Shape, arrow

Description automatically generated

Technical Coding Research Innovation, Navi Mumbai,

Maharashtra, India-410206

**(HR Employee Attrition)**

A Case-Study Submitted for the requirement of

**Technical Coding Research Innovation**

For the Internship Project work done during

**DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM**

by

Ajay Gunaji Shinde(TCRIB2R48)

Rutuja Doiphode CO-FOUNDER &CEO

TCR innovation.

# DATA SCIENCE PROJECT ON HR\_ATTRITION

**DATA SET USING MACHINE LEARNING AND PYTHON**

## INTRODUCTION

The “HR EMPLOYEE ATTRITION DATASET” consists of

the details of an employee like gender, age, business travel, department, education, relationship satisfaction, and many others. Basically, the dataset consists of exactly 2940 employees' data, and employee has 34 features. The dataset consists of both numerical and categorical data.

The dataset is having no null value so we don’t need to deal with null values .The size of dataset is 102900.

Our final project is based on the machine learning in which we have provided with a data set of an organization from which we have to predict, whether a employee will leave the organization or not. We have used different classification algorithm such as random forest, logistic regression and knn to derive the best results .our project has the capability to predict the targeted variable with around 95.46% accuracy by using random forest algorithm, this is the best result we have got among all algorithm.

IDE

For the final project we have used jupyter notebook ide:- Jupyter Notebook. Jupyter Notebook is an interactive computing environment and Python web server, providing a browser-based UI (user interface) for Jupyter users. Jupyter Notebooks are an ordered list of input/output cells, each providing a REPL (read-eval-print loop read-eval-print loop) for writing code and a window to show output in real time.

## Libraries used

For the project we have imported different libraries for different uses:-

1. Pandas
2. Numpy
3. Sklearn
4. Matplotlib
5. Seaborn
6. sklearn. model\_selection KNeighbors Classifier
7. sklearn. model\_selection import train\_test\_split

**1. NUMPY:**

NumPy is a Python library used for working with arrays.

It also has functions for working in domain of linear algebra, fourier transform, and matrices.

NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python. In Python we have lists that serve the purpose of arrays, but they are slow to process.

NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.

Arrays are very frequently used in data science, where speed and resources are very important.

1. **PANDAS:**

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and

"Python Data Analysis" and was created by Wes McKinney in 2008. Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

1. **SKLEARN:**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**.

**Features**

Rather than focusing on loading, manipulating and summarising data, Scikit-learn library is focused on modeling the data. Some of the most popular groups of models provided by Sklearn are as follows:

**Supervised Learning algorithms:** Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.

**Unsupervised Learning algorithms:** On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.

**Clustering:** This model is used for grouping unlabeled data.

**Cross Validation:** It is used to check the accuracy of supervised models on unseen data.

**Dimensionality Reduction**: It is used for reducing the number of attributes in data which can be further used for summarisation, visualisation and feature selection.

**Ensemble methods:** As name suggest, it is used for combining the predictions of multiple supervised models.

**Feature extraction:** It is used to extract the features from data to define the attributes in image and text data.

**Feature selection:** It is used to identify useful attributes to create supervised models.

**Open Source:** It is open source library and also commercially usable under BSD license.

1. **MATPLOTLIB:**

Matplotlib is a low level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter. Matplotlib is open source and we can use it freely.

Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility.

1. **SEABORN:**

**Seaborn** is a library mostly used for statistical plotting in Python. It is built on top of Matplotlib and provides beautiful default styles and color palettes to make statistical plots more attractive. In these project we have used classification algorithm which are supervised learning algorithm such as Random forest and Logistic Regression.

1. **Random forest:**

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

**From sklearn.metrics import accuracy score**

1. **KNN 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.

We have used confusion matrix for evaluating .

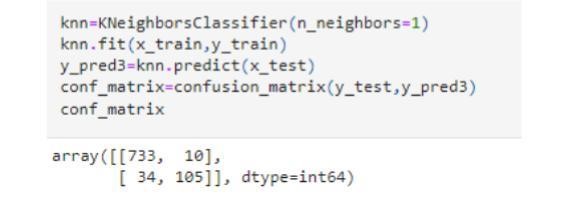
One common way to evaluate the quality of a logistic regression model is to create a confusion matrix, which is a 2×2 table that shows the predicted values from the model vs. the actual values from the test dataset.

To create a confusion matrix for a logistic regression model in Python, we can use the **confusion\_matrix()** function from the [sklearn](https://scikit-learn.org/stable/) package:

**from sklearn import metrics** **metrics.confusion\_matrix(y\_actual, y\_predicted)**

**Confusion matrix for our model:**

Conclusion



We got the best accuracy of 95.46% from Random forest classification model which is the highest accuracy among different models.

In logistic regression we got accuracy of 88.20%.

In KNN algorithm we got accuracy of 95.011%.

**Introduction to Dataset**

The “HR EMPLOYEE ATTRITION DATASET” consists of

the details of an employee like gender, age, business travel,

department, education, relationship satisfaction, and many

others. Basically, the dataset consists of exactly 2940

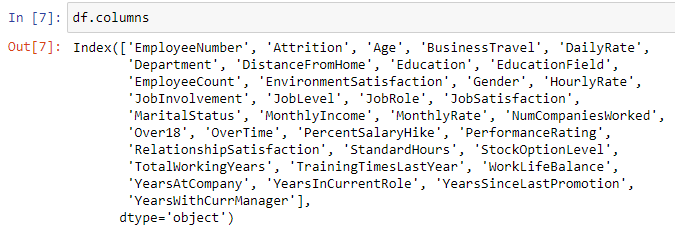
employees' data, and employee has 34 features. The dataset

consists of both numerical and categorical data. Below is an

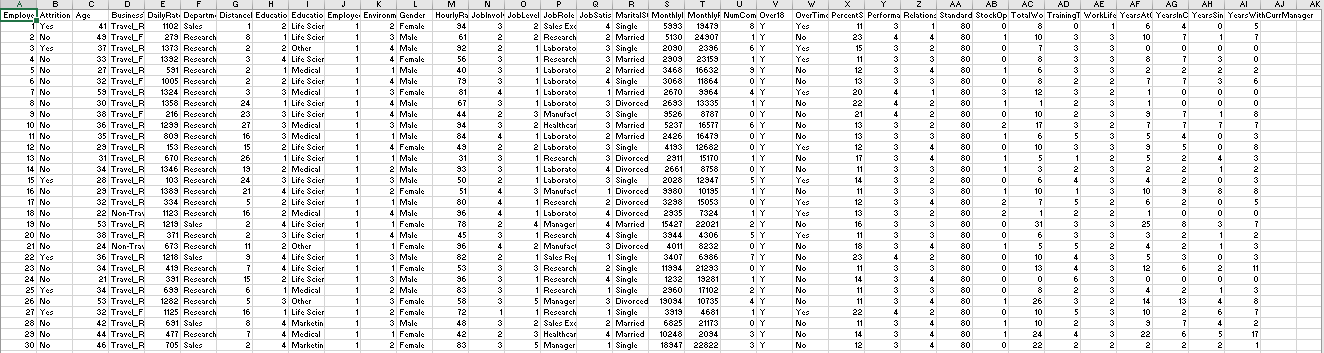
image of the dataset.

The input dataset is an Excel file with information about 2940 employees. For each employee, in addition to whether the employee left or not (attrition), there are attributes / features such as age, employee role, daily rate, job satisfaction, years at the company, years in current role, etc.

Features in the dataset :-



Dataset :-



**Exploratory Data Analysis (EDA)**

Basically, Exploratory Data Analysis(EDA) on a dataset gives

you a better understanding of the entire dataset. For an

instance, suppose anyone wants to check for any (Not Any

Value)NAN values in the dataset then EDA will help us to find

out that in the dataset. Later, we can resolve the problem of

NAN values in the dataset using various techniques like

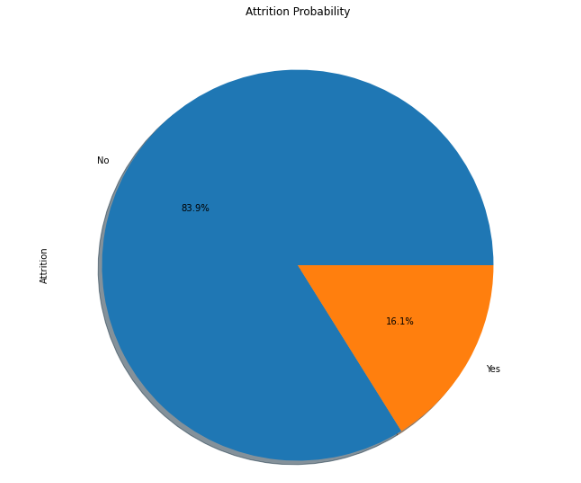
replacing the NAN values with the mean, median, or mode

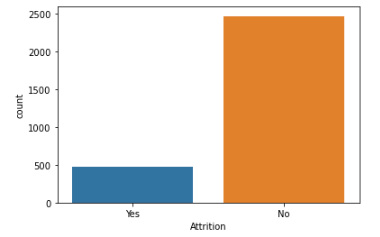
value. Luckily; in the "HR EMPLOYEE ATTRITION

DATASET", there are no NAN and outliers values. Using

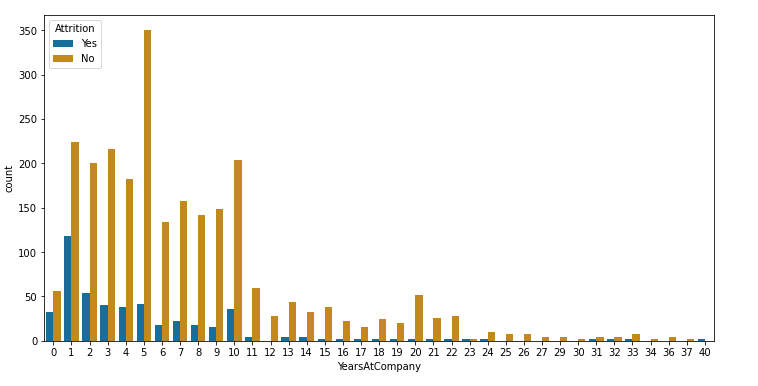
EDA, I try to find out the frequency distribution of all

categorical values w.r.t. Attrition.

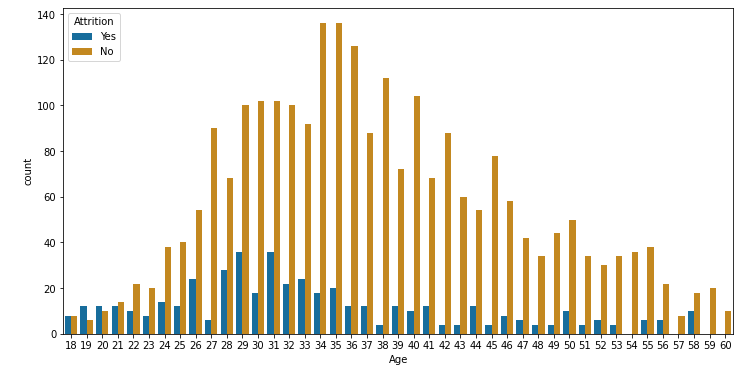




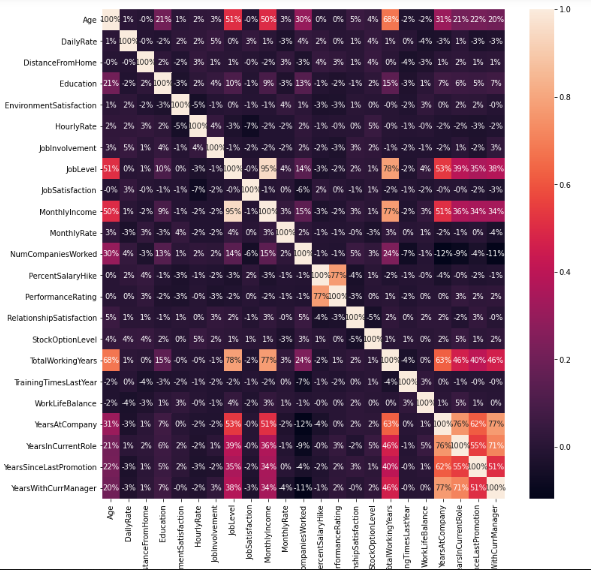
How feature (years at company) is affecting the Attrition count is shown below



How age of an employee is affecting the Attrition count is shown below



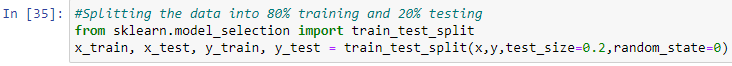
Correlation Map

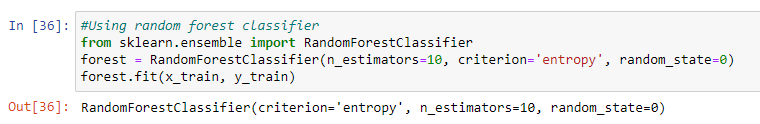


**IV. Training and Prediction of data**

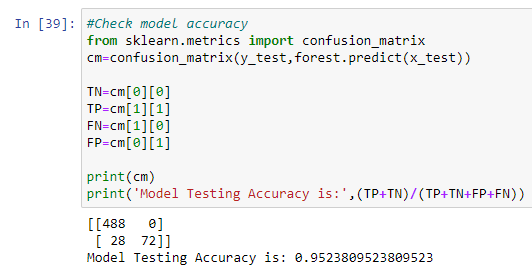
After analysing the complete dataset, I apply the Random Forest Classification algorithm to train the machine learning modelbecause this algorithm can work effectively with a large number of features.

Firstly, I split the dataset as **80 % for training and 20% for testing** as shown below :





**Accuracy of our model** after using Random Forest Classifier :



### SOME COMMON MISTAKES

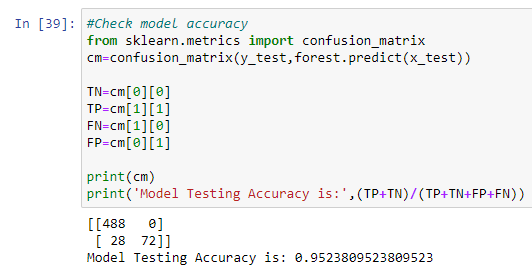
We have wasted our time on trying different machine learning algorithms . We have also wasted our time on doing unnecessary research on our data set.

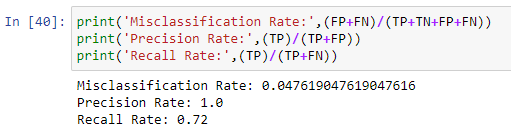
In starting , when we are trying to fit our data in different models ,we were getting wrong result due to improper cleaning of data.

**Conclusion**

In this paper, the project Employee Attrition data was analysed and various insights were drawn about the reason the employees are leaving the company along with a Random forest classifier model with testing accuracy of **95.23%** making it a Best fit.

Final Output :





### ACKNOWLEDGMENT

We would like to thank our mentors of TCR i.e saheel ramji , he help us in learning machine learning algorithms and python. We would also thanks TCR innovation for giving us the connect for our report.

### REFERENCES

1. TCR innovation study material
2. Google