**DR. BABASAHEB AMBEDKAR MARATHWADA UNIVERSITY, AURANGABAD**

**DEPARTMENT OF STATISTICS**

** 2022-2023**

**PROJECT REPORT ON**

**“Statistical Analysis of Employees Attrition**

**and Retention”**

**SUBMITTED BY**

**MISS. BALLAL GAYATRI VIVEK**

**MR. GORE MAYURESH HANUMANT**

**MISS. JADHAV SAPNA VASANTRAO**

**MISS. PRITI UTTAM SASANE**

**UNDER THE GUIDANCE OF**

**DR. MRS. CHHAYA D. SONAR,**

**ASSOCIATE PROFESSOR**

**DEPARTMENT OF STATISTICS, DR. B. A. M. U. AURANGABAD**

**DR. BABASAHEB AMBEDKAR MARATHWADA UNIVERSITY, AURANGABAD**

**DEPARTMENT OF STATISTICS**

**2022-2023**

****

**CERTIFICATE**

This is certify that

**Ballal Gayatri Vivek**

**Gore Mayuresh Hanumant**

**Jadhav Sapna Vasantrao**

**Priti Uttam Sasane**

**Have successfully completed the project entitled,**

**“Statistical Analysis of Employees Attrition**

**and Retention”**

**Submitted to the Department of Statistics, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, towards the partial fulfilment of the Master degree in science (Statistics) during the academic year 2022 – 2023**

**GUIDE EXAMINER HEAD OF DEPARTMENT**

**Asst. Prof. Dr.Mrs. Chhaya D. Sonar Prof. Dr. S. V.Kawale**

**Department of Statistics, Department of Statistics,**

**Dr. B. A. M. U., Aurangabad Dr. B. A. M. U. Aurangabad**

**ACKNOWLEDGEMENT**

We are wish to express my sincere gratitude to our project guide Dr. Mrs. Chhaya D. Sonar ma’am, Associate Professor, Dept. of Statistics, Dr. Babasaheb Ambedkar Marathwada University. Aurangabad. Whose guidance, undeterred moral support enabled us to complete our project. Her meticulous thinking. scientific judgment, critical analysis and able guidance at every stage have made this project to seep Its due results.

We are thankful to Prof. Dr. S.V. Kawale sir, Head of Department, Dept. of Statistics, Dr. B.A.M.U., Aurangabad, for helping us, and without their helping completion of this project could not have been possible. We are also thankful to Prof. Dr. Omprakash S. Jadhav sir, Prof. Dr. Ashok Y. Tayade sir who have fully co-operate and inspire us.

We are also thankful to Research Students of the department to help us in statistical analysis of our data and also all non-teaching staff, Department of Statistics, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

Lastly, we expressed our gratitude to all of our friends and contributors to our effort.

**Contents**

**Chapter 1: Introduction**

* 1. **About employees’ attrition and retention**
  2. **Importance of employees’ retention**
  3. **Objectives of the study**
  4. **Need of the study**
  5. **Review of Literature**

**Chapter 2: Research Methodology**

**2.1 Concept and Definitions**

**2.2 Method of collection of data**

**2.3 Classification Algorithms**

**2.4 Model Evaluation Metrics and Interpretation**

**Approach**

**Chapter 3: Statistical Analysis and Discussion**

**3.1 Data Visualization**

**3.2 Employees attrition dashboard**

**3.3 Data Analysis Using Classifiers**

**3.4 Testing used for analysis**

**3.5 Data analysis using Python and R**

**Chapter 4: Summary and Conclusions**

**Chapter 5: References**

**CHAPTER 1**

**INTRODUCTION**

**1.INTRODUCTION**

**1.1 About Employees Attrition and Retention**

"Attrition is said to be the gradual reduction in the number of employees through retirement, resignation or death. It can also be said as Employee Turnover or Employee Defection". Attrition means that number of employees leaves an organization from a specific group over a period of time. Retention means employees that stay with the organization during that time frame. For each employee, an organization has to invest an abundance of time and money for their training as per the organization's requirement. When an employee leaves the organization, the company is not only losing its valuable employees, but the company also loses on the amount that it has spent to recruit and select those employees and to train them for their respective jobs. On the other hand, the organization needs to invest more and more in recruitment, training, and development of new staff to fill up their vacant positions. Due to these reasons, every organization wants to control the attrition rate and retain its employees through more satisfactory company policies and work environments.

The present research work would be helpful for most companies to know about their employee's satisfaction levels and get some useful information that would aid in controlling the attrition rate. Employee attrition or voluntary turnover presents a key issue for organizations as it affects not only their productivity and work sustainability but also their long-term growth strategies. On this path, employee retention is a major challenge Building & maintaining a suitable environment is the key that contributes to stable & collaborative employees. Employees attrition means that productive. employees decide to leave the organization due to different reasons such as work pressure, unsuitable environment or not satisfying Salary. Employees attrition affects the organization's productivity because it loses a productive employer as well as other resources such as HR staff effort in recreating new employers. Recruiting new employees requires training, development & integrating them into the new environment. An employee would choose to join or depart organization depending on many causes i.e. work environment, work place, gender equity, pay equity & many other. The rest of employee. may think about personal reasons for instance relocation due to family, maternity, health issues with the co-workers in a team. The employees leaves the organization faster than they are hired. The high attrition rate shows that the employees are frequently leaving "The results of high attrition rate are loss of organizational benefits. In order to keep the organization in process, the attrition rate must be controlled.

**1.2 Importance Of Employees Retention**

The process of employee retention will benefit an organization in the following ways:

**The cost of turnover:**

The cost of employee turnover adds hundreds of thousands of money to a company's expenses. While it is difficult to fully calculate the cost of turnover (including hiring costs, training costs and productivity loss), industry experts often quote 25% of the average employee salary as a conservative estimate.

**Loss of company's knowledge:**

When an employee leaves, he takes with him valuable knowledge about the company, customers, current projects and past history (sometimes to competitors). Often much time and money has been spent on the employee in expectation of future return. When the employee leaves, the investment is not realized.

**Interruption of customer service:**

Customers and clients do business with a company in part because of the people. Relationships are developed that encourage continued sponsorship of the business. When an employee leaves, the relationships that employee built for the company are served, which could lead to potential customer loss.

**Turnover leads to more turnovers:**

When an employee terminates, the effect is felt throughout the organization. The unspoken negativity often intensifies for the remaining staff.

**Goodwill of the company:**

The goodwill of a company is maintained when the attrition rates are low. Higher retention rates motivate potential employees to join the organization.

**Method of collection of data:**

We use the secondary data set taken from Kaggle website for this project. The data set contains 1470 records and 32 features. The first thing to do is to deal with the missing and duplicates value. We do not found any missing value and duplicate value in our data. There are 32 features in our data namely age, attrition, department, daily rate, distance from home, education, education field, environment satisfaction, employee count, employee number, gender, hourly rate, job involvement, job role , job satisfaction, marital status, monthly income, monthly rate, number of companies worked, overtime, percent salary hike, performance rating, relationship satisfaction, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, years since last promotion, Years With Current Manager.

Sr. No Variables Description

1. Age The age of the employee

2. Attrition 1: yes , 0:No

3. Business Travel Non travel, travel frequently, travel rarely

4. Daily Rate Numerical value-Salary level

5. Department Human Resources, R&D,

6. Distance from Home Numerical value-The distance from work to

home

7. Education Numerical value

8. Education Field HR, Life Sciences, Medical, Marketing

Other,Technical Degree

9. Environment Satisfaction Rated on the scale of 1 to 4,4 being lowest

and 1 being Highest

10. Employee Count Numerical value

11. Employee Number Numerical value-EMPLOYEE ID

12. Gender Gender of the Employee

13. Hourly Rate Numerical value-HOURLY SALARY

15. Job involvement Rated on the scale of 1 to 4,4 being lowest

and 1 being Highest

16. Job Level Rated on the scale of 1 to 4,4 being lowest

and 1 being Highest

17. Job role (1=HREP, 2=HR, 3=LABTECHNICIAN,

4=MANAGER)

18. Job satisfaction Rated on the scale of 1 to 4,1 being highest

and 4 being lowest

19. Marital itatus Single, Married or Divorced

20. Monthly income The monthly salary of the employee

21. Monthly rate Numerical value-MONTHLY RATE

22. Over 18 (1=YES,2=N0)

23. Percent Salary Hike Numerical value- Percentage Increase In

Salary

24. Performance Rating Numerical value- Performance Rating

25. Relations Numerical value- Relations

26. Stock Options Level Numerical value- Stock Options

27. Total Working Year Numerical value

28. Over Time Whether the employee does Overtime

work or not

29. Total Working Years Total work Experience in years

30. Work Life Balance Rated on the scale of 1 to 4,4 being lowest

and 1 being Highest

31. Years At Company Total experience at current company in

years

32. Years In Current Role Years working in current profile

33. Years Since Last Promotion Years since last promotion

34. Years with current Manager Years with current Manager

**1.3 Objectives of the Study:**

Purpose of this project is study the theoretical as well as practical aspects of Statistics in real world. Such type of study can give information about job

opportunity in various fields, knowledge of data collection and knowledge of The main objectives of this study are as follows:

1. To find the reason of attrition among employee’s.

2. To determine highly contributing factors which affects employee’s attrition.

3. To determine impact of demographic factors on attrition.

4.To find the association of various parameters with attrition.

5. To find the best fitted classifier suitable for attrition data.

**1.5 Need of the study:**

Dissatisfaction and for what reason employee’s change their job. Once the levels of employee’s attitude are identified, it would be possible for the management to take necessary action to reduce attrition level. Since they are considered as backbone of the company their progression will lead to the success of the company for the long run.

**1.6 Review of literature:**

1] Title-Explaining and predicting employees attrition : a machine learning approach(Springer)

Author : Praphula Kumar Jain, Madhavi Jain, Rajendra Pamula

In this paper data mining techniques are use for prediction and classification. They selected department having more than 1000 employees and binary classifications are evaluated using binary numeric features. Then evaluating classifiers using confusion matrix and result is accuracy of random forest is 99% and it is greater than decision tree and SVM so, random forest is significantly better

2]Title: From big data to deep support people analytics for employee attrition prediction(IEEE)

Author: Nesrine Ben, Yahia, Jihan Hlel, Ricardo colomo-palacios(senior member IEEE)

In this paper survey is conduct in different countries like Tunisia, Norway, France, United States, China, Italy, Pakistan, India, England and Germany. Survey is conduct through an online Questionnaire and 15000 samples are taken. Then classifiers like XGB, Random Forest, Decision Tree, Logistic Regression, SVM ,VC are used for classification. From the analysis they found that Voter classifier (VC) performs better than other models since its accuracy is greater than others i.e. 99%.

3]Title: Employees Attrition Prediction Using Deep Neural Networks (MDPI)

Authors: Salah Al Darraji ,Dhafer G. Honi , Francesca Fallucchi, Ayad I. Abdulsada, Romeo Giuliano, Husam A. Abdulmalik.

In this paper various machine learning models have been used in employees attrition such as decision trees, random forest, naïve bayes, logistic regression, SVM, DNN. They used deep learning model to classify employees attrition. From the analysis the found that accuracy of DNN is greater than all other classifiers.So DNN performs better than all other models.

4] Employee attrition using predictive techniques (Springer)

Authors: Devesh Kumar Srivatsava , Priyanka Nair

In this paper 10000 of employee data have been taken into account. Then various classifiers are use for prediction and classification. The artificial neural network gave 40% accuracy for the data. Analysis is done to predict the employee churn accurately. From model they get 60% accurate predictions and 40% inaccurate predictions.

5] Predicting employees attrition using machine learning approaches (MDPI)

Authors: Ali Raza, Kashif Munir, Mubarak Almutairi, faizan Younas, Mian Muhammad Sadiq Fareed.

In this paper split ratio of data is 85:15. Then classifiers auch as SVM, Logistic Regression, Decision Tree, Extra Tree Classifier are use for predicting employees attrition. K fold cross validation also used. From analysis they found that accuracy of ETC i.e., 93% is greater than all other classifiers so, ETC performs better than all other models.

6]Title: A study of employees attrition among executives at BGR energy systems ltd, Chennai (IRJC)

Authors: DR. R. AKILA

In this paper to study the opinion of executives an employee retention attributes at BGR sample of 109 cadre using descriptive research design under convenience study. The questionnaire forms the basic source of primary data, while secondary data was collected through books and through online journals, magazines, project report. The data collected was analysed using ANOVA, Rank correlation, Weighted Average Method, Regression, Chi square and Percentage Analysis. The analysis on the collected data revealed that the Employees have given highest weighting to comfortableness in working hours and are not satisfied with annual increments provided. Gender has significant relationship with work life balance. Correlation between job satisfaction and working condition leads to employee retention. The departments also have influence on it.

7)Title: Employees Retention: a review of literature

Authors: Krishna Prakash, Dr. Jawahar philims

In this paper total 30 research papers were analysed. A summary appeared in chronological order. This study summarized the various researcher’s new approaches for the retaining employees in organizations ability to attract and retain valuable employee is a top priority. From analysis we found that organization needs look into work environment ,leadership ,training, development as they are most important factors to retain the employees in an organization.

8] Title: An investigation into the effect of knowledge management on employee retention in the telecom sector (MDPI)

Authors: Nurul Mohammad Zayed, Friday Ogbu Edeh, Khann Mohammad Anwarul Islam, Vitalii Nitsenko, Tetiana dubovyk, hanna doroshuk.

In this paper data is collected from Nigeria related to knowledge and management of employee’s attrition. The analysis showed that there is a significant and positive impact of knowledge management processes on innovation in Jordanian consulting firms, as well as a significant and positive effect of codification and personalization approaches on innovation, while the social network approach has a significant negative impact with innovation.

**CHAPTER 2**

**RESEARCH METHODOLOGY**

**2.RESEARCH METHODOLOGY**

**2.1 Concept and definitions:**

**Mean**:

A mean score is an average, it is denoted by X̅. It is the sum of observations divided by the number of observations.

The standard deviation of a population is defined by the following *formula*

Mean=Sum of observations/No. of observations

X̅= (X1+ X₂+ X3+...+XN) / N = [ΣXi] / N

**Standard Deviation**

The standard deviation is positive square root of mean of squares of the deviations taken from arithmetic mean.

Standard deviation is defined by the followingformula:

Standard deviation**=**

Where σ is the population standard deviation, X̅, is the population mean, Xi is the ith  element from the population, and N is the number of elements in the population

**Level of significance:**

A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the significance level, and is often denoted by a.

**Sample:**

A sample is a subset of individuals from a larger population.

**Population:**

population is defined as a group of individuals of the same species living and interbreeding within a given area.

**Degrees of freedom:**

Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample.

**Regression:**

Relationship between independent and dependent variables.

**Classification:**

Classification in data mining is a common technique that separates data points into different classes.

**Chi-square test :**

**What is a chi-square test?**

Pearson’s chi-square (Χ2) tests, often referred to simply as chi-square tests, are among the most common [**nonparametric tests**](https://www.scribbr.com/statistics/statistical-tests/#nonparametric). Nonparametric tests are used for data that don’t follow the [assumptions of parametric tests](https://www.scribbr.com/frequently-asked-questions/assumptions-of-statistical-tests/), especially the assumption of a [normal distribution](https://www.scribbr.com/statistics/normal-distribution/).

If you want to [test a hypothesis](https://www.scribbr.com/statistics/hypothesis-testing/) about the distribution of a [**categorical variable**](https://www.scribbr.com/methodology/types-of-variables/#quantitative-vs-categorical) you’ll need to use a chi-square test or another nonparametric test. Categorical variables can be [nominal](https://www.scribbr.com/statistics/nominal-data/) or [ordinal](https://www.scribbr.com/statistics/ordinal-data/) and represent groupings such as species or nationalities. Because they can only have a few specific values, they can’t have a normal distribution.

A chi-square test (a test of independence) can test whether these observed frequencies are significantly different from the frequencies expected if handedness is unrelated to nationality.

**The chi-square formula**

Both of Pearson’s chi-square tests use the same formula to calculate the [test statistic](https://www.scribbr.com/statistics/test-statistic/), chi-square (Χ2):

\begin{equation*} X^2=\sum{\frac{(O-E)^2}{E}} \end{equation*}

Where:

Χ2 is the chi-square test statistic

Σ is the summation operator (it means “take the sum of”)

O is the observed frequency

E is the expected frequency

The larger the difference between the observations and the expectations (O – E) in the equation), the bigger the chi-square will be. To decide whether the difference is big enough to be [statistically significant](https://www.scribbr.com/statistics/statistical-significance/), you compare the chi-square value to a critical value.

**2.2 Hypothesis of the study**

H01: Attrition is independent on gender.

VS

H10: Attrition is dependent on gender.

H02: Attrition is independent on marital status.

VS

H20: Attrition is dependent marital status.

H03: Attrition is independent on business travel.

VS

H30: Attrition is dependent on business travel.

H04: Attrition is independent on overtime.

VS

H40: Attrition is dependent on overtime.

**2.3 Data Pre-processing and modelling**

A dataset with skewed proportions is called imbalanced dataset. In that a class that make up a large proportion of data are called majority class. Those that make up smaller proportion are called minor class. This could lead to degradation in the prediction or classification power of model. To balance the class proportion there are two different methods Over sampling and Under sampling.

**Over sampling**

Over sampling means creating a bootstrap sample from minor class with class length around length of majority class. Furthermore, interaction between regressors and response remains unaltered. But multiple rows being repeated multiple times which will create a mass around those duplicated values resulting bias in the method.

**Under sampling**

Under sampling reduces the class length of majority class up to the length of minority class. In this procedure we collect a random sample of size around the class length of minor class from major class. This might leads to loss of valuable information.

In our dataset, the count of Yes is 237 and that for No is 1233. Since the difference of class length is very high, we went for oversampling. After over sampling minority of class increases to class width of majority class.

Our first Aim was convert categorical data to numerical data we convert this data using label encoder in python and after converting this categorical data to numerical value then we use the machine learning algorithms for this project, we use the

1) Logistic Regression

2) Random Forest

3) Support Vector Machine

4) Naïve bayes

5)KNN

6)ANN

**Test dataset and train dataset:**

Separating data into train and test datasets is an important part of evaluating data. By this separation of total dataset into two datasets we can minimize effects of data inconsistency and better understand the characteristics of the model. Use the Train, Test data set in the fitting of models, first split the data 80% Training set and 20% Testing set. Here we have, 1176 records in train datasets and 294 records in test dataset. We apply data classification and prediction on test dataset. we train the model using train data set and check this model on test data set.

**1)Logistic Regression:**

**What is Logistic Regression?**

Logistic Regression is a “Supervised machine learning” algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems. Binary Classification refers to predicting the output variable that is discrete in two classes. A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, etc. The classification function calculates the statistics for a logistic response function. It shows the relationship between the dependent variable and independent variables. 𝜋(𝑦) = 1 1+exp(−𝑧) Here P(y) is our result which is determined with the help of dependent variable y, where z represents the function of independent variables used in the dataset. The range of values that P(y) predicts is from 0 to 1 which helps us to identify our category as no or yes as results.

**Assumption of Logistic Regression**

Dependent variable is binary-If this is not true, then logistic regression outputs do not apply.

Linearity between logit and independent variables

No multicollinearity-Multicollinearity distorts tests of statistical significance on regression coefficients.

Large sample size-This is more of a rule of thumb. A nice to have here is that your data has a balanced number of classes

Terms to know:

Odds Ratio =

The odds ratio specifies is defined as the probability of success as compared to the probability of failure. It is another way to represent probability, and is key to the interpretation of logistic regression coefficients.

An odds ratio of 1 means that the probability of success is equal to probability of failure. An odds ratio of 2 means that the probability of success is twice the probability of failure. An odds ratio of .5 means that the probability of failure is twice the probability of success.

Logit=log(odds)=log(

The logit is also known as the log of odds. It maps probabilities from (0, 1) to continuous values (-∞, ∞).

Logistic regression is a type of regression model we can use to understand the relationship between one or more predictor variables and a response variable when the response variable is binary.

If we only have one predictor variable and one response variable, we can use simple logistic regression, which uses the following formula to estimate the relationship between the variables:

log[p(X) / (1-p(X))] = β0 + β1X

The formula on the right side of the equation predicts the log odds of the response variable taking on a value of 1.

Simple logistic regression uses the following null and alternative hypotheses:

• H0: β1 = 0

• HA: β1 ≠ 0

The null hypothesis states that the coefficient β1 is equal to zero. In other words, there is no statistically significant relationship between the predictor variable, x, and the response variable, y. The alternative hypothesis states that β1 is not equal to zero. In other words, there is a statistically significant relationship between x and y.

If we have multiple predictor variables and one response variable, we can use multiple logistic regression, which uses the following formula to estimate the relationship between the variables:

log[p(X) / (1-p(X))] = β0 + β1x1 + β2x2 + … + βkxk

Multiple logistic regression uses the following null and alternative hypotheses:

• H0: β1 = β2 = … = βk = 0

• HA: β1 = β2 = … = βk ≠ 0

The null hypothesis states that all coefficients in the model are equal to zero. In other words, none of the predictor variables have a statistically significant relationship with the response variable, y.

The alternative hypothesis states that not every coefficient is simultaneously equal to zero.

**2) Random Forest**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, random forest is a classifier that contains a number of decision trees on various subsets of the given datasets and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output .The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

**Assumptions for Random Forest**

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

1. There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

2. The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

o It takes less training time as compared to other algorithms.

o It predicts output with high accuracy, even for the large dataset it runs efficiently.

o It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

There are mainly four sectors where Random forest mostly used:

1. Banking: Banking sector mostly uses this algorithm for the identification of loan risk.

2. Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.

3. Land Use: We can identify the areas of similar land use by this algorithm.

4. Marketing: Marketing trends can be identified using this algorithm.

Advantages of Random Forest

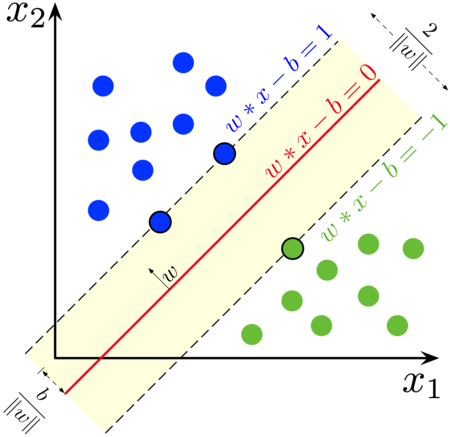
Random Forest is capable of performing both Classification and Regression tasks.

It is capable of handling large datasets with high dimensionality.

It enhances the accuracy of the model and prevents the overfitting issue.

**3) Support Vector Machine (SVM):**

Support vector machine is a classifier which is represented by a line that splits the data between the two differently classified groups of data representing the training data into two-dimensional planes as data points. We plot the features of the dataset Modern Approach for Loan Sanctioning in either side of the plane forming wo different categories. When the testing data lands on either side of the line, we can classify the data for approval as yes or no. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories divided by a clear gap that is as wide as possible.



**Fig. support vector machine**

**The followings are important concepts in SVM −**

**Support Vectors** − Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.

**Hyperplane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes

**Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

**Advantages:**

1. SVM works relatively well when there is a clear margin of separation between classes.

2. SVM is more effective in high dimensional spaces.

3. SVM is effective in cases where the number of dimensions is greater than the number of samples.

4. SVM is relatively memory efficient.

**4) Naïve bayes classifier:**

The Procedure for classifying a new record is as follows

Step-1: Find all the other records with same predictor records whether predictor values are same.

Step-2: Determine what classes the records belongs to and which class is most relevant (common).

Step-3: Assign the class to new record. The naïve bayes modification

For the basic classification procedure is as follows:

Let X=(X1,X2,…..,Xn) be n dimensional attribute vector describing n measurements made on the tuple from n attributes respectively say A1,A2,…..,An.

Suppose that there are m classes C1,C2,…….,Cm given a tuple X the classifier will predict that X belongs to the class having the highest posterior probability condition on X

i.e. according to naïve bayes classifier tuple X belongs to class Ci if and only if P(CI|X)>P(CJ|X) for

The class Ci for which P(Ci|X) is maximized is called maximum posterior probability.

By bayes theorem, P(CI|X)=

Given datasets with many attributes it would be extremely computationally extensive to compute P(X|Ci).To reduce the computation in evaluating it, naïve bayes assumption of class conditional independence is made and thus

P(X|Ci)=P(X1|C1).P(X2|C2)….P(Xn|Cn)

=∏ P(Xj|Cj) ,where Xj is the value of attribute for tuple X.

i

If Aj is categorical then P(Xj|Ci) is the number of tuples of class Ci in D having the value Xj for Aj divided by tuples of class C in D. If Aj is continuous value then typically it is assume to have gaussian distribution with mean μ and variance .

P(Xj|Ci)=f( where and is the average and variance of the values of attribute Aj for training tuples of class C.To predict the class label of X P(X|Ci).P(Ci) is evaluated for each class Ci ,the classifier predicts that the class label of tuple X is class Ci if and only if P(X|Ci).P(Ci) > P(X|Cj).P(Cj) for all j=1,2,….,n and i≠j.

**Advantages of Naive Bayes Classifier**

It is simple and easy to implement.

It doesn't require as much training data.

It handles both continuous and discrete data.

It is highly scalable with the number of predictors and data points

It is fast and can be used to make real-time predictions.

**5) K- Nearest Neighbourhood**

**What is K-Nearest Neighbours Algorithm?**

K-Nearest Neighbours is a part of general technique known as instance based learning which does no tone of the most basic yet essential classification algorithms in Machine Learning. It belongs to the [supervised learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) domain and finds intense application in pattern recognition, [data mining](https://www.geeksforgeeks.org/data-mining/), and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a [Gaussian distribution](https://www.geeksforgeeks.org/mathematics-probability-distributions-set-3-normal-distribution/) of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

**Intuition Behind KNN Algorithm**

If we plot these points on a graph, we may be able to locate some clusters or groups. Now, given an unclassified point, we can assign it to a group by observing what group its nearest neighbours belong to. This means a point close to a cluster of points classified as ‘Red’ has a higher probability of getting classified as ‘Red’. Intuitively, we can see that the first point (2.5, 7) should be classified as ‘Blue’ and the second point (5.5, 4.5) should be classified as ‘Red’.

**Distance Metrix Used in KNN Algorithm**

As we know that the KNN algorithm helps us identify the nearest points or the groups for a query point. But to determine the closest groups or the nearest points for a query point we need some metric. For this purpose, we use below distance matrix.

**Euclidean Distance**

This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object.

**How to choose the value of k for KNN Algorithm?**

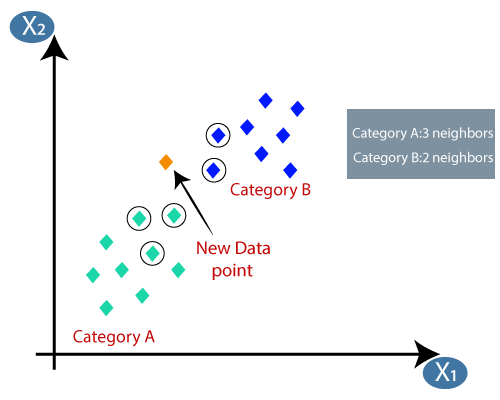
The value of k is very crucial in the KNN algorithm to define the number of neighbours in the algorithm. The value of k in the k-nearest neighbours (k-NN) algorithm should be chosen based on the input data. If the input data has more outliers or noise, a higher value of k would be better. It is recommended to choose an odd value for k to avoid ties in classification. [Cross-validation](https://www.geeksforgeeks.org/cross-validation-machine-learning/) methods can help in selecting the best k value for the given dataset.

**Advantages of the KNN Algorithm**

**Easy to implement** as the complexity of the algorithm is not that high.

**Adapts Easily** – As per the working of the KNN algorithm it stores all the data in memory storage and hence whenever a new example or data point is added then the algorithm adjusts itself as per that new example and has its contribution to the future predictions as well.

**Few Hyperparameters** – The only parameters which are required in the training of a KNN algorithm are the value of k and the choice of the distance metric which we would like to choose from our evaluation matrix.

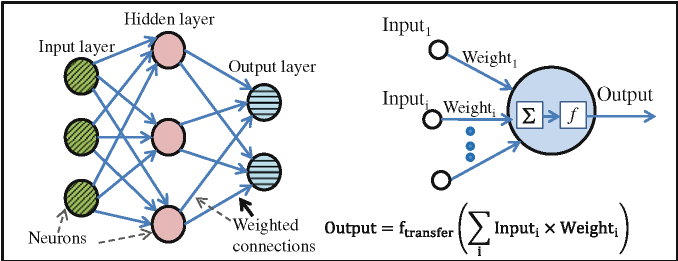


**Fig. K-Nearest neighbour**

**6) Artificial Neural Networks**

Artificial Neural Networks contain artificial neurons which are called units. These units are arranged in a series of layers that together constitute the whole Artificial Neural Network in a system. A layer can have only a dozen units or millions of units as this depends on how the complex neural networks will be required to learn the hidden patterns in the dataset. Commonly, Artificial Neural Network has an input layer, an output layer as well as hidden layers. The input layer receives data from the outside world which the neural network needs to analyse or learn about. Then this data passes through one or multiple hidden layers that transform the input into data that is valuable for the output layer. Finally, the output layer provides an output in the form of a response of the Artificial Neural Networks to input data provided.

In the majority of neural networks, units are interconnected from one layer to another. Each of these connections has weights that determine the influence of one unit on another unit. As the data transfers from one unit to another, the neural network learns more and more about the data which eventually results in an output from the output layer.



**Fig. Artificial Neural Networks**

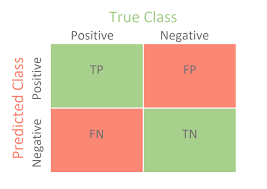
**2.4 Model evaluation metrics and interpretation approach**

**Confusion matrix**

**What is a confusion matrix?**

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

For a binary classification problem, we would have a 2 x 2 matrix as shown below with 4 values:



**Fig. confusion matrix**

Table shows a few parameters such as True Positive values (TP), True Negative values (TN), False Positive values (FP), and False Negative values (FN).

Based on these parameters, model accuracy has been calculated using the following evaluation metrics:

**True Positive (TP)**

The predicted value matches the actual value

The actual value was positive and the model predicted a positive value

**True Negative (TN)**

The predicted value matches the actual value

The actual value was negative and the model predicted a negative value

**False Positive (FP) – Type 1 error**

The predicted value was falsely predicted

The actual value was negative but the model predicted a positive value

Also known as the Type 1 error

**False Negative (FN) – Type 2 error**

The predicted value was falsely predicted

The actual value was positive but the model predicted a negative value

Also known as the Type 2 error

**Different Metrices use on Confusion Matrix**

**Accuracy**: It means the total number of two correct predictions divided by the total number of data set.

**Accuracy = 𝑇𝑝+𝑇𝑁 /𝑇𝑃+𝐹𝑃+𝑇𝑁+𝐹𝑁**

**Sensitivity**: It tells us how many of the correctly predicted cases actually turned out to be positive

**Sensitivity = 𝑇𝑃 /𝑇𝑃+𝐹𝑁**

**Specificity**: It tells how many of the actual positive cases we were able to predict correctly with our model

**Specificity = 𝑇𝑁 /𝑇𝑁+𝐹𝑃**

 **F1-Score**: F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. But there is a catch here. The interpretability of the F1-score is poor.

**F1-Score = 2∗(𝑅𝑒𝑐𝑎𝑙𝑙∗𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛) (𝑅𝑒𝑐𝑎𝑙𝑙+𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 )**

**Chapter 3**

**Statistical Analysis and Discussion**

**3.1 Data Visualization**

**3.1.1] Age wise distribution of employee’s attrition**:

A bar chart or bar graph is a chart or graph that present categorical data with rectangular bars which heights and values proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is some times called as column chart.

Here we plot a bar plot between the two variables and these are attrition rate and age.

On X axis we take a variable age start from 18 years increasing by 2 years per unit.

On Y axis we take a variable attrition Rate start from 0 employee’s increasing by 2 employees per unit.

Fig. 3.1.1 BAR CHART OF ATTRITION COUNT BY AGE

From the graph, we observe that attrition rate is high between age 24 to 37. And medium at age 18 to 23 and low between the age 38 to 60.

That is the employees are don’t want to take risk at the age of 38 to 60 they want to live steady life. Hence attrition rate is low.

**3.1.2] Business travel wise distribution of employee’s attrition:**

Here we plot a bar plot between the two variables and these are attrition rate and business travel.

On X axis we take a variable business travel, there are three class of business travel are: non-travel, travel-frequently and travel-rarely.

On Y axis we take a variable attrition rate start from 0 employee’s increasing by 2 employees per unit.

Fig.3.1.2 BAR CHART OF ATTRITION RATE BY BUSINESS TRAVEL

From the graph we observe there are 12 employees are not travelling, 69 employees are travelling frequently and 156 employees travel rarely.

The attrition rate is increases with increasing business travel.

**3.1.3] Education field wise distribution of employee’s attrition:**

Here we plot a stacked bar plot between the two variables and these are attrition rate and education field.

On X axis we take a variable attrition rate is increasing by 20 employees per unit.

On Y axis we take a variable education field, there are 6 class of education field are: Human resources, life science, marketing, medical, other, technical degree.

Fig.3.1.3 STACKED BAR PLOT ATTRITION RATE BY EDUCATION FIELD

From the graph, we observe that Attrition rate is high in the field of medical, life sciences and low in the field of human resources and other field and medium in the technical degree and marketing.

**3.1.4] Job role wise distribution of attrition rate:**

Here we plot a stacked bar plot between the two variables and these are attrition rate and job role.

On X axis we take a variable attrition rate

On Y axis we take a variable job role.

**Fig 3.1.4** Job role wise distribution of attrition rate

From the graph, we observe that attrition rate is high in sales representative, sales exclusive, research scientist, Laboratory Technician job role.

And low in research director, manufacturing director, manager human resources and healthcare representative.

**3.1.5] Department wise distribution of attrition rate:**

Here we plot a stacked bar plot between the two variables and these are attrition rate and department.

On X axis we take a variable attrition rate by taking limit 20 employee per unit.

On Y axis we take a department have three classes are human resources research and development, sales.

**Fig 3.1.5** Department wise distribution of attrition rate.

From the graph, we observe that attrition rate is high in sales department. 92 employees leave there job and research and development department 133 employees leave their job.

And low in human resources department 12 employee’s leave there job.

**3.1.6] Gender wise distribution of attrition count:**

A doughnut chart shows the relationship of parts to a whole but a doughnut chart contain more than one data series. Each data series that you plot in doughnut chart adds a ring to the chart.

Here we plot a doughnut chart of gender wise attrition rate.

**Fig 3.1.6** Gender wise distribution of attrition Rate.

From the graph, we observe that out of 237 employees there are 87 (37%) female employees leave the job and 150 (63%) male employee’s leave there job.

That is the attrition rate in males is greater than the attrition rate in females.

**3.1.7] Job satisfaction wise distribution of employee’s attrition:**

Here we plot a stacked bar plot between the two variables and these are attrition rate and job satisfaction.

On X axis we take a variable attrition rate 20 employee’s per unit.

On Y axis we take a variable job satisfaction rated on the scale of 1 to 4.

1 being highest and 4 being lowest satisfaction.

**Fig 3.1.7** Job satisfaction wise distribution of employee’s attrition.

From the graph, we observe that 66 employee’s are highly satisfied but they leave there job. 46 employee’s are satisfied but they leave there job. 73 employee’s are low satisfied and they leave there job. 52 employee’s are un satisfied therefore they leave there job.

**3.1.8] Relationship satisfaction wise distribution of attrition count:**

Employee satisfaction definition is a term that is used to describe if employees are happy and fulfilling their desires and needs at work. The crucial factor with employee satisfaction is that satisfied employees must do the job and make the contributions that the employer needs.

Here we plot a bar plot between the two variables and these are attrition rate and relationship satisfaction.

On X axis we take a variable relationship satisfaction were 1 is highly satisfied and 4 is un satisfied.

On Y axis we take a variable attrition rate 10 employee’s per unit.

**Fig. 3.1.8** Relationship satisfaction wise distribution of attrition rate.

From the graph, we observe that 57 highly satisfied employee’s are leave their job, 45 satisfied employees leave their job and 71 unsatisfied employees leave their job as well as 64 highly unsatisfied employee’s leave their job.

**3.1.9] Work life balancewise distribution of employee’s attrition:**

Work-life balance is typically defined as the amount of time you spend doing your job versus the amount of time you spend with loved ones or pursuing personal interests and hobbies.

Here we plot a bar plot between the two variables and these are attrition count and job role.

On X axis we take a variable work life balance .1 is balanced and 4 is unbalanced.

On Y axis we take a variable attrition rate scale is 20 employee’s per unit.

**Fig.3.1.9** Work life balancewise distribution of employees attrition.

From the graph we observe that the attrition rate is high in unbalanced work life and low in highly balanced work life.

**3.1.10] Years in current rolewise distribution of employees attrition:**

A line chart or line graph, also known as curve chart, is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields.

Here we plot a line chart between the two variables and these are attrition rate and job role.

On X axis we take a variable years in current role. And scale is 1year per unit.

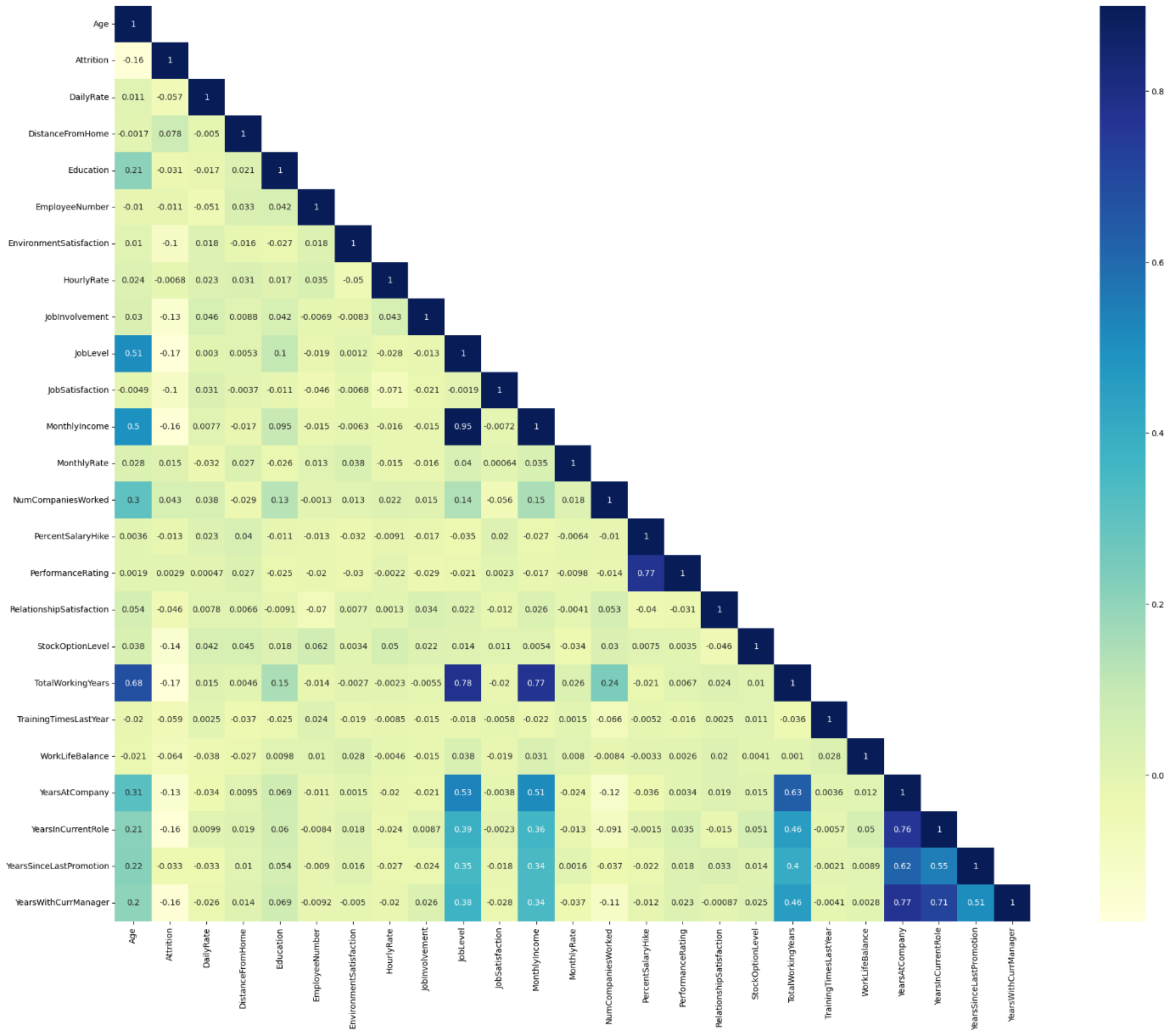
On Y axis we take a variable attrition rate and scale is 10 employee’s per unit.

**Fig 3.1.10** Years in current rolewise distribution of employees attrition.

From the graph, we observe that the attrition rate is high at first 4 years and then it minimises with increas in years in current role.

**3.1.11] Heatmap**

A heat map is a 2-dimensional data visualization technique that represents the magnitude of individual values within a dataset as a color. The variation in color may be by hue or intensity. "Heat map" is a relatively new term, but the practice of shading matrices has existed for over a century.



**Fig.3.1.12 Heatmap**

This figure shows the correlation between two features.Correlation of age with itself is 1 it indicates that strong positive correlation. Age and Attrition are negatively correlated.Education and distance from home are positively correlated.Like this interpretation is for all features.

**3.2 Employees Attrition Dashboard**

**Analysis:** By analysing various metrics such as employee count, attrition, attrition rate, average age of employees leaving the organization, average salary, average years worked at company, attrition by gender, attrition by age group, job satisfaction rating, attrition by job roles, education field wise attrition.

**Aim:** To achieve valuable insights by creating meaningful relationships between these attributes.

**Objective:** To gain insights into the attrition patterns within the organization and helping the organization to gain a deeper understanding of employee attrition by visualizing attrition data and trends.

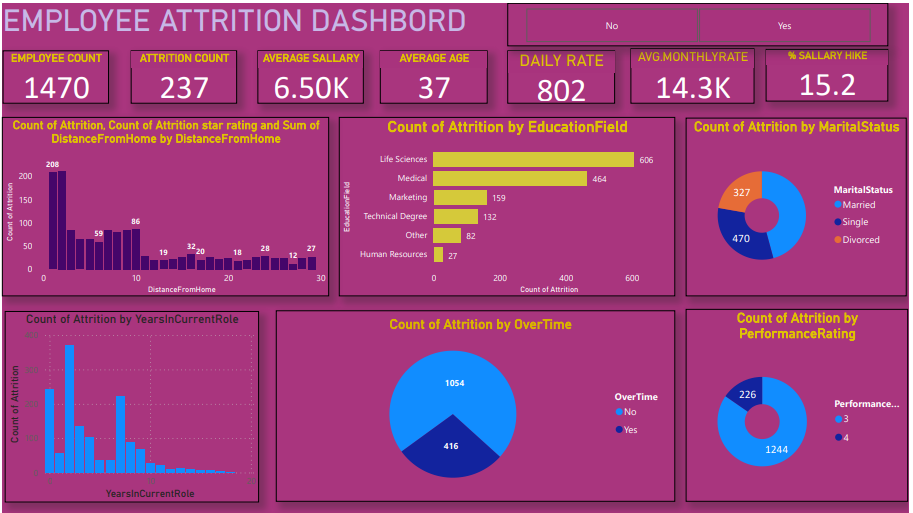
**Process:**

1.Importing the dataset into Power Bi, transforming the data, data cleaning (removing duplicates, null values).

2.Using DAX, calculated Average age, Average salary etc. and created additional columns (conditional column) as per requirement.

3.I have included important KPIs such as Total employees, Attrition count, Attrition rate, Average age, average salary and Years at company. 4. Finally created different types of visualizations such as donut chart, stacked bar chart, stacked column chart, tree map, matrix, area chart to get. meaningful insights.

**3.2.1** Employees Attrition Dashboard of Employees With And Without Attrition:

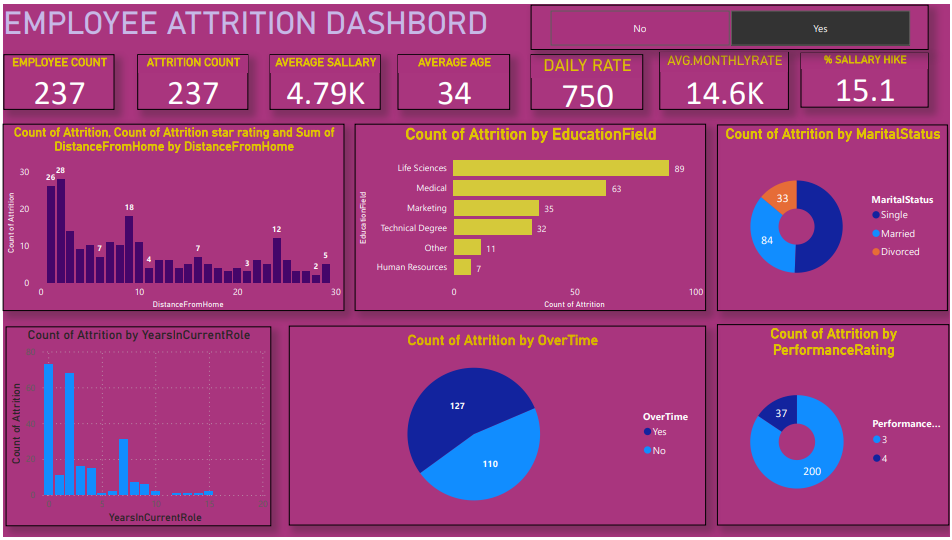


**Fig. 3.2.1** Employees Attrition Dashboard of Employees With And Without Attrition

Insights:

* In the first plot we obesrve that count of attrition is high in employees whose distance from distance is between 0 to 2.
* In the second plot,count of attrition is high in life sciences field and low in human resources field.
* In the third plot,count of attrition is high in married employees and low in divorced employees.
* In the fourth plot, count of attrition is high in 3 years in current role and low in 18 yers in current role.
* In the fifth plot out of 1470 employees, 1054 work with overtime and 416 work without overtime.
* In the sixth plot there are 1244 employees having high performance rating and 226 employees having very high performance rating.

**3.2.2** Employees Attrition Dashboard of Employees With Attrition

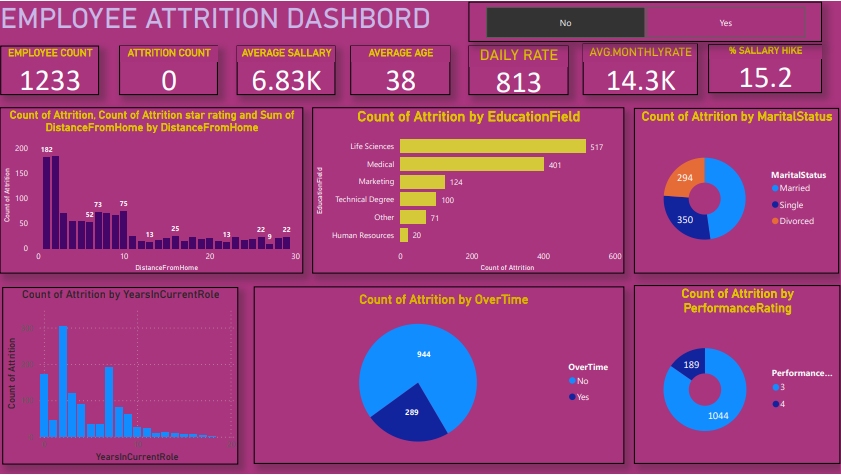


**Fig.3.2.2** Employees Attrition Dashboard of Employees With Attrition

Insights:

* In the first plot we obesrve that count of attrition is high in employees whose distance from distance is between 0 to 2.
* In the second plot,count of attrition is high in life sciences field and low in human resources field.
* In the third plot,count of attrition is high in single employees and low in divorced employees.
* In the fourth plot, count of attrition is high in o years in current role and low in 15 years in current role.
* In the fifth plot out of 237 employees, 127 work with overtime and 110 work without overtime.
* In the sixth plot there are 200 employees having high performance rating and 37 employees having very high performance rating.

**3.2.3** Employees Attrition Dashboard Of Employees Without Attrition

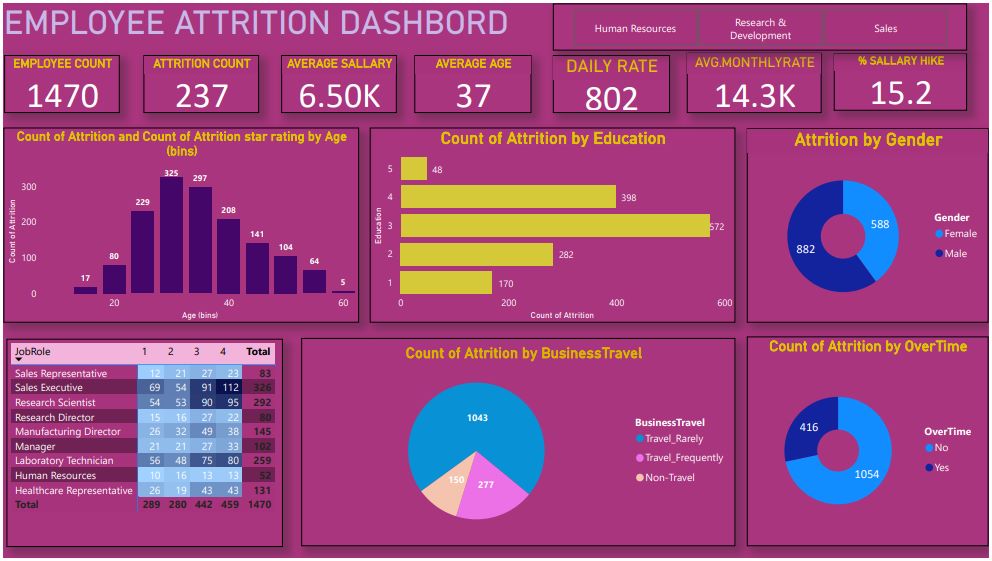


**Fig.** **3.2.3** Employees Attrition Dashboard Of Employees Without Attrition

Insights:

* In the first plot we obesrve that count of attrition is high in employees whose distance from distance is between 0 to 2.
* In the second plot,count of attrition is high in life sciences field and low in human reources field.
* In the third plot,count of attrition is high in married employees and low in divorced employees.
* In the fourth plot, count of attrition is high in 2 years in current role and low in 18 years in current role.
* In the fifth plot out of 1233 employees, 289 work with overtime and 944 work without overtime.
* In the sixth plot there are 1044 employees having high performance rating and 189 employees having very high performance rating.

**3.2.4** Employees Attrition Dashboard Of Employees Attrition In Department



**Fig.3.2.4** Employees Attrition Dashboard Of Employees Attrition In Department.

Insights:

* Total employee count is 1470 among which 237 workers are leaving the organization out of which 133 of them are from Research & development department.
* Average age of employees is 37, whereas the highest number of employees leaving the company are between the age group of 25-35.
* In the second plot count of attrition is high in medium level of education.
* In the third plot out of 1470 employees attrition count in females is 588 and in males is 882.
* In the fourth table job satisfaction i.e. low satisfied,moderately satisfied,high satisfied and very high satisfied and there job roles are shown.
* In the fifth plot count of attrition is 1043 in employees travel rarely , 277 travel frequently and 150 in non-travel employees.
* In the fifth plot out of 1470 employees, 416 work with overtime and 1054 work without overtime.

**Chapter 4**

**Statistical analysis and results**

**3.3 Data Analysis Using Classifiers**

To analyze the the attrition and retention in the company we will be use chi square test,t test, Descriptive statistics, Regression, Machine learning, Classification, Datamining and softwares like Python, Rsoftware, Tableu, SPSS.By using this we will be identify which factors are significant and determining relationship between this variables and find reasons behind the attrition and solutions to retain them.

**1) Logistic Regression**

ROC-curve of imbalanced dataset-

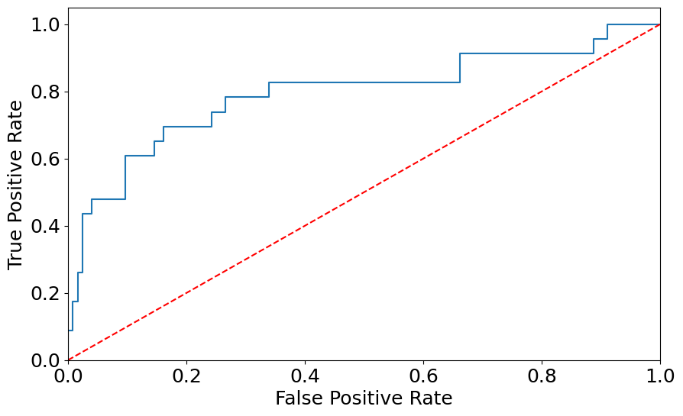


Fig. ROC-curve

Firstly we apply logistic regression on imbalanced dataset but area under curve of logistic regression is 0.5.so,we balance the data using random oversampling methd and both the yes and no class becomes 1233.

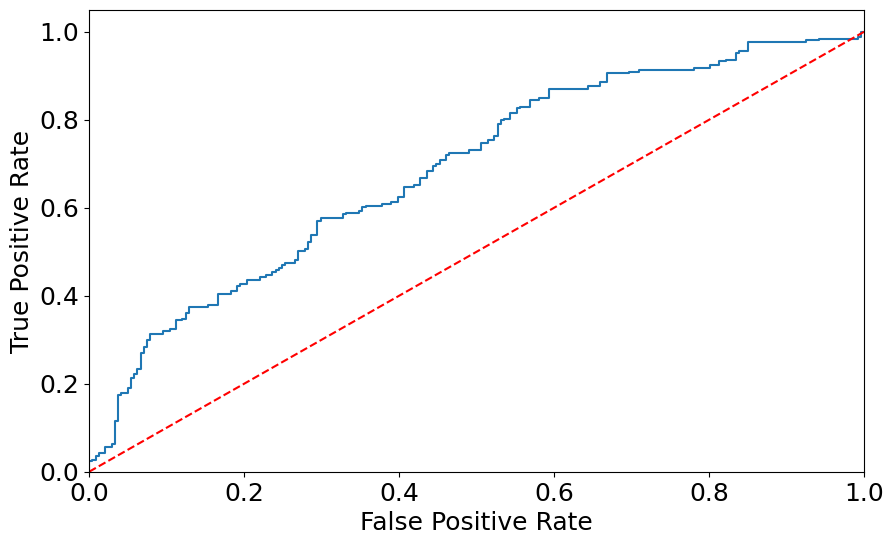


Fig. ROC-curve

After applying oversampling method area under curve of ROC becomes 0.61.

**1) Logistic Regression**

**Before Oversampling**

Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 255 | 0 |
| 0 | 39 | 0 |

Fig. Confusion Matrix

This confusion matrix shows that correctly classified attrition count is 255 ,incorrectly positive classified attrtion count is 39,correctly negative classified attrition count is 0,incorrectly negative classified attrition count is 0.

|  |  |
| --- | --- |
| Accuracy | 0.84 |
| Sensitivity | 0.8673 |
| Specificity | 0 |
| Precision | 0.80 |
| F1 score | 0.91 |
| Area under ROC curve | 0.50 |

Since, accuracy of logistic model is 0.84 it implies that 84 % of times test results are correct.Area under ROC curve is 0.50 means logistic regression correctly classifies only 50% observations so we are using oversampling method to increase area under ROC Curve.

After Oversampling

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 143 | 98 |
| 0 | 95 | 158 |

Fig. Confusion Matrix

This confusion matrix shows that correctly classified attrition count is 143 ,incorrectly positive classified attrtion count is 95,correctly negative classified attrition count is 98,incorrectly negative classified attrition count is 158.

|  |  |
| --- | --- |
| Accuracy | 0.6093 |
| Sensitivity | 0.61 |
| Specificity | 0.9919 |
| Precision | 0.0105 |
| F1 score | 0.91 |
| Area under ROC curve | 0.6089 |

Since, accuracy of logistic model is 0.6093 it implies that 60.93 % of times our test results are correct.Area under ROC curve is 0.61 means logistic regression correctly classifies only 61% observations so by using oversampling area under curve increases and accuracy decreases.

**2) Random Forest (RF):**

Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 247 | 0 |
| 0 | 45 | 2 |

Fig. Confusion matrix

This confusion matrix shows that correctly classified attrition count is 247 ,incorrectly positive classified attrtion count is 45,correctly negative classified attrition count is 0,incorrectly negative classified attrition count is 2.

|  |  |
| --- | --- |
| Accuracy | 0.85 |
| Sensitivity | 1 |
| Specificity | 0.042 |
| Precision | 0.85 |
| F1 score | 0.92 |

Since, accuracy of random forest model is 0.85 it implies that 85 % of times our test results are correct.

**3) Support Vector Machine (SVM) :**

|  |  |
| --- | --- |
| Accuracy | 0.88 |
| Sensitivity | 0.98 |
| Specificity | 0 |
| Precision | 0.88 |
| F1 score | 0.93 |

Since, accuracy of support vector machine model is 0.88 it implies that 88 % of times our test results are correct.

**4) Naïve bayes classifier:**

|  |  |
| --- | --- |
| Accuracy | 0.813 |
| Sensitivity | 0.88 |
| Specificity | 0.89 |
| Precision | 0.89 |
| F1 score | 0.88 |

Since, accuracy of naïve bayes classifier model is 0.813 it implies that 81.3 % of times our test results are correct.

**5)K- Nearest Neighbours (KNN):**

Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 232 | 15 |
| 0 | 39 | 8 |

Fig. Confusion matrix

This confusion matrix shows that correctly classified attrition count is 232 ,incorrectly positive classified attrtion count is 39,correctly negative classified attrition count is 15,incorrectly negative classified attrition count is 8.

|  |  |
| --- | --- |
| Accuracy | 0.82 |
| Sensitivity | 0.9392 |
| Specificity | 0.1702 |
| Precision | 0.86 |
| F1 score | 0.90 |

Since, accuracy of K-nearest neighbour model is 0.82 it implies that 82 % of times our test results are correct.

**6) Artificial neural network :**

Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 207 | 29 |
| 0 | 32 | 26 |

Fig. Confusion matrix

This confusion matrix shows that correctly classified attrition count is 207 ,incorrectly positive classified attrtion count is 32,correctly negative classified attrition count is 29,incorrectly negative classified attrition count is 26.

|  |  |
| --- | --- |
| Accuracy | 0.79 |
| Sensitivity | 0.88 |
| Specificity | 0.4482 |
| Precision | 0.87 |
| F1 score | 0.87 |

Since, accuracy of artificial neural network model is 0.79 it implies that 79 % of times our test results are correct.

From the above analysis, we observe that accuracy of support vector machine is maximum and accuracy of artificial neural network is minimum among above 6 classifiers.So,Support vector machine classifier performs better than other 6 classifiers.

**3.4 Test Used For Analysis:**

**1.Chi square test of independence of attributes (Gender)**

**Hypothesis:**

H01: Attrition is independent on gender.

Vs

H10: Attrition is dependent on gender.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Crosstab** | | | | |
|  |  | **Female** | **Male** | **Total** |
| **Gender** | **No** | 501 | 732 | 1233 |
| **Yes** | 87 | 150 | 237 |
| **Total** | | 588 | 882 | 1470 |

Pearson chi square test = 1.2752

Since p value = **0.2588** is greater than level of significance (0.05).So, we accept H01 at

**Conclusion:-** Attrition is independent on gender.

**2.Chi square test of independence of attributes (Marital status)**

**Hypothesis:**

H02: Attrition is independent on marital status.

Vs

H20: Attrition is dependent on marital status.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
| **Count** | | | | | |
|  |  | **Divorced** | **Married** | **Unmarried** | **Total** |
| **Marital status** | **No** | 294 | 589 | 350 | 1233 |
| **Yes** | 33 | 84 | 120 | 237 |
| **Total** | | 327 | 673 | 470 | 1470 |

Pearson chi square =46.164

Since p value = **9.456e-11** is less than level of significance (0.05).So, we reject H02 at

**Conclusion:-** Attrition is dependent on marital status.

**3.Chi square test of independence of attributes (Business travel)**

**Hypothesis:**

H03: Attrition is independent on business travel.

Vs

H30: Attrition is dependent on business travel .

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
| **Count** | | | | | |
|  |  | **Non-travel** | **Travel-rarely** | **Travel-frequently** | **Total** |
| **Business travel** | **No** | 138 | 208 | 887 | 1233 |
| **Yes** | 12 | 69 | 156 | 237 |
| **Total** | | 150 | 277 | 1343 | 1470 |

Pearson chi square= 24.182

Since p value = **5.609e-06** is less than level of significance (0.05).So, we reject H03 at

**Conclusion:-** Attrition is dependent on business travel.

**4.Chi square test of independence of attributes (Overtime)**

**Hypothesis**:

H04: Attrition is independent on overtime.

Vs

H40: Attrition is dependent on overtime.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Crosstab** | | | | |
| **Count** | | | | |
|  |  | **No** | **Yes** | **Total** |
| **Overtime** | **No** | 944 | 289 | 1233 |
| **Yes** | 110 | 127 | 237 |
| **Total** | | 1054 | 416 | 1470 |

Pearson chi square= 89.044

Since p value = 2.2e-16 is less than level of significance (0.05).So, we reject H04 at

**Conclusion**:- Attrition is dependent on overtime.

**Chapter 4**

**Summary and conclusions**

**Results by visualization:**

1] Attrition count is high at age 35 and low at age 57. Attrition count is high in the field of life sciences and low in the field of human resources.

2] Attrition count attrition count is high in job role of sales executive and low in human resources.

3] In males is greater than the attrition count in females.

4] Attrition count is high in low satisfied with their with the staff members in the company.

5] Attrition count is high in low satisfied relationhip with the staff members in the company.

6] Attrition count is high in employees whose work life balance is better and low in employees whose work life balance is good.

7] Attrition count is high in employees having 2 years in current role and low in 18 years in current role.

**Employees attrition dashboard**

1] Count of attrition is high in employees whose distance from distance is between 0 to 2.

2] Count of attrition is high in life sciences field and low in human resources field.

3] Count of attrition is high in married employees and low in divorced employees.

4] Count of attrition is high in 3 years in current role and low in 18 yers in current role.

5] Out of 1470 employees, 1054 work with overtime and 416 work without overtime.

6] There are 1244 employees having high performance rating and 226 employees having very high performance rating.

7] Attrition count is high in the field of life sciences and low in the field of human resources.

8] Attrition count attrition count is high in job role of sales executive and low in human resources.

9] Attrition count in males is greater than the attrition count in females.

10]Attrition count is high in low satisfied with their with the staff members in the company.

11] Attrition count is high in low satisfied relationhip with the staff members in the company.

12] Attrition count is high in employees whose work life balance is better and low in employees whose work life balance is good.

13] Attrition count is high in employees having 2 years in current role and low in 18 years in current role.

**Employees attrition dashboard of employees with attrition**

1] In the first plot we obesrve that count of attrition is high in employees whose distance from distance is between 0 to 2.

2] In the second plot,count of attrition is high in life sciences field and low in human resources field.

3] In the third plot,count of attrition is high in single employees and low in divorced employees.

4] In the fourth plot, count of attrition is high in o years in current role and low in 15 years in current role.

5] In the fifth plot out of 237 employees, 127 work with overtime and 110 work without overtime.

6] In the sixth plot there are 200 employees having high performance rating and 37 employees having very high performance rating.

**Employees attrition dashboard of employees without attrition**

1] In the first plot we obesrve that count of attrition is high in employees whose distance from distance is between 0 to 2.

2] In the second plot,count of attrition is high in life sciences field and low in human resources field.

3] In the third plot,count of attrition is high in married employees and low in divorced employees.

4] In the fourth plot, count of attrition is high in 2 years in current role and low in 18 years in current role.

5] In the fifth plot out of 1233 employees, 289 work with overtime and 944 work without overtime.

6] In the sixth plot there are 1044 employees having high performance rating and 189 employees having very high performance rating.

**Employees attrition dashboard of employees attrition in department**

1] In the first plot attrition count is high in employees whose age between 25-35.

2] In the second plot count of attrition is high in medium level of education.

3] In the third plot out of 1470 employees attrition count in females is 588 and in males is 882.

4] In the fourth table job satisfaction i.e. low satisfied,moderately satisfied,high satisfied and very high satisfied and there job roles are shown.

5] In the fifth plot count of attrition is 1043 in employees travel rarely , 277 travel frequently and 150 in non-travel employees.

6] In the fifth plot out of 1470 employees, 416 work with overtime and 1054 work without overtime.

**Results by applying different classifiers**:

1] Before applying over sampling method logistic regression correctly classifies only 50%

observations

2] After applying over sampling method logistic regression correctly classifies 61%

observations

3] Accuracy of random forest model is 0.85 it implies that 85 % of times our test results are

correct.

4] Accuracy of support vector machine model is 0.88 it implies that 88 % of times our test

results are correct.

5] Accuracy of naïve bayes classifier model is 0.813 it implies that 81.3 % of times our test

results are correct.

6] Accuracy of K-nearest neighbour model is 0.82 it implies that 82 % of times our test

results are correct.

7] Accuracy of artificial neural network model is 0.79 it implies that 79 % of times our test

results are correct.

**Conclusions**

1. SVM model would be best fitted model for the dataset in hand.
2. Highly contributing factors which affects employees attrition are educational field,distance from home,years in current role,education.
3. Demographic factor like age,employees of age 25 to 35 having maximum rate of leaving the company Employees in the life science field having maximum rate of leaving the company.
4. Satisfaction level toward job is high in sales executive.
5. Factors such as low job satisfaction and low environment satisfaction that make employees dissatisfied.
6. Need of employees for searching new job is because of overtime.
7. Reasons for attrition among emplyees are medium education level , overtime, greater distance from home,low satisfaction level.
8. Employees whose education is medium having maximum rate of leaving the company.
9. The accuracy of SVM is better and this classifiers can be deployed to predict the employee’s are likely to leave the company.
10. Attrition is independent of gender and dependent of marital status,overtime and business travel.

**3.5 Data analysis using python and R:**

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer

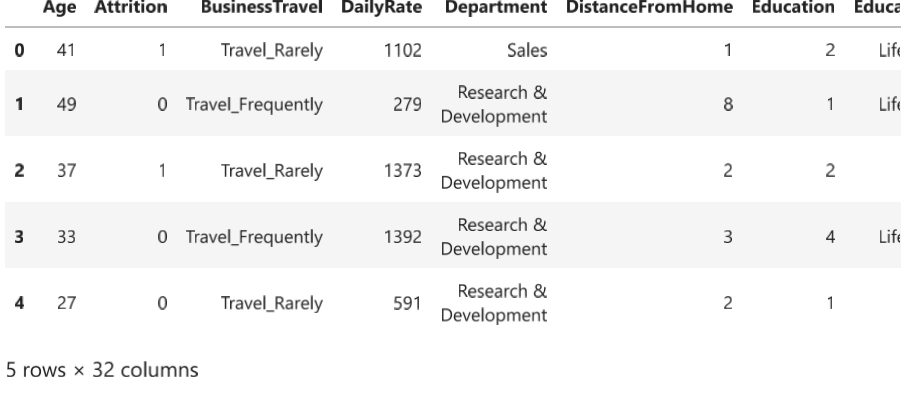
Description automatically generated with medium confidence

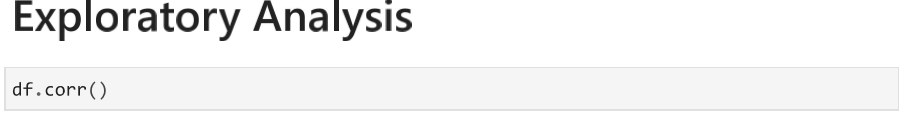
A screenshot of a computer

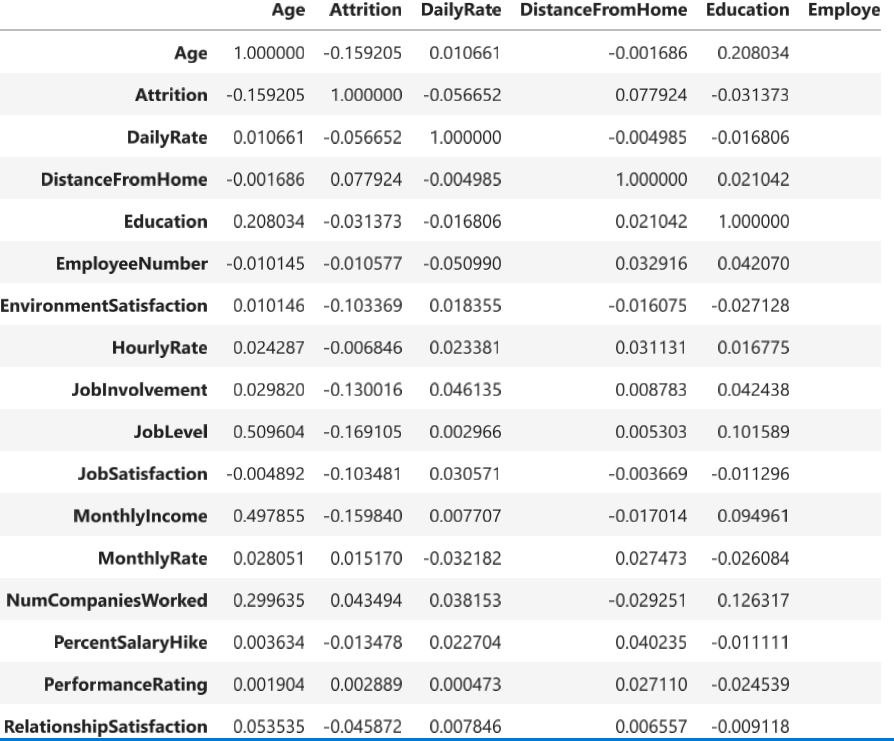
Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with low confidence



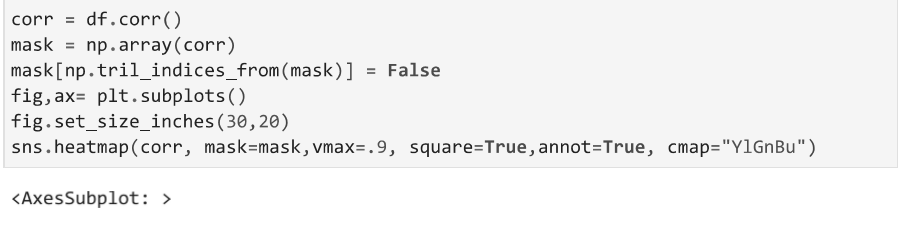




A screenshot of a calculator

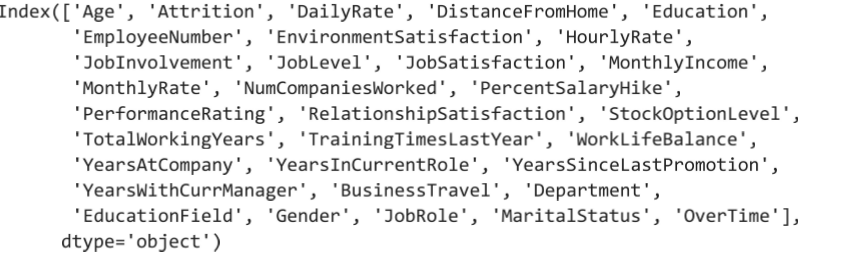
Description automatically generated with medium confidence

A picture containing screenshot, text, diagram

Description automatically generated

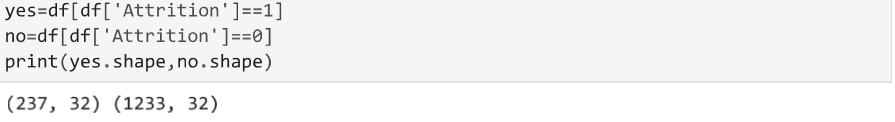


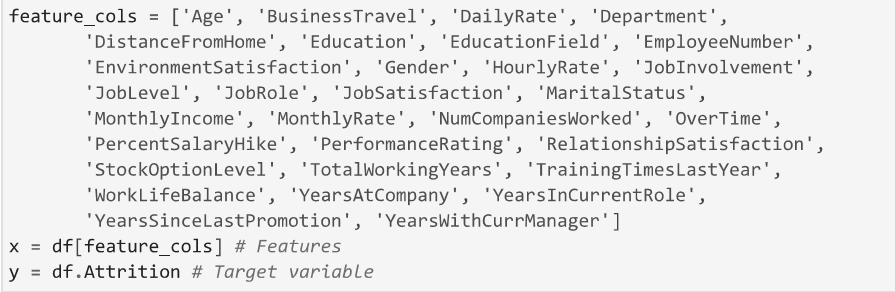




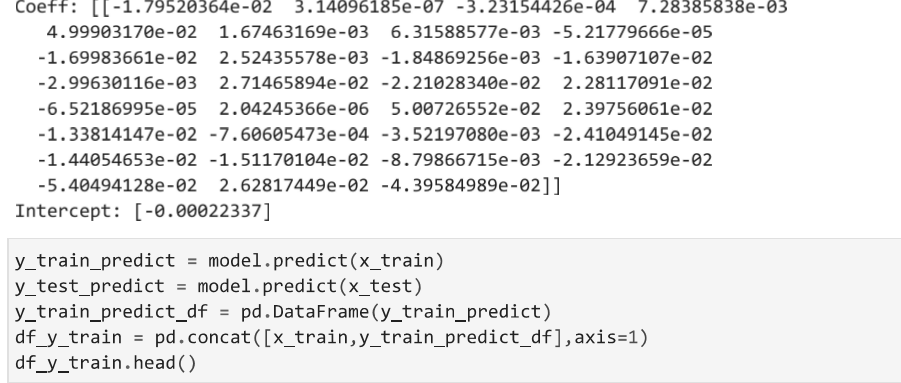
A picture containing text, white, font, document

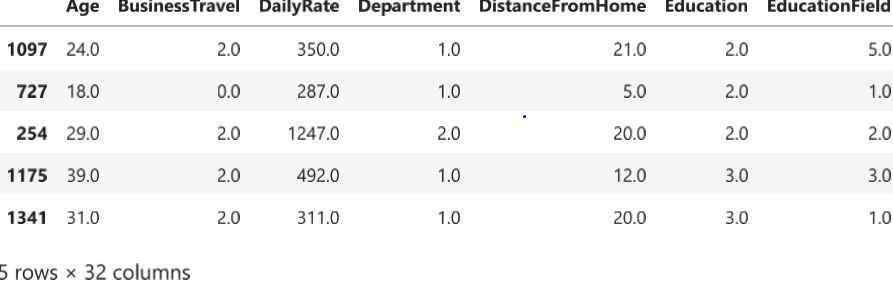
Description automatically generated

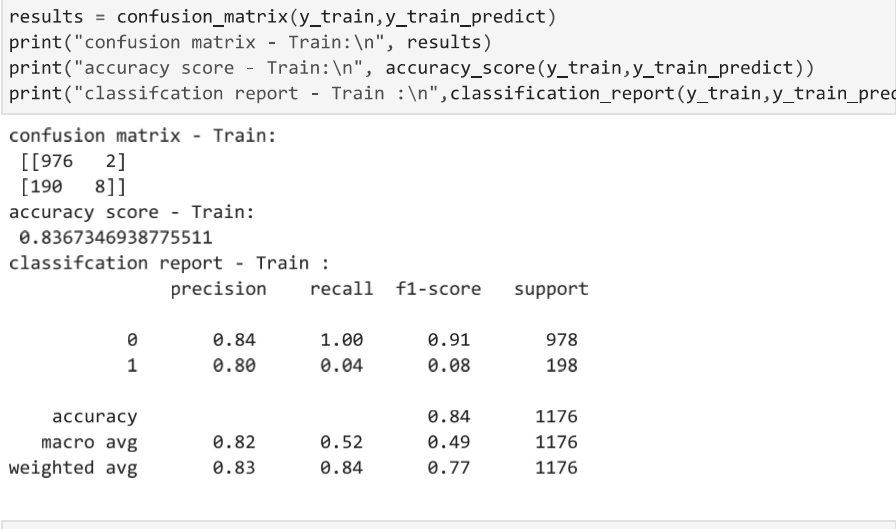






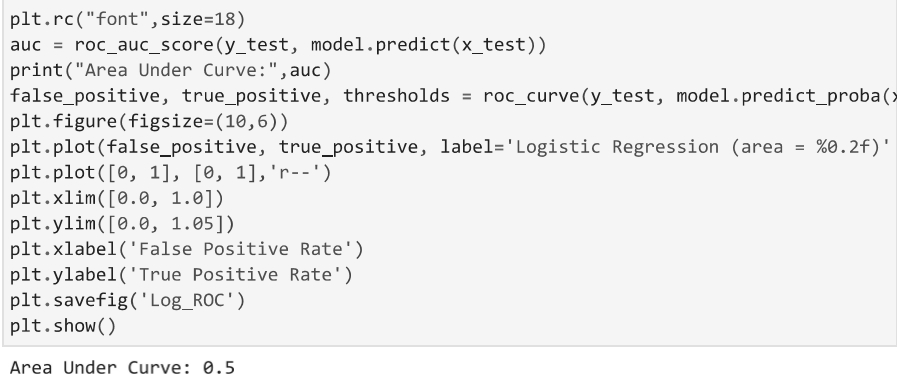


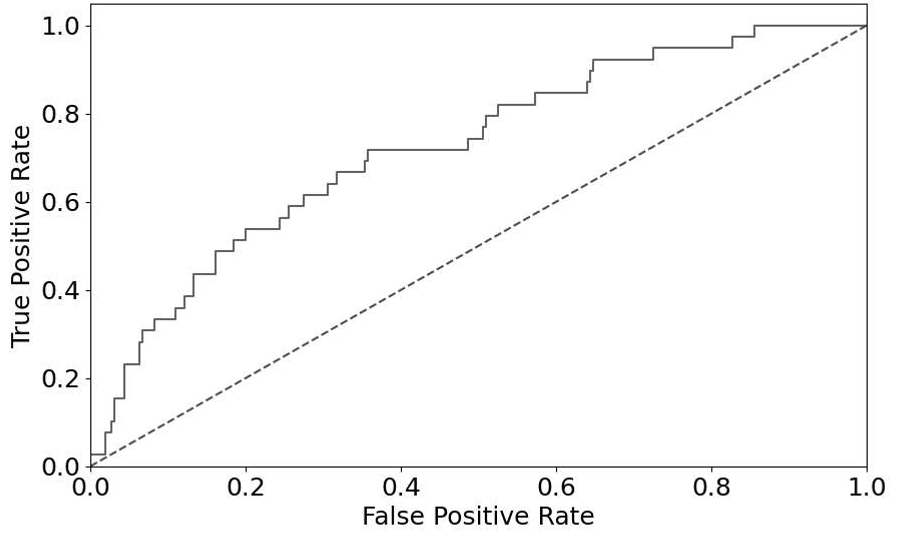


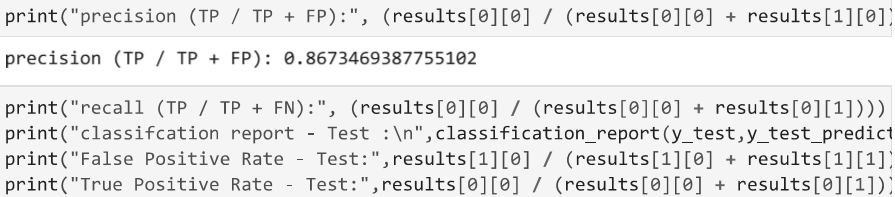


A screenshot of a computer code

Description automatically generated with low confidence





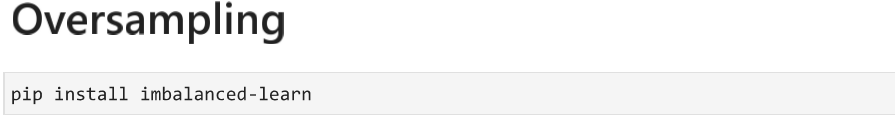


A picture containing text, receipt, font, screenshot

Description automatically generated

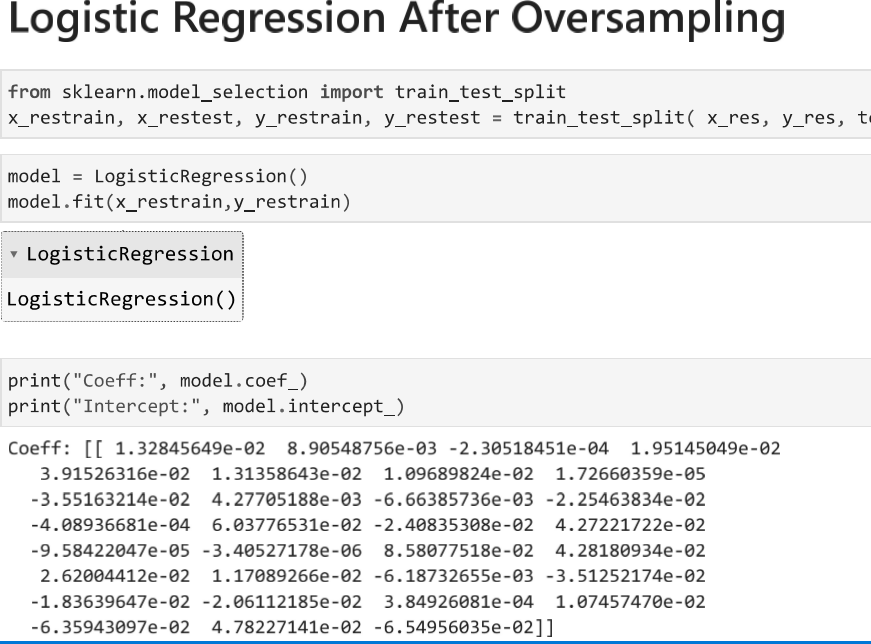
A picture containing text, receipt, font, white

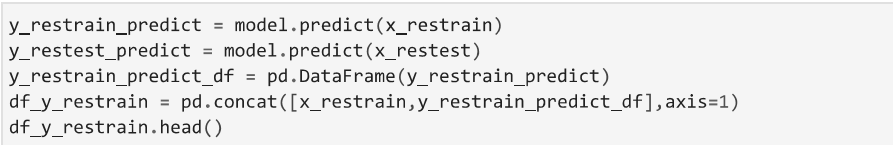
Description automatically generated

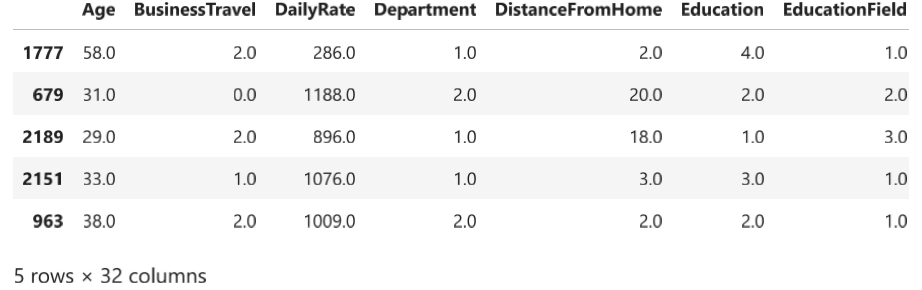


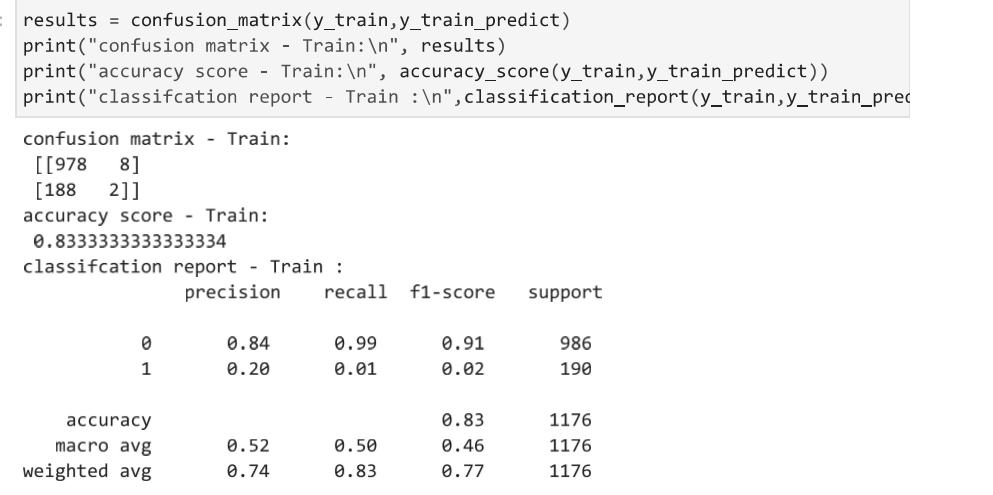
A screenshot of a computer code

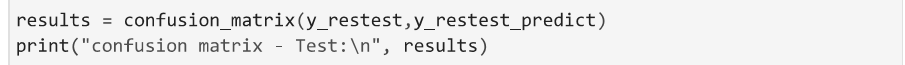
Description automatically generated with low confidence



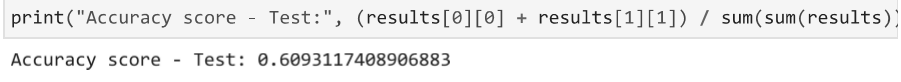






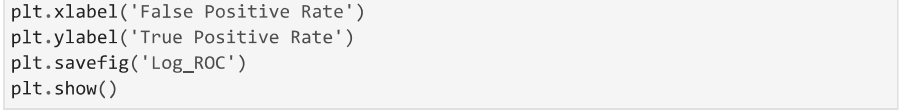






A picture containing text, receipt, white, font

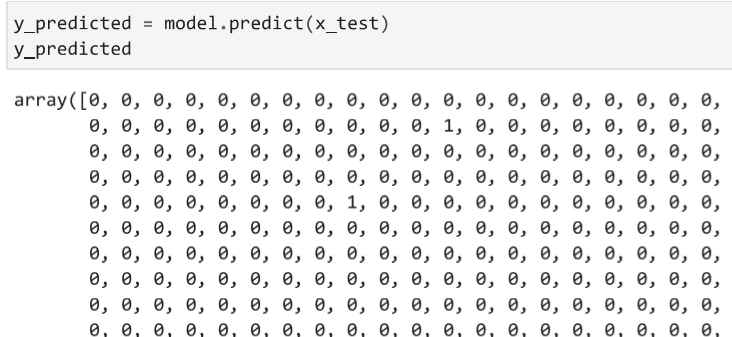
Description automatically generated



A picture containing text, diagram, line, plot

Description automatically generated



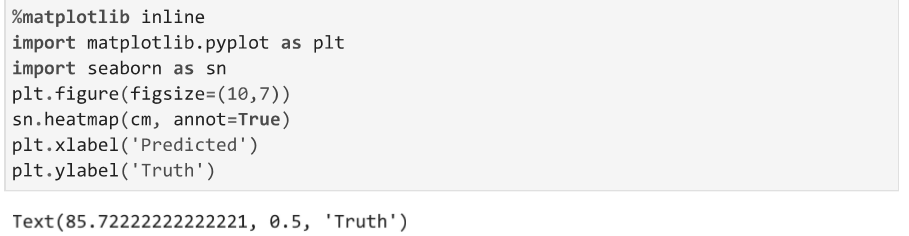


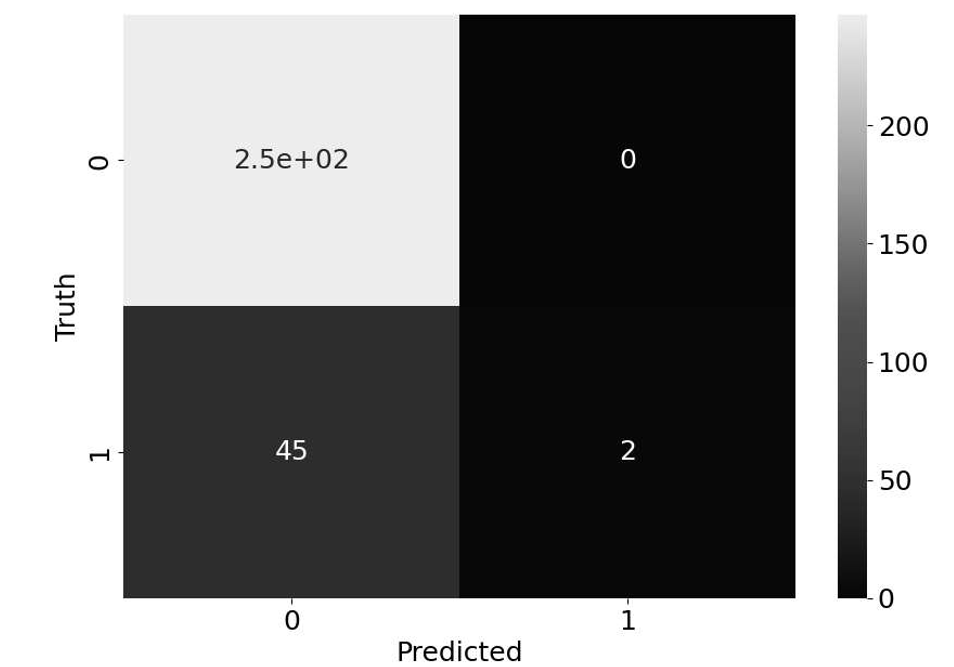
A screenshot of a computer code

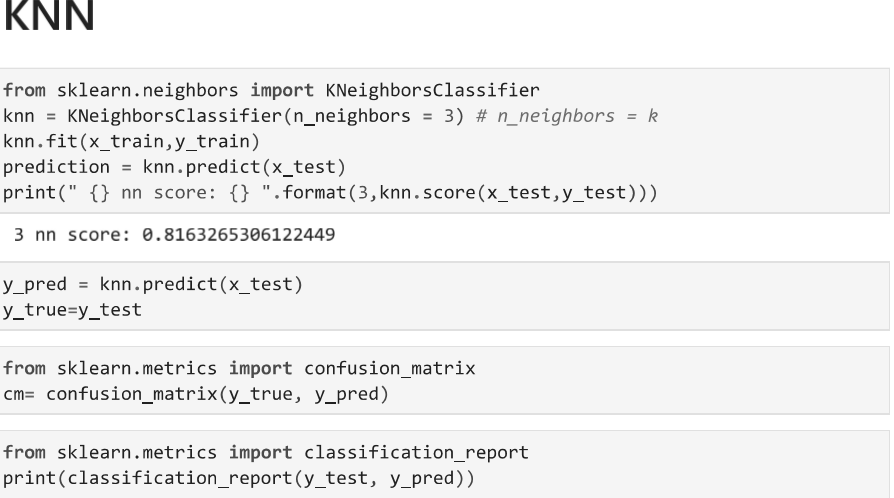
Description automatically generated with low confidence

A picture containing text, receipt, screenshot, font

Description automatically generated

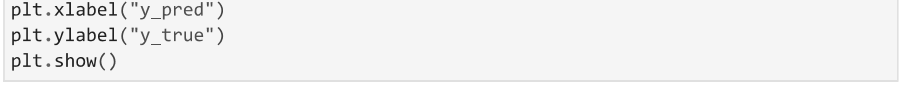






A screenshot of a computer

Description automatically generated with medium confidence



A picture containing text, screenshot, number, design

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

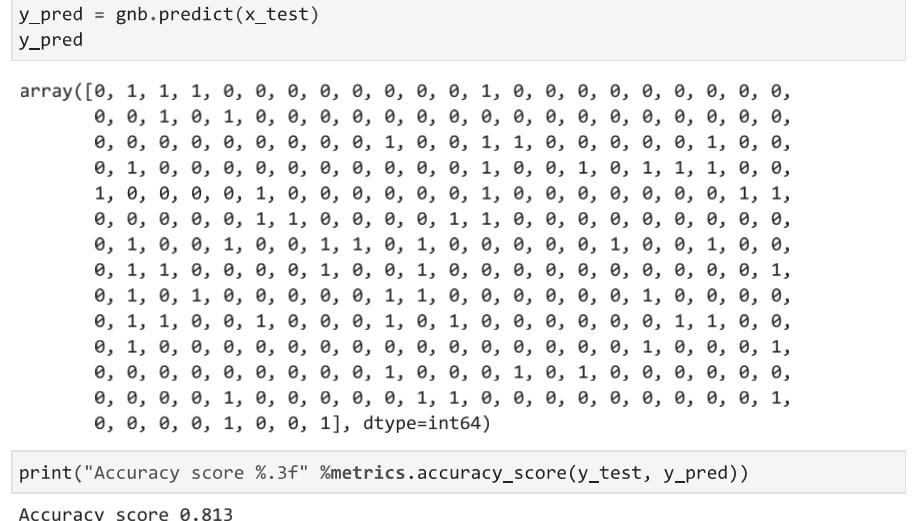
Description automatically generated with low confidence

A screenshot of a computer

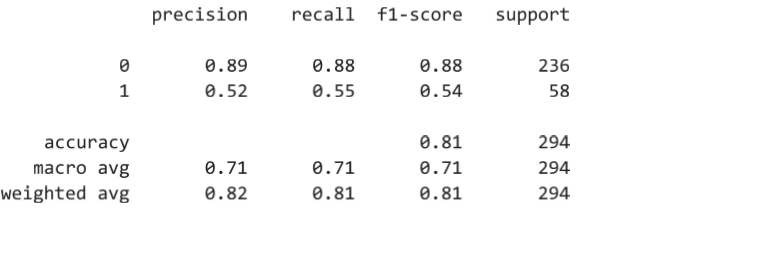
Description automatically generated with medium confidence

A screenshot of a computer

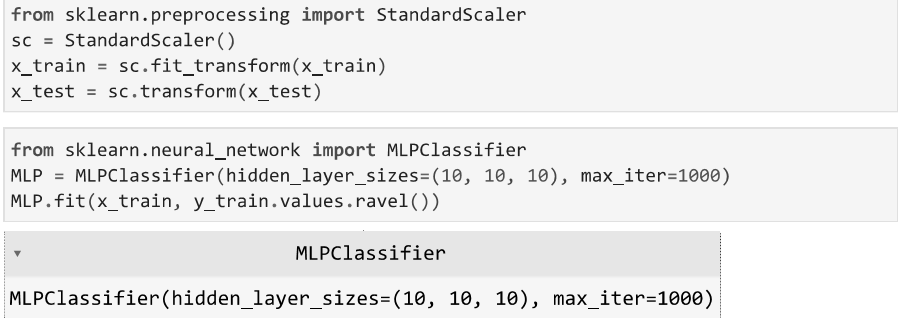
Description automatically generated with low confidence











A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, receipt, screenshot, white

Description automatically generated

A picture containing text, receipt, screenshot, font

Description automatically generated

**Chi- square test of independent of attributes using R software**

>data=read.csv("C:/Users/HP/Desktop/WA\_Fn-UseC\_-HR-Employee-

Attrition.csv")

> head(data)

Age Attrition BusinessTravel DailyRate

1 41 Yes Travel\_Rarely 1102

2 49 No Travel\_Frequently 279

3 37 Yes Travel\_Rarely 1373

4 33 No Travel\_Frequently 1392

5 27 No Travel\_Rarely 591

6 32 No Travel\_Frequently 1005

Department DistanceFromHome Education

1 Sales 1 2

2 Research & Development 8 1

3 Research & Development 2 2

4 Research & Development 3 4

5 Research & Development 2 1

6 Research & Development 2 2

EducationField EmployeeCount EmployeeNumber

1 Life Sciences 1 1

2 Life Sciences 1 2

3 Other 1 4

4 Life Sciences 1 5

5 Medical 1 7

6 Life Sciences 1 8

EnvironmentSatisfaction Gender HourlyRate

1 2 Female 94

2 3 Male 61

3 4 Male 92

4 4 Female 56

5 1 Male 40

6 4 Male 79

JobInvolvement JobLevel JobRole

1 3 2 Sales Executive

2 2 2 Research Scientist

3 2 1 Laboratory Technician

4 3 1 Research Scientist

5 3 1 Laboratory Technician

6 3 1 Laboratory Technician

JobSatisfaction MaritalStatus MonthlyIncome

1 4 Single 5993

2 2 Married 5130

3 3 Single 2090

4 3 Married 2909

5 2 Married 3468

6 4 Single 3068

MonthlyRate NumCompaniesWorked OverTime

1 19479 8 Yes

2 24907 1 No

3 2396 6 Yes

4 23159 1 Yes

5 16632 9 No

6 11864 0 No

PercentSalaryHike PerformanceRating

1 11 3

2 23 4

3 15 3

4 11 3

5 12 3

6 13 3

RelationshipSatisfaction

1 1

2 4

3 2

4 3

5 4

6 3

StockOptionLevel TotalWorkingYears

1 0 8

2 1 10

3 0 7

4 0 8

5 1 6

6 0 8

TrainingTimesLastYear WorkLifeBalance

1 0 1

2 3 3

3 3 3

4 3 3

5 3 3

6 2 2

YearsAtCompany YearsInCurrentRole

1 6 4

2 10 7

3 0 0

4 8 7

5 2 2

6 7 7

YearsSinceLastPromotion YearsWithCurrManager

1 0 5

2 1 7

3 0 0

4 3 0

5 2 2

6 3 6

> table(data$Attrition,data$Gender)

Female Male

No 501 732

Yes 87 150

> chisq.test(data$Attrition,data$Gender,correct=FALSE)

Pearson's Chi-squared test

data: data$Attrition and data$Gender

X-squared = 1.2752, df = 1, p-value = 0.2588

> table(data$Attrition,data$MaritalStatus)

Divorced Married Single

No 294 589 350

Yes 33 84 120

> chisq.test(data$Attrition,data$MaritalStatus,correct=FALSE)

Pearson's Chi-squared test

data: data$Attrition and data$MaritalStatus

X-squared = 46.164, df = 2, p-value = 9.456e-11

> table(data$Attrition,data$BusinessTravel)

Non-Travel Travel\_Frequently Travel\_Rarely

No 138 208 887

Yes 12 69 156

> chisq.test(data$Attrition,data$BusinessTravel,correct=FALSE)

Pearson's Chi-squared test

data: data$Attrition and data$BusinessTravel

X-squared = 24.182, df = 2, p-value =5.609e-06

> table(data$Attrition,data$OverTime)

No Yes

No 944 289

Yes 110 127

> chisq.test(data$Attrition,data$OverTime,correct=FALSE)

Pearson's Chi-squared test

data: data$Attrition and data$OverTime

X-squared = 89.044, df = 1, p-value < 2.2e-16

**Chapter 5**

**References**

* **Books for Reference:**
* Statistical methods and use of R software

Prof. P. G. Dixit, Prof. V. R. Pawgi , Prof. P. S. Kapre

* Fundamentals of mathematical statistics

S. C. Gupta

* Descriptive Statistics

Prof. P. G. Dixit, Prof. Dr. Mrs. V. R. Prayag

* **Websites**

www.google.com

www.youtube.com

[www.tutorialspoint.com](http://www.tutorialspoint.com)

https://towardsdatascience.com

[www.scribbr.com](http://www.scribbr.com)

www.geeksforgeeks.org <https://www.youtube.com/watch?v=xLkk6MUrvrw&ab_channel=5MinutesEngineering>

<https://www.youtube.com/watch?v=c7LrqSxjJQQ&ab_channel=LeilaGharani>

<https://www.youtube.com/watch?v=6cV3OwFrOkk&t=1145s&ab_channel=RishabhMishra>

<https://www.youtube.com/watch?v=bjLIA1vSqGs&ab_channel=LeanExcelSolutions>

https://www.youtube.com/watch?v=Dc3MjUFsEa4&ab\_channel=ExcelconceptuallearningwithM.ArifAslam

* **Other references**

Research articles

* Explaining and predicting employee’s attrition: a machine learning approach (Springer) .
* Employees attrition analysis using predictive techniques Employees attrition prediction using deep neural networks (MDPI).

THANK YOU