpret and produces predictions more slowly because it includes several choices. However, it produces the greatest outcomes than the decision tree. I have chosen this as it will give the best possible performance and will be able to correctly predict the result. A rationale for using Logistics Regression: It is an excellent option for both classification and regression. Like linear regression, logistic regression has an equation. The dependent variable is splitting, thus there are only two ways to classify the data. The sole purpose for choosing this algorithm was that I needed my dependent variable to predict the default of credit cards clients.

2) Dataset 2: HR Analytics : Employee promotion data: The Logistic Regression and Gaussian Naive Bayes models were used to create the model for this dataset. Using this dataset, I was able to compare the performance of logistics regression in various types of data, which was a valuable learning experience for me. The following are the reasons for using Naive Bayes: - Since the dataset had a great number of categorical variables, and because Naive Bayes performs well with categorical variables, we used it. When independence is satisfied, Naive Bayes outperforms other models like logistic regression and takes less training data. Generally, the existence of one feature in a class is assumed to be independent of all other features in that class.

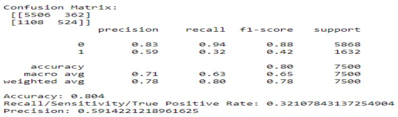
3) Dataset 3: HR Analytics: Job change of Data Scientist: Once the dataset was cleaned created, I used a Random Forest Classifier and a Decision Tree to classify the data. The Random Forest Classifier has also been employed in this dataset to test its performance in a variety of other datasets. The following are the reasons for selecting Decision Tree: It is possible to utilize a decision tree to solve the problems of classification and regression as well. Considering that this dataset has a significant number of categorical variables, a decision tree is a hierarchical tree structure that seems similar to a flowchart, but where each internal node represents a function. A decision tree is a white box in machine learning since the underlying decision-making logic is shared, while neural networks are a black box. The decision tree works best with three data layers. Its capability of using a variety of feature subsets and decision criteria at various phases of the categorization process.

IV. Evaluation

At this step, the results and outputs of data mining algorithms used in the preceding section are evaluated to determine whether or not acceptable and usable patterns have been discovered. As part of the evaluation process, a number of visualizations and prediction evaluations will be created. The evaluations of each model developed during the data mining phase of KDD are listed below. The models are evaluated on the following parameters listed below:

- Confusion Matrix: It assists to determine if a model correctly predicts and classifies an item when it is provided with that item to predict. The following terms make up the confusion metrics:
  True positive(TP): Should forecast and be positive.
  False positives (FP): Results that should be positive but aren't.
  True negatives (TN):Predictions that should be wrong and are wrong.
  False negatives (FN): Predictions that should be negative but turn out to be positive.
- Accuracy: It is a metric that summarizes how well a model works across all sets of data. It is based on the link between accurate forecasts and total predictions.
- Recall: It aids in determining how many instances are positive out of the overall number of genuine positive events. According to this formula, the number of positive cases divided by the number of negative cases is TP/(TP+FN).
- Precision: It aids in determining how many positive cases there are out of the total projected positive occurrences (prediction accuracy). It answers the problem of how much the model is correct. The formula is TP/(TP + FP).
- F1 Score: It is the average of precision and recall . More the F1 score is better.

1) Dataset 1: Default of Credit Card Clients : As the dataset was already balanced so we have not used any resampling method. We have used K-Nearest Neighbor(KNN, Random Forest Classifier and Logistics regression on this dataset. After running the algorithms on the data we saw that KNN got an accuracy of 80.40% on the transformed dataset. A precision of 59.14%, recall score of 32.10% and AUC score of 0.72 demonstrates that large values on Y-Axis lower false negatives and higher true positives. The figure 11 shows the model details for KNN. For the model Random forest classifier, we received an accuracy of 81.08%. A precision of 68.08%, a recall score of 24.57% and an AUC value of 0.77. The figure 12 and 13 shows the model details for Random Forest Classifier. For the model Logistics regression, we got an accuracy of 81.46%. A precision score of 22.67%, a recall score of 33.27% and an AUC value of 0.72. The figure 14 shows the model details for Logistics Regression. In general, logistics regression proved successful in classifying if an individual is going to default his credit card payment in advance. As a result, Logistics Regression has been selected as the most appropriate model for answering our research question.



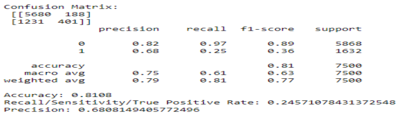
```
Confusion Matrix:
 [[5506  362]
 [1108  524]]
              precision    recall  f1-score   support

           0       0.83      0.94      0.88      5868
           1       0.59      0.32      0.42      1632

    accuracy                           0.80      7500
   macro avg       0.71      0.63      0.65      7500
weighted avg       0.78      0.80      0.78      7500

Accuracy: 0.804
Recall/Sensitivity/True Positive Rate: 0.32107843137254904
Precision: 0.5914221218961625
```

Fig. 11. Confusion Matrix for KNN for Dataset 1



```
Confusion Matrix:
 [[5680  188]
 [1231  401]]
              precision    recall  f1-score   support

           0       0.82      0.97      0.89      5868
           1       0.68      0.25      0.36      1632

    accuracy                           0.81      7500
   macro avg       0.75      0.61      0.63      7500
weighted avg       0.79      0.81      0.77      7500

Accuracy: 0.8108
Recall/Sensitivity/True Positive Rate: 0.24571078431372548
Precision: 0.6808149405772496
```

Fig. 12. Confusion Matrix for Random Forest Classifier for Dataset 1

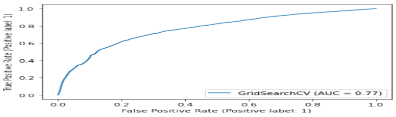
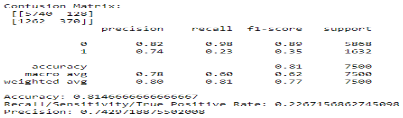


Fig. 13. AUC Of Random Forest for Dataset 1



```
Confusion Matrix:
 [[5740  128]
 [1262  370]]
              precision    recall  f1-score   support

           0       0.82      0.98      0.89      5868
           1       0.74      0.23      0.35      1632

    accuracy                           0.81      7500
   macro avg       0.78      0.60      0.62      7500
weighted avg       0.80      0.81      0.77      7500

Accuracy: 0.8146666666666667
Recall/Sensitivity/True Positive Rate: 0.2267156862745098
Precision: 0.7429718875502008
```

Fig. 14. Confusion Matrix for Logistics Regression for Dataset 1

2) Dataset 2: HR Analytics : Employee promotion data: We utilized the SMOTE method to resample the data and make it seem more balanced since the data was severely unequal. Since an unbalanced dataset will not provide the intended outcome. In our tests, we obtained an accuracy of 78.01% for Logistic Regression and 67.47% for Gaussian Naive Bayes machine learning models on the dataset after running them both. With an accuracy score of 77.07%, a recall score of 80.50%, and an F1 score of 78.75%, the Logistics regression model outperformed as shown in the Figure 15. When naive Bayes was used to train the model on this dataset, the results were unsatisfactory, with an accuracy score of 61.85%, and an F1 score of 74.38 % for the model.To check the model's performance. Confusion measures were used to distinguish between true positive and predictive events. As seen in Fig. 16, the Naive Bayes model did not predict false negatives. Instead of predicting 0 occurrences, it guessed 1. The recall result of 93.27%, on the other hand, was favourable for Naive Bayes. The model was able to accurately distinguish between real positives and true negatives. However, when it came to false negatives, the model failed to forecast them correctly once again. The final comparison between the models is shown in the Figure 17.In general, logistic regression proved successful in classifying prime individuals who should be promoted based on the data provided. As a result, logistic regression has been selected as the most appropriate model for answering our research question.

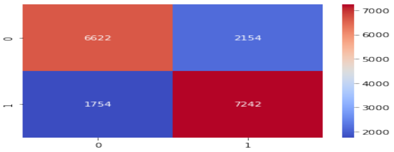


Fig. 15. Confusion Matrix for Logistics Regression for Dataset 2

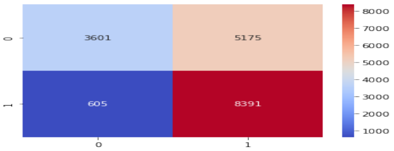
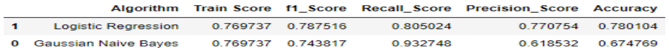


Fig. 16. Confusion Matrix for Naive Bayes for Dataset 2



	Algorithm	Train Score	f1_Score	Recall_Score	Precision_Score	Accuracy
1	Logistic Regression	0.769737	0.787516	0.805024	0.770754	0.780104
0	Gaussian Naive Bayes	0.769737	0.743817	0.932748	0.618532	0.674769

Fig. 17. Comparison of Models in Dataset 2

3) Dataset 3: HR Analytics: Job change of Data Scientist: Following the construction of our model using Decision Trees and Random Forest Classifiers, evaluation techniques such as the confusion matrix, Precision Score, F1 score, recall, ROC AUC curves are used to assess the performance of our machine learning model. After training the dataset using decision trees and random forests, the following is the result: Fig. 18 and 19, respectively. In this, Random Forest has a high accuracy of 80.91%, whereas the decision tree has an accuracy of just 75.67%, as can be observed. Random forest, on the other hand, performed well in all areas. The F1 score of 81.09%, the recall score of 82.22%, and the precision score of 80% are all excellent. When the model was told to forecast to zero, it predicted all zeros, and when the model was told to predict one, it predicted one. The model correctly predicted 1, 804 times when it was told to predict 0, while it incorrectly projected 0, 592 times when it was told to predict 1. Additionally, when the model was trained using a random forest classifier, it accurately predicted the real positive. Using confusion metrics, you can see whether or not the model has done well enough and is generating the desired outcomes. Also after plotting the Area Under Curve (AUC) shown in the figure 20 we got a value of 0.88 for random forest and a value of 0.83 for decision tree. The final comparison between the models is shown in the Figure 21. Consequently, it can be said that the Random Forest classifier model is capable of classifying the number of employees who want to remain with the organization as well as those who are seeking a new position. The Random Forest Classifier has been determined to be the best-suited model for solving our research question in this case.

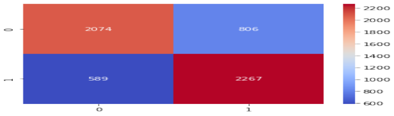


Fig. 18. Confusion Matrix for Decision Tree for Dataset 3

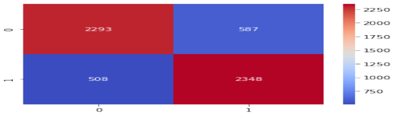


Fig. 19. Confusion Matrix for Random Forest for Dataset 3

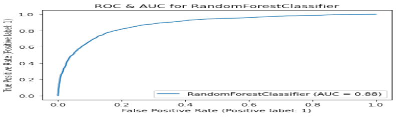
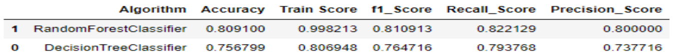


Fig. 20. AUC Of Random Forest for Dataset 3



	Algorithm	Accuracy	Train Score	f1_Score	Recall_Score	Precision_Score
1	RandomForestClassifier	0.809100	0.998213	0.810913	0.822129	0.800000
0	DecisionTreeClassifier	0.756799	0.806948	0.764716	0.793768	0.737716

Fig. 21. Comparison of Models in Dataset 3

V. Conclusions and Future Work

The KDD technique to data mining was used in this research to conduct a comparative examination of the outputs of five machine learning models. All five machine learning models have been successfully tested on three different data sets, with satisfying results in terms of developing dataset-specific models. We hope that this work will help us learn more about what customers want and what employees do. To estimate which consumers would default on their credit card payments in the first dataset, three machine learning models are utilized, namely the KNN, the Random Forest Classifier, and logistic regression. The accuracy of logistic regression was found to be greater at 81.46% and the AUC was found to be 0.72 when compared to KNN and Random Forest, which also had good accuracy of 80.40% and 81.08%, respectively. Due to the fact that the basic idea of detecting which customers are likely to fail on their payments remains the same, this model may be utilized among both banks and other financial issuers in addition to loan lenders. This has the potential to save banks money by identifying likely clients who will fail in a timely manner. In the future, this model might be modified to deal with credit card scams that result in credit card defaults, as well as other scenarios. This model may also be used to forecast loan defaults, which is another use. Logistic regression and Gaussian Nave Bayes are employed in the second dataset to determine whether or not an employee deserves a promotion or not. Logistics regression again showed to be the superior model, with an accuracy of 78.01%, compared to Naive Bayes 61.85% . In the workplace, this model may be used to determine whether or not an individual is worthy of a promotion. If the correct individual is given a promotion, this may significantly aid a company's growth. Random Forest Classifier and Decision Tree are employed in the third dataset to predict whether or not an employee is seeking for a new job. Random Forest had a greatest accuracy of 80.91% with an AUC of 0.88, whereas Decision had an AUC of 0.83. This model may be used in businesses to determine if a worker is actively seeking a new position or is content to remain with his or her current employer. More advanced machine learning techniques may be used to these models, which would ultimately provide a better outcome.

References

[1] "Wilton attorney pleads guilty to bank, credit fraud," The Hour, Aug. 07, 2008. https://www.thehour.com/wilton/article/Wilton-attorney-pleads-guilty-to-bank-credit-8266456.php (accessed Apr. 24, 2022).

[2] A. Hameed and A. Waheed, "Employee Development and Its Affect on Employee Performance A Conceptual Framework," vol. 2, no. 13, p. 6.

[3] R. Liu, "Machine Learning Approaches to Predict Default of Credit Card Clients," Modern Economy, vol. 09, no. 11, pp. 1828–1838, 2018, doi: 10.4236/me.2018.911115.

[4] A. Husejinovic, D. Kečo, and Z. Mašetić, "Application of Machine Learning Algorithms in Credit Card Default Payment Prediction," International Journal of Scientific Research, vol. 7, pp. 425–426, Oct. 2018, doi: 10.15373/22778179#husejinovic.

[5] Y. Li, Y. Li, and Y. Li, "What factors are influencing credit card customer's default behavior in China? A study based on survival analysis," Physica A: Statistical Mechanics and its Applications, vol. 526, p. 120861, Jul. 2019, doi: 10.1016/j.physa.2019.04.097.

[6] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi, and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," in 2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA), Oct. 2018, pp. 1–4. doi: 10.1109/ICACCAF.2018.8776802.

[7] Y. Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction," in 2020 International Conference on Computing and Data Science (CDS), Aug. 2020, pp. 212–218. doi: 10.1109/CDS49703.2020.00050.

[8] J. N. Houle, J. M. Collins, and M. D. Schmeiser, "Flu and Finances: Influenza Outbreaks and Loan Defaults in US Cities, 2004–2012," Am J Public Health, vol. 105, no. 9, pp. e75–e80, Sep. 2015, doi: 10.2105/AJPH.2015.302671.

[9] M. Leow and J. Crook, "A new Mixture model for the estimation of credit card Exposure at Default," European Journal of Operational Research, vol. 249, no. 2, pp. 487–497, 2016, doi: https://doi.org/10.1016/j.ejor.2015.10.001.

[10] O. Netzer, A. Lemaire, and M. Herzenstein, "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," Journal of Marketing Research, vol. 56, no. 6, pp. 960–980, 2019, doi: 10.1177/0022243719852959.

[11] A. Tadesse, "The Effect of Employee Promotion Practice on Job Satisfaction: The Case of Dashen Bank S.C.," masters, Addis Ababa University, 2017. Accessed: Apr. 22, 2022. [Online]. Available: http://thesisbank.jhia.ac.ke/5094/.

[12] Y. Long, J. Liu, M. Fang, T. Wang, and W. Jiang, "Prediction of Employee Promotion Based on Personal Basic Features and Post Features," in Proceedings of the International Conference on Data Processing and Applications - ICDPA 2018, Guangdong, China, 2018, pp. 5–10. doi: 10.1145/3224207.3224210.

[13] L. Liu, S. Akkineni, P. Story, and C. Davis, "Using HR Analytics to Support Managerial Decisions: A Case Study," in Proceedings of the 2020 ACM Southeast Conference, New York, NY, USA, Apr. 2020, pp. 168–175. doi: 10.1145/3374135.3385281.

[14] S. Shet and B. Nair, "Quality of hire: expanding the multi-level fit employee selection using machine learning," International Journal of Organizational Analysis, vol. ahead-of-print, no. ahead-of-print, Jan. 2022, doi: 10.1108/IJOA-06-2021-2843.

[15] A.Sarker,S.M.Shamim,P.D.M.S.Zama, and M.M.Rahman, "Employee's Performance Analysis and Prediction Using K-means Clustering & Decision Tree Algorithm," Global Journal of Computer Science and Technology, Feb. 2018, Accessed: Apr. 22, 2022. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/1660.

[16] T. Attri, "Why an Employee Leaves: Predicting using Data Mining Techniques," masters, Dublin, National College of Ireland, 2018. Accessed: Apr. 22, 2022. [Online]. Available: http://norma.ncirl.ie/3434/.

[17] A. C. C. de Jesus, M. E. G. D. Júnior, and W. C. Brandão, "Exploiting linkedin to predict employee resignation likelihood," in Proceedings of the 33rd Annual ACM Symposium on Applied Computing, Pau France, Apr. 2018, pp. 1764–1771. doi: 10.1145/3167132.3167320.

[18] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," Computers, vol. 9, no. 4, Art. no. 4, Dec. 2020, doi: 10.3390/computers9040086.

[19] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in 2017 International Conference on Inventive Computing and Informatics (ICICI), Nov. 2017, pp. 1016–1020.doi: 10.1109/ICICI.2017.8365293.

[20] O. T. Tran and L. P. Nguyen, "Trainee Churn Prediction using Machine Learning: A Case Study of Data Scientist Job," p. 9, vol.3026.

[21] "KDD Process/Overview. "http://www2.cs.uregina.ca/ d̃bd/cs831/notes/kdd/1_kdd.html.