A Project Report

On

EMPLOYEE RETENTION USING DATA SCIENCE

Submitted in partial fulfilment of the requirements for the award of the degree in

Bachelor of Technology

In

Department of Electronics and Computer Engineering

Submitted

By

Aitham Suhas **Roll No.18671A1903**

Under the Guidance of

Dr. Roshan Kavuri

ASSOCIATE PROFESSOR

Department of Electronics and Computer Engineering

J.B INSTITUTE OF ENGINEERING AND TECHNOLOGY

(UGC Autonomous), Yenkapally, Moinabad – 500075

2018-2022

EMPLOYEE RETENTION USING DATA SCIENCE

A Dissertation Submitted in the Partial Fulfilment of the Academic Requirements for the Award of the Degree of

Bachelor of Technology

In

Electronics & Computer Engineering

AITHAM SUHAS **18671A1903**

Under the Guidance of

Dr. ROSHAN KAVURI

ASSOCIATE PROFESSOR

Department of Electronics and Computer Engineering

J.B. INSTITUTE OF ENGINEERING AND TECHNOLOGY

(UGC Autonomous & Accredited by NBA& NAAC, Approved by AICTE & Affiliated to JNTU, Hyderabad)

Yenkapally, Moinabad Mandal, R.R. Dist., Hyderabad-50007

2021-2022

J.B. INSTITUTE OF ENGINEERING AND TECHNOLOGY

(UGC Autonomous & Accredited by NBA&NAAC, Approved by AICTE & Affiliated to JNTU, Hyderabad)

CERTIFICATE

This is to certify that the dissertation work entitled **"EMPLOYEE RETENTION USING DATA SCIENCE"** was carried out by **Aitham Suhas** bearing **Roll No.18671A1903** in partial fulfilment of the requirements for the degree of Bachelor of Technology in Electronics and Computers Engineering of the J.B. Institute of Engineering and Technology, Hyderabad, during the academic year 2021-22, is a bonafide record of work carried out under our guidance and supervision.

The results embodied in this report have not been submitted to any other University or Institution for the award of any degree or diploma.

Dr. K. Roshan
Associate Professor
Internal guide

Dr. K. Roshan
Associate Professor
HOD-ECM

External Examiner

ACKNOWLEDGEMENT

There are many people who helped us directly and indirectly to complete our project successfully. We would like to take this opportunity to thank one and all.

First of all, we would like to express our deep gratitude towards our internal guide**,** **Dr. K. Roshan,** Associate Professor**,** Department of Electronics and Computer Engineering for his support in the completion of our dissertation. We wish to express our sincere thanks to **Dr. K. Roshan,** HOD, Department of Electronics and Computer Engineering and also to our Principal **Dr. P. C. Krishnamachary** for providing the facilities to complete the dissertation. We would like to thank the college management for the support in the completion of our dissertation.

We would like to thank all our faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral and financial support and encouragement to achieve goals.

Aitham Suhas **18671A1903**

DECLARATION

I **Aitham Suhas** bearing Roll No.: **18671A1903**, hereby declare that the major project entitled **"Employee Retention using Data Science"** is submitted in the partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Electronics and Computer Engineering from J. B. Institute of Engineering and Technology (UGC Autonomous). The results embodied in this project have not been submitted to any other university or Institution for the award of any degree or diploma.

Aitham Suhas **18671A1903**

PLACE: Hyderabad
DATE:09/02/2022

CONTENTS

ABSTRACT

Now a day's data science predictions are used in IT industries, for the improvement in market investment, employee management etc. Retention of valuable employees within an organization has become an important issue as it is hard to find out the reasons that why employees are leaving an organization and keep them satisfied is a big challenge, for this a report is made to predict the retention of an employee in an organization using the python programming with data science methods. The main idea of this report is to find out that which valuable employee will leave the company and the features which are affecting him/her to making this decision like salary level, no. of hours spending in a week, promotion, no. of work accident etc. The application was developed in python programming language and prediction are made with the help of data science and machine learning models. The design criteria and the implementation details are presented in this report.

1.INTRODUCTION

➢ DEFINITION AND OBJECTIVE OF PROJECT:

Employee retention is the organizational goal of keeping talented employees and reducing turnover by fostering a positive work atmosphere to promote engagement, showing appreciation to employees, and providing competitive pay and benefits and healthy work-life balance. Employers are particularly interested in retaining employees during periods of low unemployment and heightened competition for talent. To retain employees, organizations use human resources technology for recruiting, onboarding, engaging and recognizing workers and offer more work flexibility and modern benefits like physical and financial wellness programs.
Employee retention strategies:

Organizations that are focused on retaining employees usually start with the employee hiring and onboarding process by giving new workers adequate training and orientation in the culture of the organization. They also give new employees an opportunity to ask questions and engage in dialogue with supervisors about their work.

Some organizations use systematic recognition and rewards strategies to show they value employees. Some employers rely on employee engagement software that uses gamification and other techniques to recognize workers and provide rewards and perks such as retail discounts. Employers also focus on competitive pay using employee compensation management software that compares pay rates against benchmarks for given regions, job titles and performance ratings.

Employers seek to distinguish themselves in the hiring arena by offering slates of varied benefits offerings, both voluntary benefits, or employee-paid, and those paid for or subsidized by the organization. Newer types of benefits include lower premium high-deductible health insurance plans, pet insurance, education debt repayment programs and legal counselling.

2.LITERATURE SURVEY

➢ INTRODUCTION:

Data mining is the next big in the world of Information Technology, usage of data extraction is increasing day by day. Data science is the process of mining of useful insights from larger amount of data to use it for the development purpose. To extract data several algorithms, methods and analyzing processes are used depending upon the kind of data we have and what the analyst intended to do with the data. The data we get is in the form of raw data, it needs to get pre-processed to make it in the form to apply algorithm on it. Pre-processing techniques includes collection, noise removal, data reduction, transformation etc. data science methodologies are mainly classified in two categories as making prediction and pattern discovery, prediction making is the process of producing estimate result by analyzing previous results known as regression or supervised learning and pattern discovery is that method when we apply different approaches to find out similarities and dissimilarities in the given data by assigning class notations which is known as clustering or unsupervised learning. Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. A Data Analyst as a rule clarifies what is happening by handling history of the information. Then again, Data Scientist not exclusively does the exploratory investigation to find bits of knowledge from it, yet in addition utilizes different propelled machine learning calculations to recognize the event of a specific occasion later on. A Data Scientist will take a gander at the information from numerous edges, at times edges not known before. Along these lines, Data Science is essentially used to settle on choices and forecasts making utilization of prescient causal examination, prescriptive investigation (prescient in addition to choice science) and machine learning. We know that larger companies contain more than thousand employees working for them, so taking care of the needs and satisfaction of each employee is a challenging task to do, it results in valuable and talented employees leave the company without giving the proper reason. This paper provides solution for the given problem as it gives a prediction model that can be used to predict which employee will leave the company and which will not leave. It also helps in finding the exact reasons which are motivating the employees for shifting companies like lower salary, less promotions or heavy work load etc. To find the result in the form of yes or no, we have used logistic regression method, which predicts result in binary values that are 0 or 1, 0 means employee will not leave the company and 1 means he/she will.

➢ EXISTING SYSTEM:

Retention of valuable employee within an organization is a major issue in the companies, so several efforts are made to find out the proper employee management policies in the

companies, we are discussing some work from them – Piotr Płoński (MLJAR) proposed the analytic methods those can improve Human Resources (HR) management for companies with large number of employees by providing approaches to predict employee attrition with machine learning. They used 1200 employee's data for training datasets, which contains description, but the retention is unknown, which is predicted using binary classification. Le Zhang and Graham Williams proposed that employee retention is the biggest challenge for a company, so it is important for company to recognize behavioural patterns to understand their employees better. They used R for predictions by feature extraction methods as word-to-vector, term frequency, or term frequency and inverse document frequency, R packages such as tm etc. They finally concluded that ensemble techniques can be deployed to effectively boost model performance. Ashish Mishra proposed that it is first important to recruit right person to do talent management, the easily available data source for present and past candidates is their resume. This paper provides a method to calculate the employee score using his educational and business experience scores. They concluded that information like number of years of education, number of organizations worked for, number of positions held in the past, and age can be easily translated into a score for every employee which can be used for predicting retention. Rupesh Khare, Dimple Kaloya and Gauri Gupta proposed that a risk equation can be develop, which can be used assess attrition risk with current set of employees that a company is having. They concluded by stating that among the various attrition predictive techniques available in the market, Logistic Regression and Discriminant Analysis are the closest to give a solution which produced highly accurate results. Randy Lao states that a company which make healthy environment and provide equal opportunities for employees to glow, grows rapidly. Their goal is to create a model that help in improving retention strategies on targeted employees. He used R programming language and, they concluded by saying that employees having higher satisfaction and evaluation rate will have fewer chance to leave the company.

➢ DISADVANTAGES OF EXISTING SYSTEM:

There are various disadvantages of using data science concepts and machine learning algorithms in the existing systems are:

- Efficiency of data training and the prediction of our model.
- The proposed system only used logistic regression for prediction.

➢ PROPOSED SYSTEM:

The proposed system mainly focusses on accuracy, the older model was able to get an accuracy ranging between 75-80. The accuracy mainly depends on data pre-processing and the ML algorithms we used. I used various supervised learning algorithms, such as KNN, MLP classifier, Random forest, Decision Tree Classifier etc. With the help of these various algorithms I was able to achieve an accuracy of over 80 with two of those algorithms. My goal is to create a model that help in improving retention strategies on targeted employees. I used python programming language and, I concluded by saying that employees having higher job satisfaction level, promotion levels and better standards of living will have fewer chance to leave the company.

3.ANALYSIS

➢ INTRODUCTION

Retention of a positive and motivated employee is very important for the organization's success. High employee turnover increases the expenses and also has a negative impact on the organization's morale. Implementation of an employee retention program is an effective way of making sure that the pivotal workers remain employed while balancing and maintaining job performance and productivity.

- Recruitment Enhancement − Effective retention strategies often begin during the employee recruitment process.
- Employee Turnover Management − Employers implement retention strategies to manage employee turnover and attract quality employees.
- Performance and Productivity Maintenance − Employee retention practices help support an organization's productivity.
- Cost Effective − An organization can significantly get benefit from employee retention programs because of a direct effect on an employer's strategies.
- Increases Morale − Employees who enjoy what they do and the atmosphere in which they work are more likely to remain employed with their organization over a longer period of time.

Retaining a Valuable Employee is Essential :

The organization and management should understand the difference between a valuable employee and an employee who does not contribute much to the organization. Sincere efforts must be made to encourage the employees so that they stay happy in the current organization and do not look for a change.

- An organization invests time and money in grooming an individual and make him ready to work and understand the corporate culture.
- An employee, who resigns from the present organization, may join the competitor.
- It is essential for the organization to retain the valuable employees showing potential.
- The employees working for a longer period of time are more familiar with the company's policies, guidelines and thus they adjust better.
- Hiring is not an easy process.
- It has been observed that individuals staying in an organization for a longer time are more loyal towards the management and the organization.

➢ SOFTWARE REQUIREMENT SPECIFICATIONS:

❖ SOFTWARE REQUIREMENTS

- Operating System : Windows 10
- Python IDE :python 3.2.7,pycharm

❖ HARDWARE REQUIREMENTS

- RAM :4GB and Higher
- Processor :i3processor
- Hard Disk :500GB

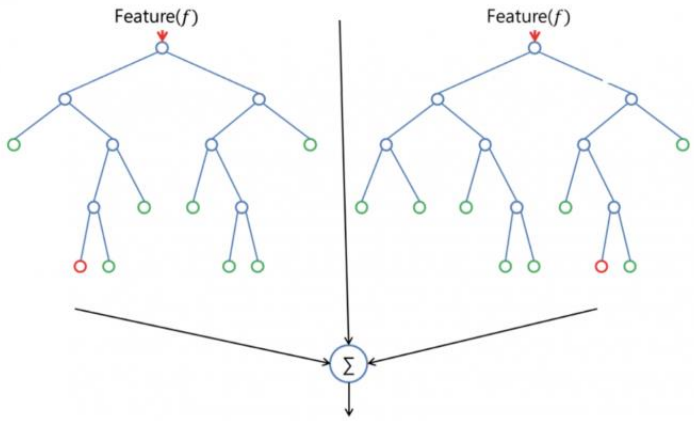
➢ ALGORITHMS AND FLOWCHARTS:

Utilizing this expectation demonstrate, which intends to foresee whether a representative will proceed or leave the association based upon the investigation of the information of past workers. The expectation factors incorporate fulfillment level, last assessment, normal month to month hours, compensation, work mischance, advancement, time spent at the organization and division, in view of these parameters, diverse machine learning models like calculated relapse, choice tree order and so forth are connected to foresee which worker will leave straightaway and the variables that are most huge in this choice. In measurable demonstrating, relapse investigation is an arrangement of factual procedures for assessing the connections among factors. It incorporates numerous systems for displaying and dissecting a few factors, when the emphasis is on the connection between a reliant variable and at least one free factors (or 'indicators'). More particularly, relapse examination causes one to see how the run of the mill estimation of the needy variable (or 'model variable') changes when any of the free factors is fluctuated, while the other autonomous factors are held settled. Most regularly, relapse investigation evaluates the restrictive desire of the needy variable given the autonomous factors – that is, the normal estimation of the reliant variable when the free factors are settled. Less regularly, the attention is on a quantile, or other area parameter of the restrictive conveyance of the reliant variable given the autonomous factors. In all cases, a component of the free factors called the relapse work is to be evaluated. In relapse investigation, it is additionally important to portray the variety of the needy variable around the forecast of the relapse work utilizing a likelihood conveyance. A related however particular approach is Necessary Condition Analysis (NCA), which gauges the most extreme (instead of normal) estimation of the needy variable for a given estimation of the autonomous variable (roof line as opposed to focal line) to recognize what estimation of the free factor is important yet not adequate for a given estimation of the reliant variable. Relapse investigation is broadly utilized for expectation and estimating, where its utilization has considerable cover with the field of machine learning. Relapse examination is likewise used to comprehend which among the autonomous factors are identified with the needy variable, and to investigate the types of these connections. In confined conditions, relapse investigation can be utilized to induce causal connections between the autonomous and ward factors. However, this can prompt figments or false connections, so alert is advisable; for instance, relationship does not demonstrate

causation. Numerous strategies for completing relapse investigation have been created. Well-known techniques, for example, straight relapse and common minimum squares relapse are parametric, in that the relapse work is characterized as far as a limited number of obscure parameters that are evaluated from the information. Nonparametric relapse alludes to strategies that permit the relapse capacity to lie in a predefined set of capacities, which might be endless dimensional. Through this expectation show an organization can choose its arrangements to keep great representatives from leaving the organization. Information science part that utilized as a part of this venture is to take crude information from csv record and then apply distinctive preparing system to settle on information valuable in settling on choices from it like arrangement of dataset, Label Encoding, Onehot Encoding and highlight scaling. Relapse is the most widely recognized technique utilized for making expectation utilizing python programming dialect. Relapse examination likewise enables us to look at the impacts of factors estimated on various scales, for example, the impact of value changes and the quantity of limited time exercises. These advantages help economic specialists/information experts/information researchers to dispose of and assess the best arrangement of factors to be utilized for building prescient models. For this project we used five different algorithms, They are:

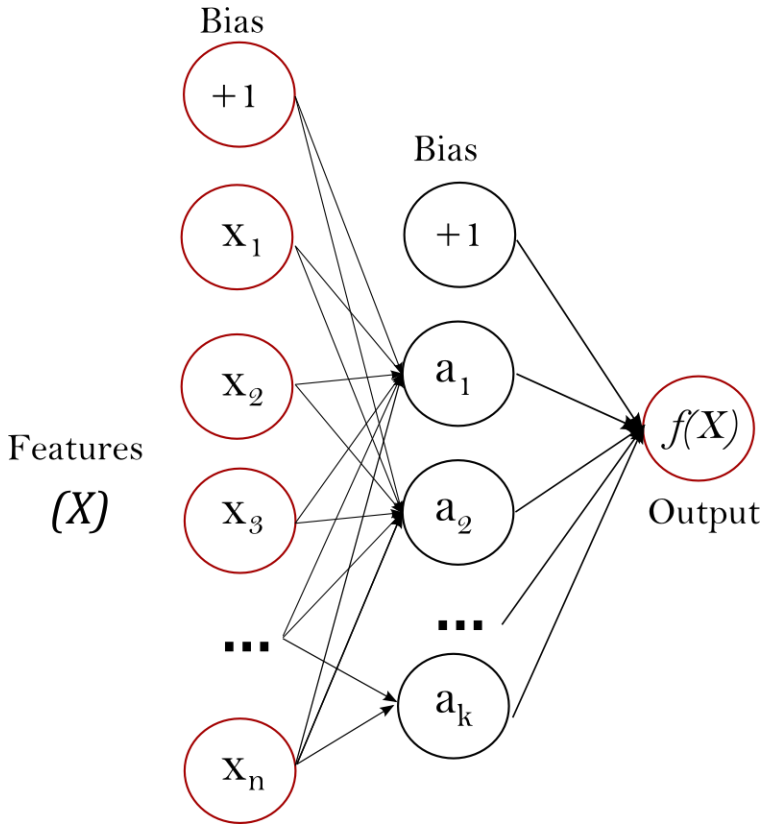
A. Random Forest Classifier:

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.
Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:



B. MLP Classifier:

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function f(·):Rm→Ro by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features X=x1,x2,...,xm and a target y, it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.



The leftmost layer, known as the input layer, consists of a set of neurons {xi|x1,x2,...,xm} representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation w1x1+w2x2+...+wmxm, followed by a non-linear activation function g(·):R→R - like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.

The module contains the public attributes coefs_ and intercepts_. coefs_ is a list of weight matrices, where weight matrix at index i represents the weights between layer i and layer i+1. intercepts_ is a list of bias vectors, where the vector at index i represents the bias values added to layer i+1.

The advantages of Multi-layer Perceptron are:

- Capability to learn non-linear models.
- Capability to learn models in real-time (on-line learning) using partial_fit.

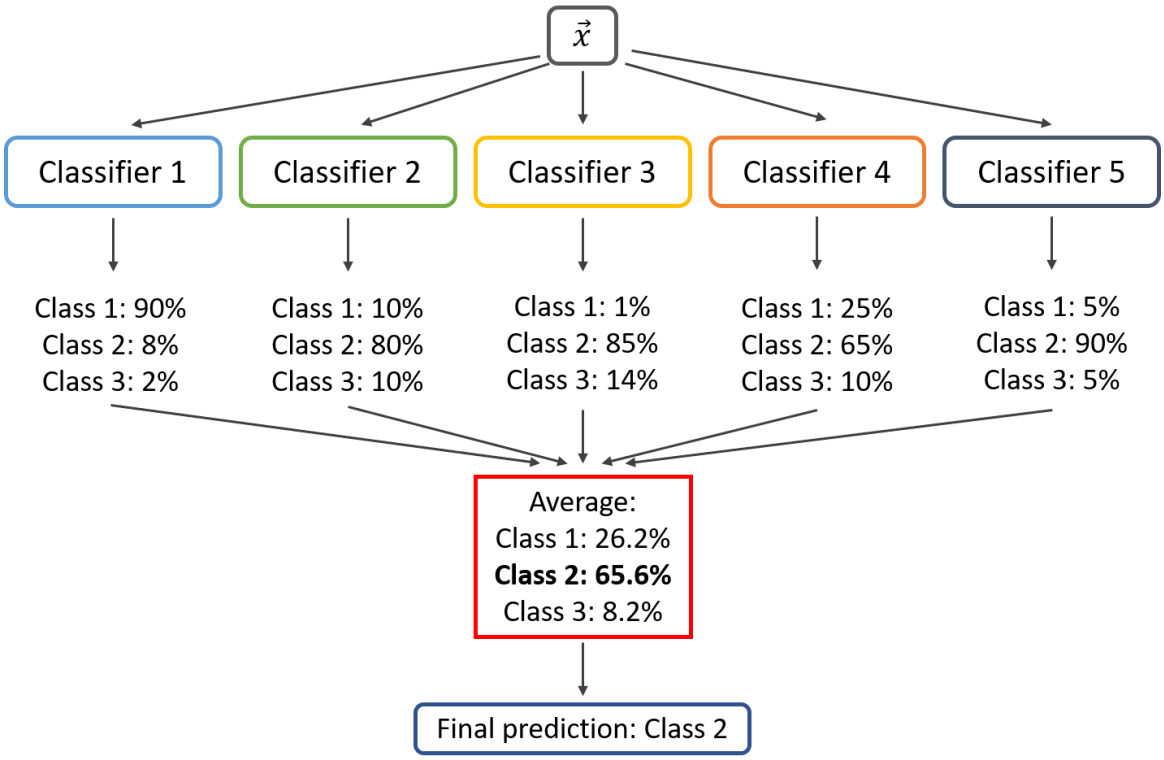
C. Voting Classifier:

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.
It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

Voting Classifier supports two types of votings.

Hard Voting: In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the output class(A, A, B), so here the majority predicted A as output. Hence A will be the final prediction.

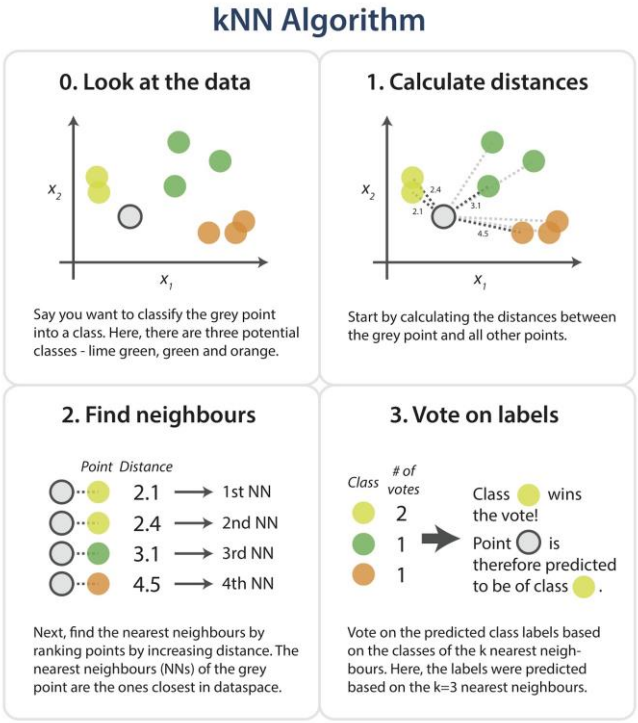
Soft Voting: In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.



D. K-Nearest Neighbor Algorithm:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique .K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm..K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



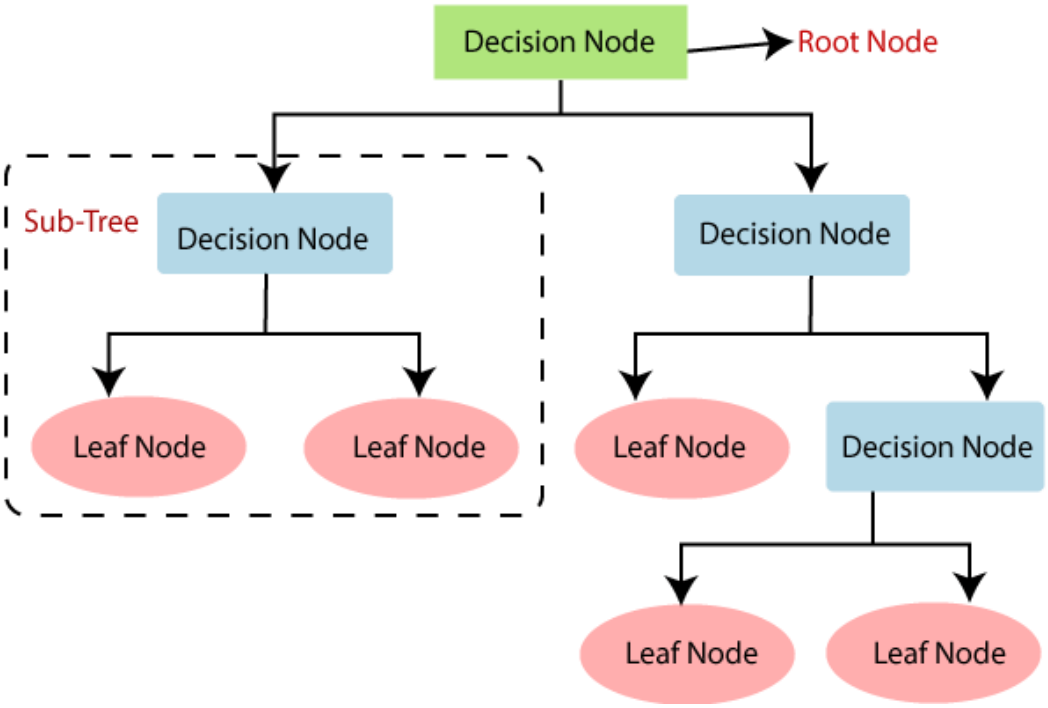
E. Decision Tree Algorithm:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.

- Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.



➢ CONCLUSION:

Machine learning is the process of making the machine to learn itself through patterns and training data sets. Training data sets are data which is given to machine for understanding the hidden patterns within data and make relations for own understanding. It helps in working of machines efficiently by making them processed like a human brain. Pattern recognition is the most challenging task for developers to use such algorithms that allows different machines to work according to the requirement. This paper emphasizes on making prediction of retention of an employee within an organization such that whether the employee will leave the company or continue with it. It uses the data of previous employees which have worked for the company and by finding pattern it predicts the retention in the form of yes or no. It uses various parameters of employees such as salary, number of years spent in the company, promotions, number of hours, work accident, financial background etc. Considering new processing innovations, machine adapting today isn't care for machine learning of the past. It was conceived from design acknowledgment and the hypothesis that PCs can learn without being customized to perform assignments; specialists intrigued by manmade brainpower needed to check whether PCs could gain from information. The iterative part of machine learning is essential claiming as models are presented to new information, they can freely adjust. They gain from past calculations to deliver solid, repeatable choices and results. It's a science that is not new – but rather one that is increasing crisp energy. While numerous machine learning calculations have been around for quite a while, the capacity to

naturally apply complex scientific computations to huge information again and again, quicker and speedier is a current advancement.

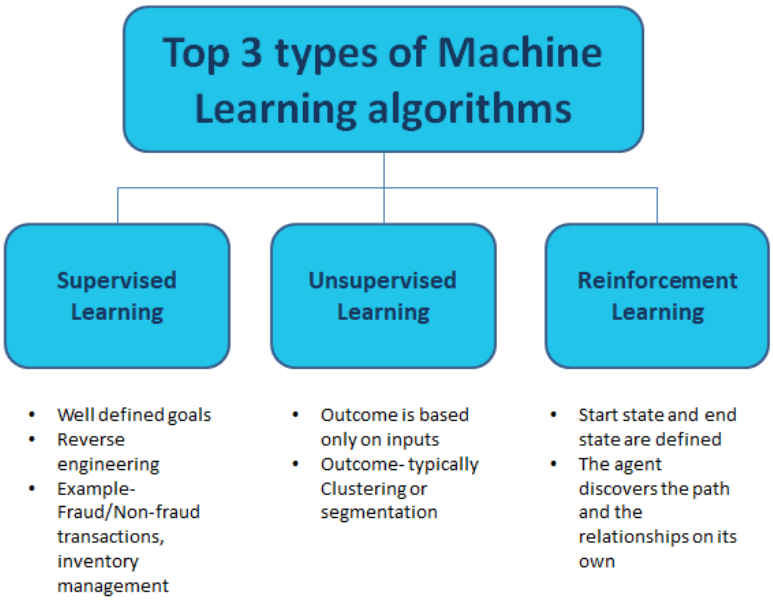


Fig. 1. Types of Machine Learning

Machine learning algorithms are differentiated as supervised or unsupervised. A. Supervised machine learning calculations can apply what has been realized in the past to new information utilizing marked cases to anticipate future occasions. Beginning from the examination of a known preparing dataset, the learning calculation creates a surmised capacity to make expectations about the yield esteems. The framework can give focuses to any new contribution after adequate preparing. The learning calculation can likewise contrast its yield and the right, planned yield and discover mistakes to adjust the model appropriately. B. In differentiate, unsupervised machine learning calculations are utilized when the data used to prepare is neither grouped nor named. Unsupervised learning contemplates how frameworks can induce a capacity to portray a concealed structure from unlabelled information. The framework doesn't make sense of the correct yield; however, it investigates the information and can attract derivations from datasets to depict concealed structures from unlabelled information. C. Semi-directed machine learning calculations fall some place in the middle of regulated and unsupervised learning, since they utilize both marked and unlabelled information for preparing – ordinarily a little measure of named information and a lot of unlabelled information. The frameworks that utilization this strategy can significantly enhance learning precision. For the most part, semi-administered learning is picked when the procured named information requires gifted and significant assets to prepare it/gain from it. Something else, obtaining unlabelled information by and large doesn't require extra assets. D. Reinforcement machine learning calculations is a learning technique that interfaces with its condition by creating activities and finds mistakes or rewards. Experimentation seek and postponed compensate are the most pertinent attributes of fortification learning. This technique enables machines and programming operators to naturally decide the perfect conduct inside a setting to augment its execution. Basic reward input is required for the specialist to realize which activity is ideal; this is known as the support f

4.DESIGN

➢ INTRODUCTION:

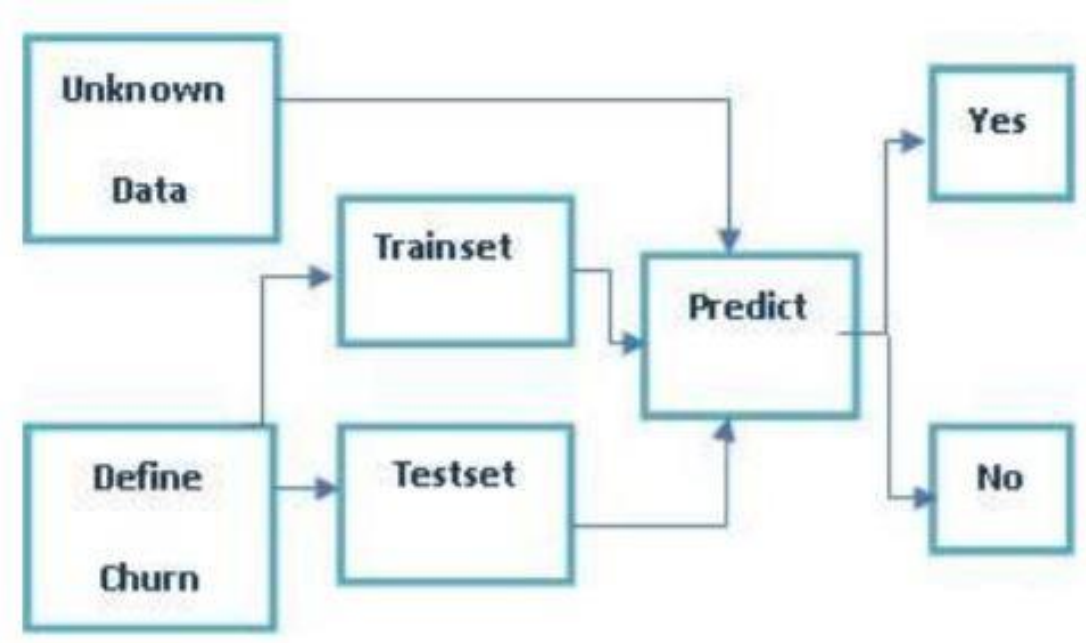


Fig. 2. Prediction Methodology

We have utilized Python programming dialect, which is a translated, progressively written dialect and least difficult in grammar. Python is utilized for every one of the applications like in IOT advancement, information science field, web improvement, scripting reason and so forth. Consequently, now it is being utilized generally over the globe. Python contains various number of libraries accessible in it, this makes it simple to use for each application like for web rejecting delightful cleanser, for GUI improvement TKinter, for web network urlib2, for machine learning sklearn, numpy, pandas and so on. Python is one of the for the most part utilized dialect for Data Science applications since it gives libraries, for example, Pandas, nltk which can oversee substantial number of datasets into fitting way, it gives representation libraries like Matplotlib, Bokeh, Seaborn and so on that are exceedingly expressive regarding charts and plots portrayals. The sklearn library is one which gives bigger number of machine learning calculations, for example, direct and various relapse, polynomial relapse, choice tree characterization and so on., to make expectations, bunching and grouping of information in number of billions Machine learning is a branch in software engineering that reviews the outline of calculations that can learn. Run of the mill errands are idea learning, work learning or "prescient demonstrating", bunching and finding prescient examples. These undertakings are found out through accessible information that were seen through encounters or directions, for instance. The expectation that accompanies this teach is that including the experience into its assignments will in the end enhance the learning. However, this change needs to occur such that the learning itself ends up programmed with the goal that people like ourselves don't have to meddle any longer is a definitive objective. Scikit-learn is the most helpful

library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a great deal of efficient devices for machine learning and factual displaying including arrangement, relapse, bunching and dimensionality lessening. Scikit-learn gives a scope of directed and unsupervised learning calculations through a reliable interface in Python. It is authorized under a lenient disentangled BSD permit and is circulated under numerous Linux appropriations, empowering scholastic and business utilize.

➢ DATA PRE-PROCESSING, TEST AND TRAIN:

In straightforward words, pre-preparing et.al [9] alludes to the changes connected to the information before nourishing it to the calculation. In python, scikit-learn library has a pre-assembled usefulness under sklearn. pre-processing. The information we get from client is as crude information, so it needs to get perfect, change and decrease to make it proper for applying strategies on it, this procedure is known as pre-processing. require scientific sandbox in which you can perform examination for the whole term of the task. You have to investigate, pre-process and condition information preceding demonstrating. Further, you will perform ETLT (remove, change, stack and change) to get information into the sandbox. It enhances the general nature of the information and effectiveness of the model to deliver comes about. There are numerous more alternatives for pre-preparing as –

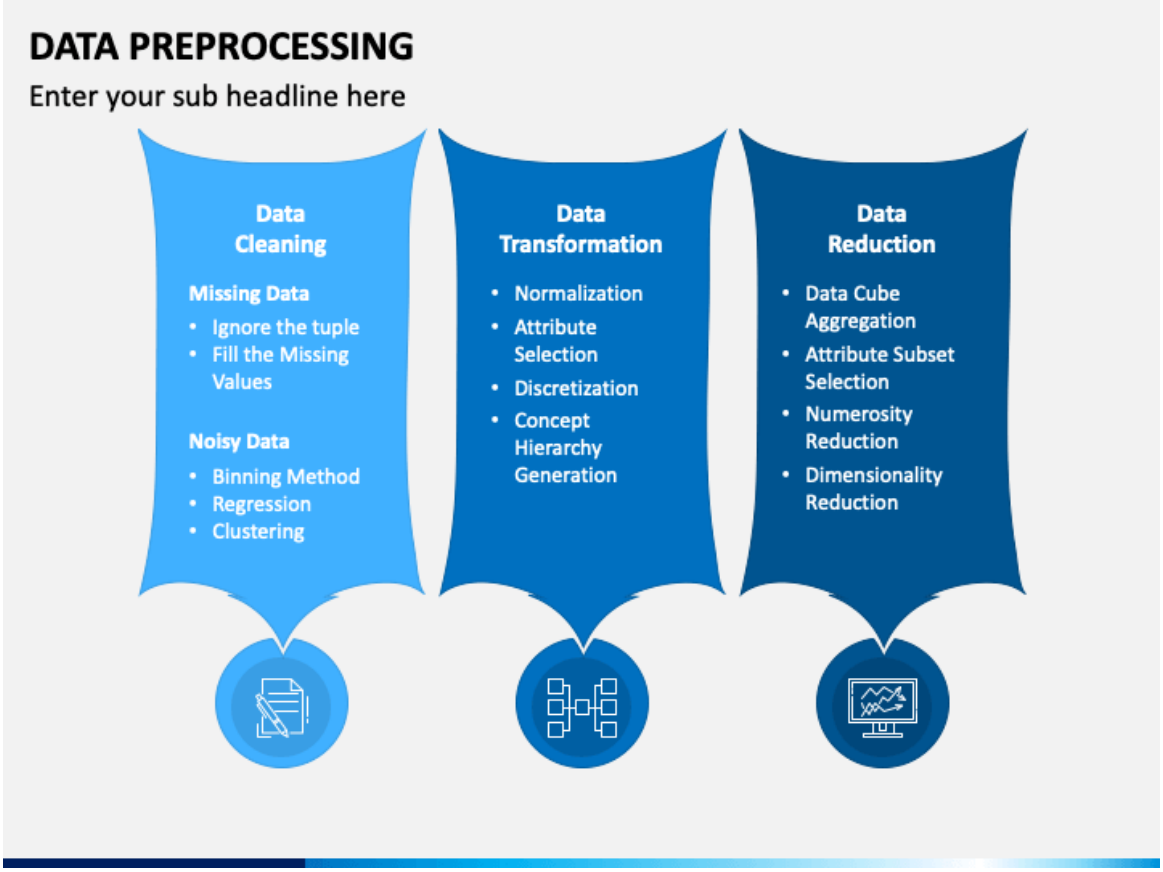


Fig. 3. Pre-processing Techniques

A. Feature Scaling: Highlight scaling is the strategy to restrict the scope of factors with the goal that they can be thought about on basic grounds. It is performed on constant factors.

B. Label Encoding: Sklearn gives an extremely proficient device to encoding the levels of an all-out highlights into numeric esteems. Name Encoder encode names with an incentive about 0 and classes

C. One-Hot Encoding: One-Hot Encoding changes each clear-cut component with n conceivable esteems into n parallel highlights, with just a single dynamic. Most of the ML calculations either take in a solitary weight for each component or it figures remove between the examples.

Data Cleaning:

Data Cleaning is particularly done as part of data preprocessing to clean the data by filling missing values, smoothing the noisy data, resolving the inconsistency, and removing outliers.

1. Missing values

Here are a few ways to solve this issue:

- Ignore those tuples

This method should be considered when the dataset is huge and numerous missing values are present within a tuple.

- Fill in the missing values

There are many methods to achieve this, such as filling in the values manually, predicting the missing values using regression method, or numerical methods like attribute mean.

2. Noisy Data

It involves removing a random error or variance in a measured variable. It can be done with the help of the following techniques:

- Binning

It is the technique that works on sorted data values to smoothen any noise present in it. The data is divided into equal-sized bins, and each bin/bucket is dealt with independently. All data in a segment can be replaced by its mean, median or boundary values.

- Regression

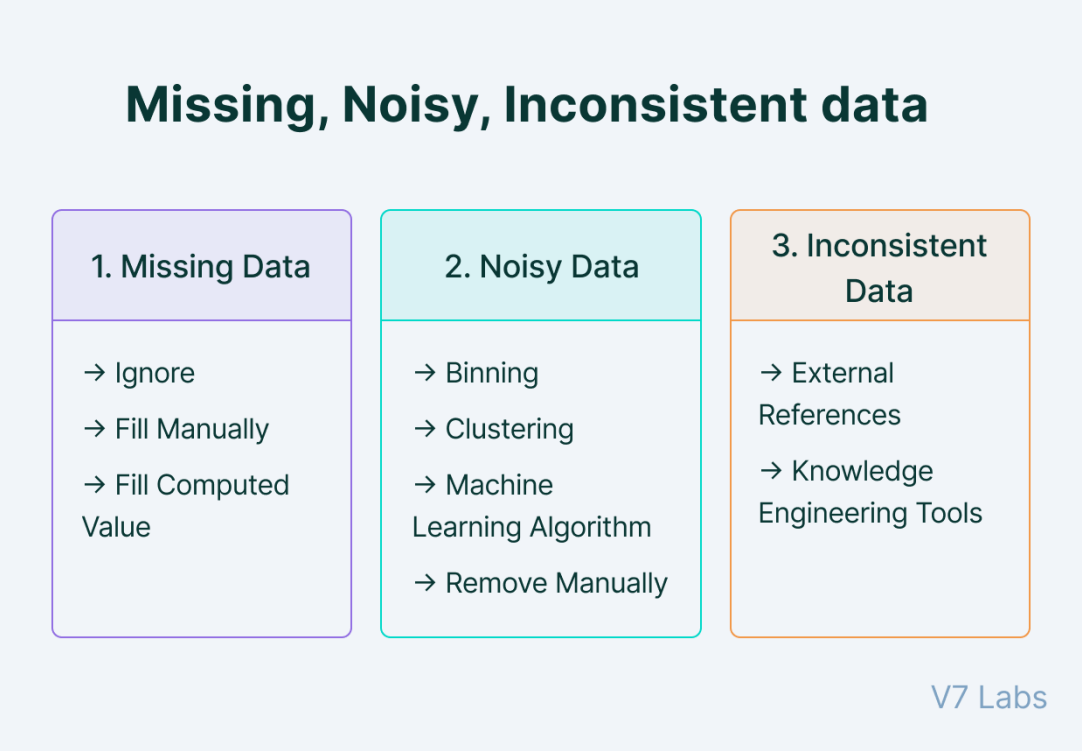
This data mining technique is generally used for prediction. It helps to smoothen noise by fitting all the data points in a regression function. The linear regression equation is used if there is only one independent attribute; else Polynomial equations are used.

- Clustering

Creation of groups/clusters from data having similar values. The values that don't lie in the cluster can be treated as noisy data and can be removed.

3. Removing outliers

Clustering techniques group together similar data points. The tuples that lie outside the cluster are outliers/inconsistent data.



5.IMPLEMENTATION AND RESULTS

➢ Input Dataset:

- Source: Kaggle
- https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

➢ Output:

Model Comparison

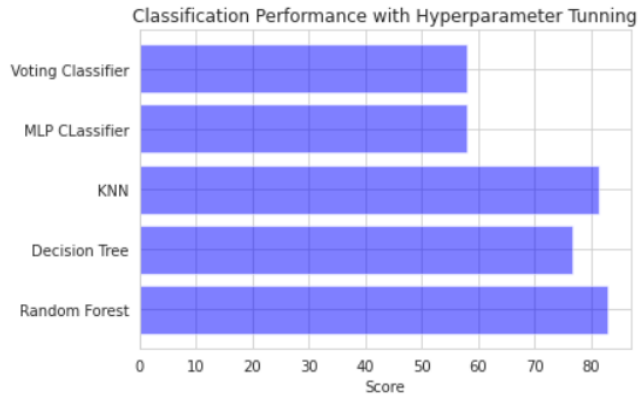
Accuracy

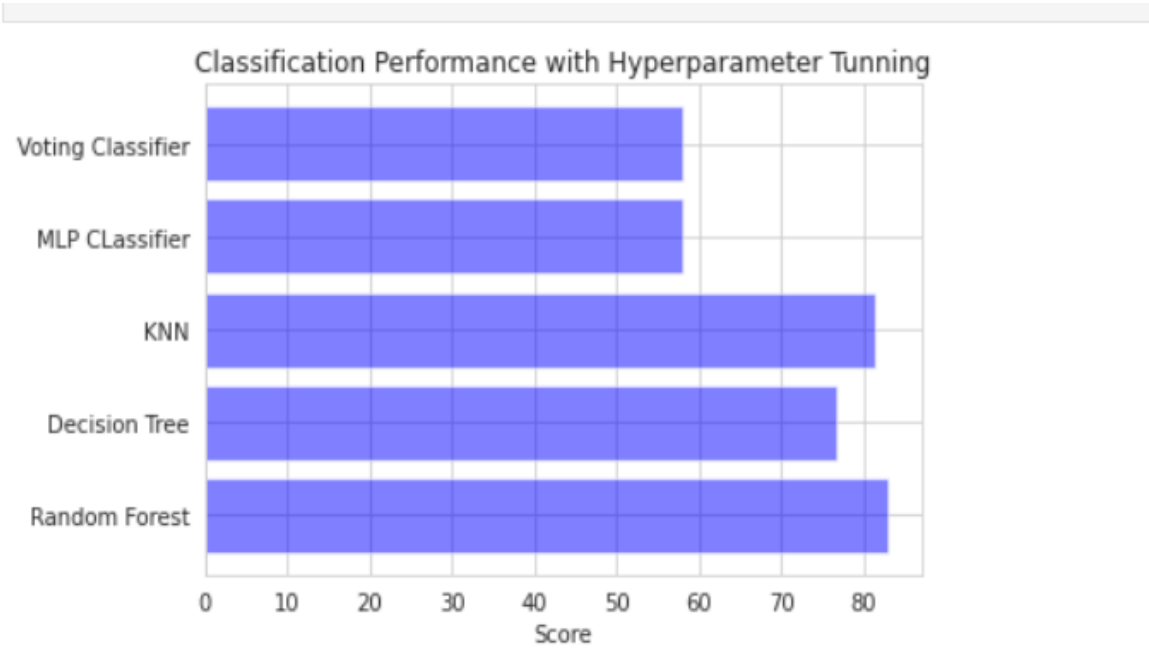
In [108…
```
score = [val1,val2,val3,val4,val5]
```

In [109…
```
#make variabel for save the result and to show it
classifier = ('Random Forest','Decision Tree','KNN','MLP CLassifier','Voting Classifier')
y_pos = np.arange(len(classifier))
print(y_pos)
print(score)
```

```
[0 1 2 3 4]
[82.88043478260869, 76.63043478260869, 81.25, 57.88043478260869, 57.88043478260869]
```

In [110…
```
import matplotlib.pyplot as plt2
plt2.barh(y_pos, score, align='center', alpha=0.5,color='blue')
plt2.yticks(y_pos, classifier)
plt2.xlabel('Score')
plt2.title('Classification Performance with Hyperparameter Tunning')
plt2.show()
```





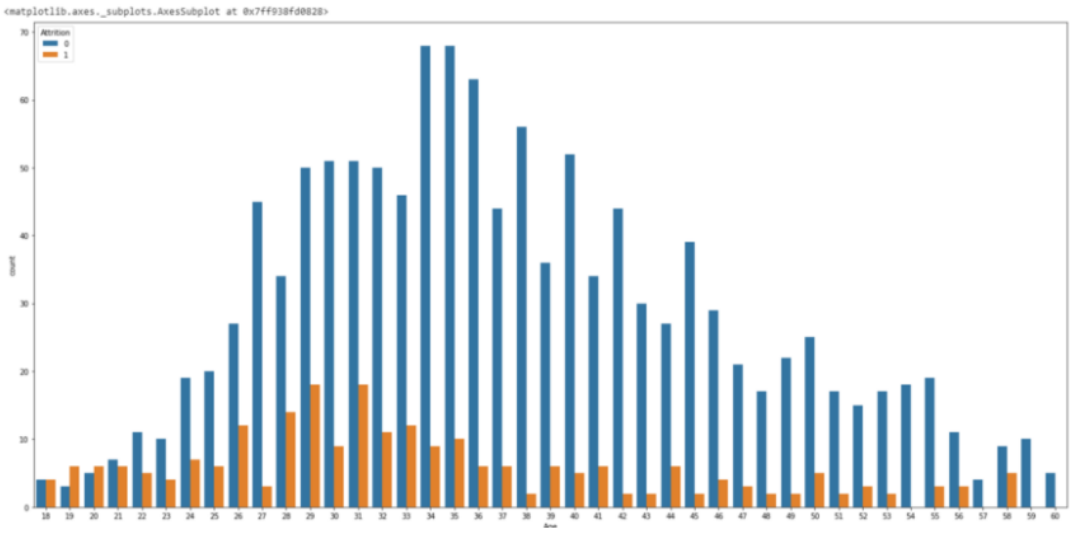
➢ PARAMETER COMPARISON:

Data Processing

In [3]:
```
data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
Age                         1470 non-null int64
Attrition                   1470 non-null object
BusinessTravel              1470 non-null object
DailyRate                   1470 non-null int64
Department                  1470 non-null object
DistanceFromHome            1470 non-null int64
Education                   1470 non-null int64
EducationField              1470 non-null object
EmployeeCount               1470 non-null int64
EmployeeNumber              1470 non-null int64
EnvironmentSatisfaction     1470 non-null int64
Gender                      1470 non-null object
HourlyRate                  1470 non-null int64
JobInvolvement              1470 non-null int64
JobLevel                    1470 non-null int64
JobRole                     1470 non-null object
JobSatisfaction             1470 non-null int64
MaritalStatus               1470 non-null object
MonthlyIncome               1470 non-null int64
MonthlyRate                 1470 non-null int64
NumCompaniesWorked          1470 non-null int64
Over18                      1470 non-null object
OverTime                    1470 non-null object
PercentSalaryHike           1470 non-null int64
PerformanceRating           1470 non-null int64
RelationshipSatisfaction    1470 non-null int64
StandardHours               1470 non-null int64
StockOptionLevel            1470 non-null int64
TotalWorkingYears           1470 non-null int64
TrainingTimesLastYear       1470 non-null int64
WorkLifeBalance             1470 non-null int64
YearsAtCompany              1470 non-null int64
YearsInCurrentRole          1470 non-null int64
YearsSinceLastPromotion     1470 non-null int64
YearsWithCurrManager        1470 non-null int64
dtypes: int64(26), object(9)
memory usage: 402.0+ KB
```

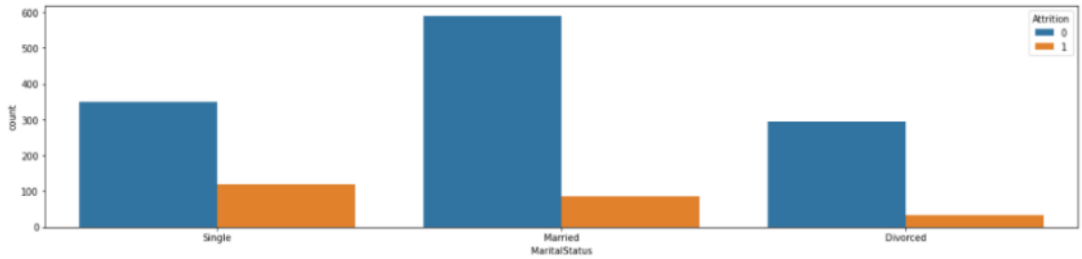
1. Age vs Attrition Analysis



Age vs Attrition (Image by Author)

People of age of 29 and 31 years left the company more frequently. Although the number of employees in age group of 18 to 23 is less but the attrition rate is also high in this group. Also, as age increases the chances of leaving the company decreases.

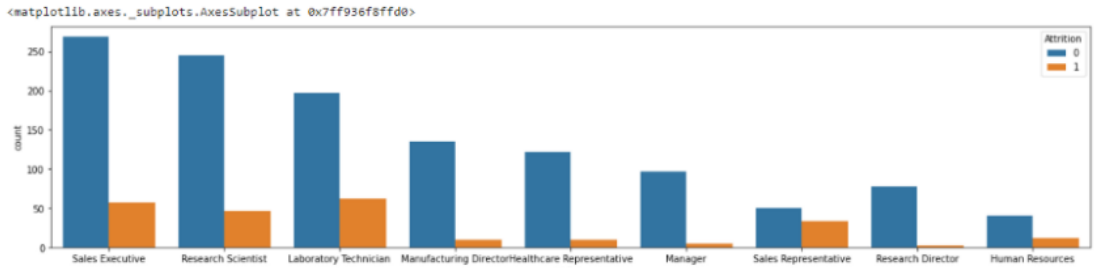
2. Marital Status vs Attrition



Marital Status vs Attrition (Image by Author)

Single employees tend to leave compared to married and divorced

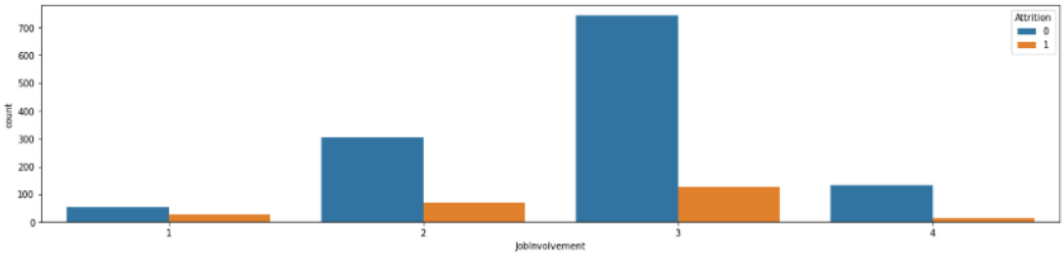
3. Job Role vs Attrition



Job Role vs Attrition (Image by Author)

Sales Representatives and Lab Technician tend to leave compared to any other job.

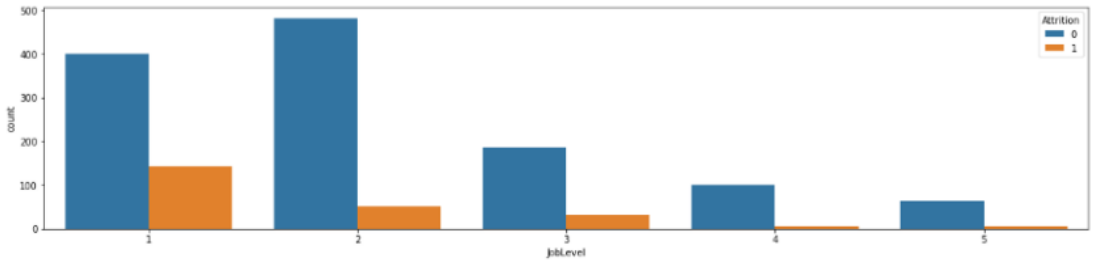
4. Job Involvement vs Attrition



Job Involvement vs Attrition (Image by Author)

Less involved employees tend to leave the company. If you notice Jobinvolvement = 1, it has more attrition as compared to total population under this category.

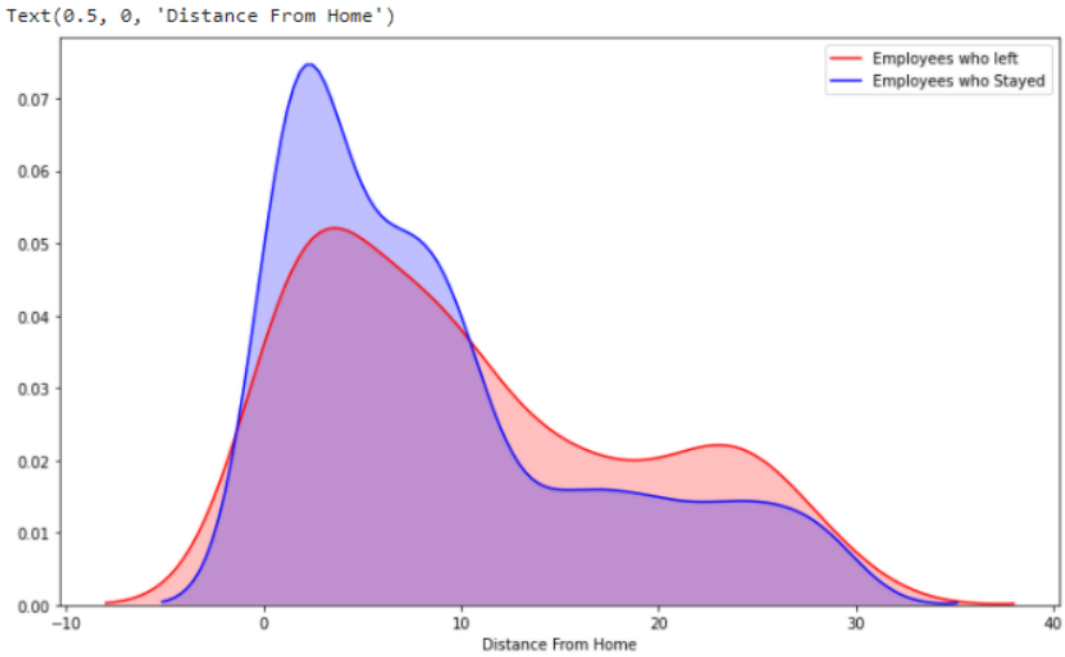
5. Experienced vs Attrition



Experienced vs Attrition (Image by Author)

Less experienced (low job level i.e JobLevel =1) tend to leave the company.

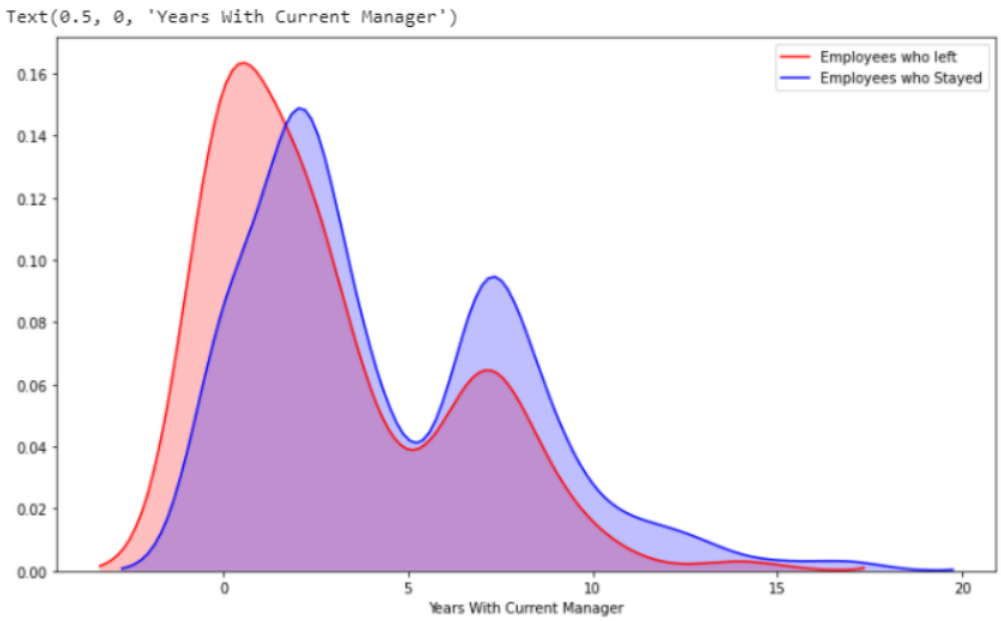
6. Distance from Home vs Attrition



Experienced vs Attrition (Image by Author)

People staying far (more than 10km) from office more likely to leave company. You can notice the red line is above blue line after 10 in the x-axis i.e Distance from Home.

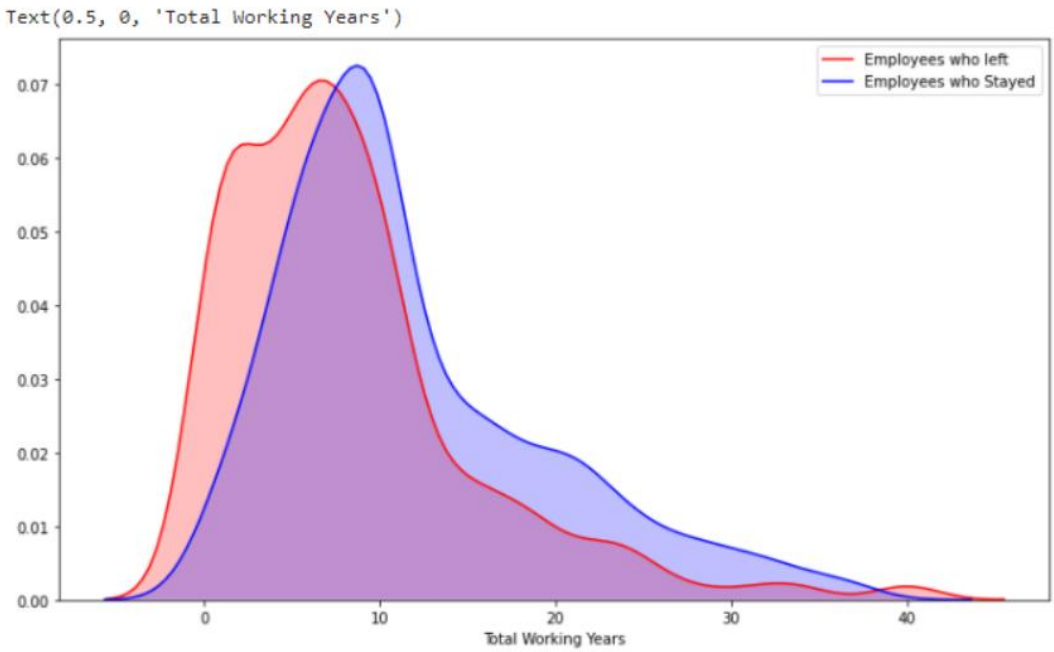
7. Years with Current manager vs Attrition



Years with Current manager vs Attrition (Image by Author)

Employee with small span of time with Current manager are more likely to leave the company. You can notice the red line is above blue line at the starting of x-axis i.e Years with Current manager. However as we increase the number of years, the blue line tends to supersede the red line, which means that as you go beyond 4 to 15 years, the number of employees who actually tend to stay is more than the number of employees who actually leaves the company.

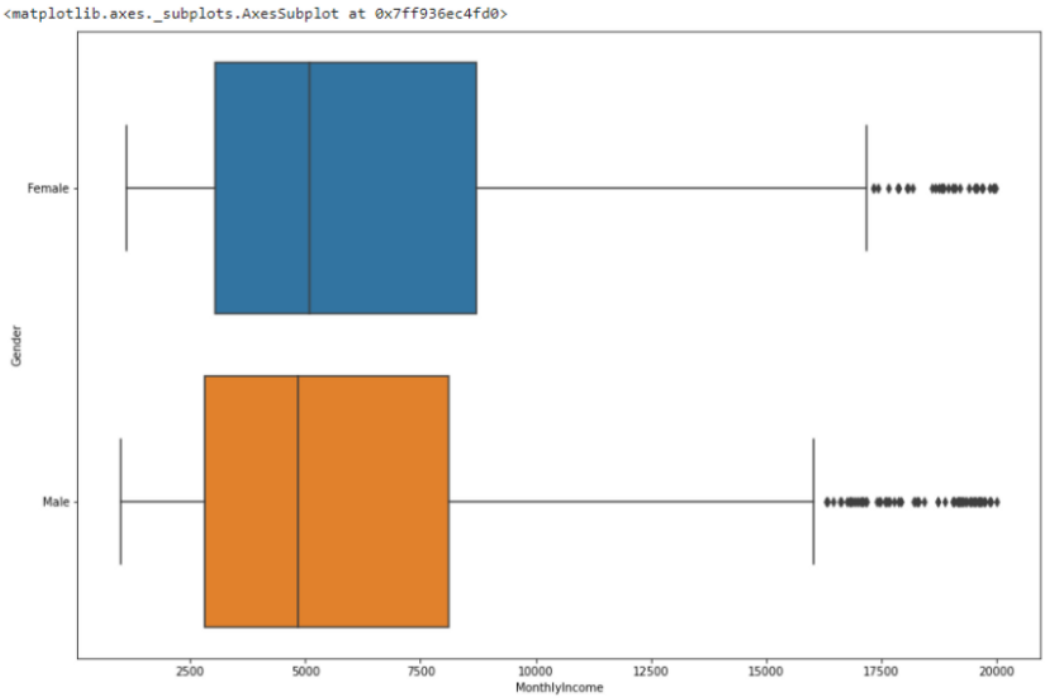
8. Total Working Years vs Attrition



Total Working Years vs Attrition (Image by Author)

Employees with less number of years(0 to 6 years) with the company tend to leave the company. You can notice the red line is above blue line at the starting of x-axis i.e Total Working Years. However as you go beyond 6 years, you will find that the blue line tend to supersede which means the employees tend to stay as you increase the total working years.

9. Gender vs. Monthly Income

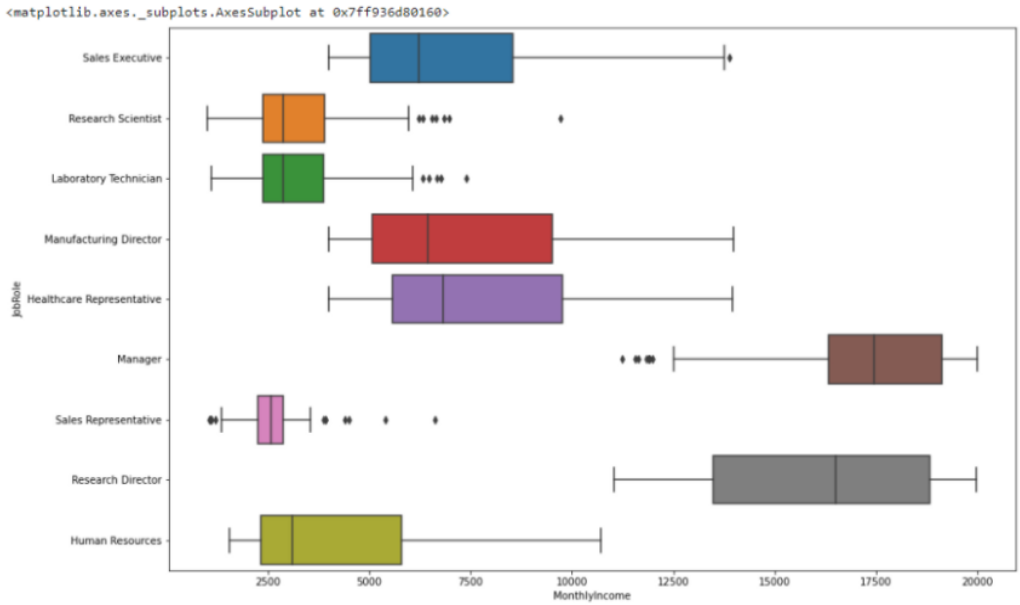


Gender vs. Monthly Income (Image by Author)

You can see here the average salary is almost quite comparable between male and female, that's actually a great thing.

Gender pay equality is actually critical and very important thing for any company. Actually, by looking into box-plot it seems that females actually get paid even more than the males here.

10. Monthly income vs. job role



monthly income vs. job role (Image by Author)

The above box-plot shows that the employees who work as managers tend to get a lot more which obviously makes sense. And it seemed that if you work as Research scientist and Lab technician, the salary range is almost same for both job role.

If you're doing Sales Representative work, you get paid a lot less compared to the Managers and the Research Directors. The human resources are somewhere in between here as well.

6.CONCLUSION

- You can see how data analysis using different charts and visualizations helped in getting answers to many questions. We used five various ML algorithms to predict the accuracy of our model. Out of the five algorithms we used KNN and Random Forest Classifier has achieved greater accuracies than other algorithms. These Algorithms help IT firms to concentrate on particular aspects and these are trustworthy because of the accuracies they provide.

- This analysis will help the company to make some employee policies and modify them if required. Even this will help to make some new employee engagement program that help employee stay more longer.

7.REFERENCES

Textbooks:

Programming Python, Mark Lutz
Head First Python, Paul Barry
Core Python Programming,R. Nageswara Rao
Learning with Python, Allen B. Downey

Journals:

[1] Piotr Płoński (MLJAR), "Human-first Machine Learning Platform," Human Resource Analytics Predict Employee Attrition.
[2] Le Zhang and Graham Williams (Data Scientist, Microsoft), "Employee Retention with R based Data Science Accelarator".
[3] Ashish Mishra (Data Scientist, Experfy), "Using Machine Learning to Predict and explain Employee Attrition".
[4] Rupesh Khare, Dimple Kaloya and Gauri Gupta, "Employee Attrition Risk Assessment using Logistic Regression Analysis," from 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence.
[5] Randy Lao, "Predicting Employee Kernelover," Kaggle.

Websites:

https://www.w3schools.com/python/
https://www.tutorialspoint.com/python/index.htm
https://www.javatpoint.com/python-tutorial
https://www.learnpython.org/
https://www.pythontutorial.net/