**HR Analytics Project**

**Understanding the Attrition in HR**

**1. Introduction**

Companies regularly hire new employees and focus on improving the effectiveness of their workforce. The HR department plays a pivotal role, not only in hiring but also in training both new and existing employees through various programs aimed at enhancing productivity. HR analytics is employed to evaluate the impact of these training programs and provide valuable insights. One critical aspect of HR analytics is attrition, which refers to the gradual loss of employees over time. High attrition is problematic for organizations, as it leads to increased costs for recruitment, onboarding, training, and overall business disruption. This project aims to uncover insights into employee attrition and help companies make more informed hiring and retention decisions. We will use exploratory data analysis (EDA) and machine learning models to guide the decision-making process.

**2.Exploratory Data Analysing**

A screenshot of a computer

Description automatically generated

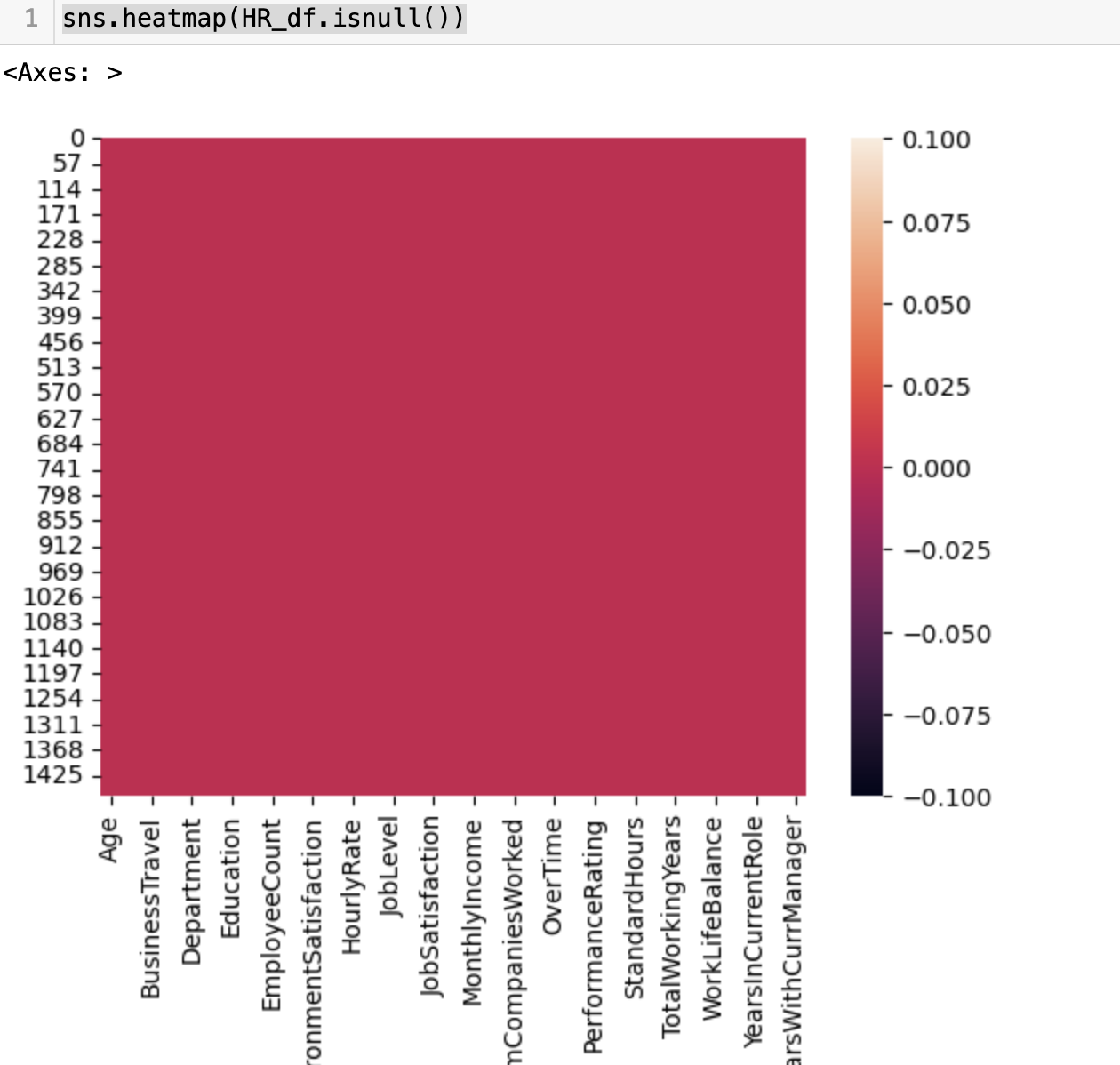
A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Based on this summary, the dataset has a total of 35 columns with 1470 unique entries. There is no null data or duplicates in the dataset, making it more usable.



The graphical presentation also shows that there is no null value present in the dataset.

A screenshot of a computer screen

Description automatically generated

After doing all the required changes in the dataset and as there no negative values present in the dataset it looks good to continue the data analysing. Following observations can be done on the bases of describe function. i. The mean age of the employees documented in the dataset is 36.92. The average Environment Satisfaction rating 2.72; average hourly rate is 65.89. ii Count of all the columns are equal. iii Right skewness of data is present because in few columns as mean is more than median. iii Left skewness of data is also there because in few columns as median is more than mean. iv There are also chances of outlier in some columns because there is major difference in the values of 75% and max in some columns like Age, Employee Number. Most of the columns are self-explanatory, according to the sources of the dataset, for ratings, a higher number means a higher rating/satisfaction.

**3. Data Visualization**

Various graphs are used to in order to understand how the data is distributed, it’s skewness, scattered. A screen shot of a graph

Description automatically generated

This output indicates that there are 1233 instances where "Attrition" is "No" and 237 instances where "Attrition" is "Yes". In this case, the bar for "No" is significantly taller than the bar for "Yes", indicating that a much larger number of employees in the dataset did not experience attrition compared to those who did.

A screenshot of a computer

Description automatically generated

This graph visualizing the relationship between "Attrition" and "Gender" in the dataset named "HR\_df". The output indicates that there are 882 male employees and 588 female employees in the dataset. Following observations can be made from the graph:

i. More male employees than female employees overall.

ii A higher proportion of male employees experienced attrition compared to female employees.

iii A larger number of female employees did not experience attrition.

A screenshot of a computer

Description automatically generated

The count plot visualizing the relationship between "Attrition" and "Marital Status" in the dataset named "HR\_df". This output indicates that there are 673 married employees, 470 single employees, and 327 divorced employees in the dataset. The graph shows following observations:

1. A larger number of married employees compared to single or divorced employees.
2. A higher proportion of single employees experienced attrition compared to married or divorced employees.
3. A larger number of married employees did not experience attrition.

A screenshot of a computer

Description automatically generated

A graph of different colored bars

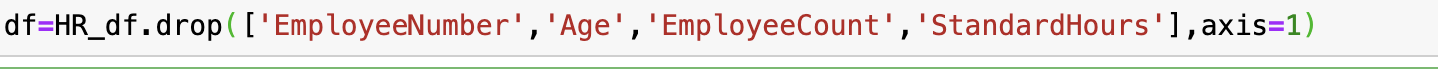
Description automatically generated

The count plot visualizing the relationship between "Attrition" and "Job Role" in the dataset named "HR\_df".It plot shows the distribution of attrition across different job roles, the number of employees who experienced attrition and those who did not, for each job role and Potential differences in attrition rates among different job roles. Following observations can made from the above graph:-

1. Sales Executive and Research Scientist have the highest number of employees overall.
2. Research Scientist has the highest number of employees who experienced attrition.
3. Sales Executive has the highest number of employees who did not experience attrition.
4. Manufacturing Director and Healthcare Representative have relatively low attrition rates.
5. Human Resources has a relatively high attrition rate compared to some other job roles.

**4. Data Preparation And Modelling**

4.1 Data Preparation



Dropping the columns which are not required in further steps



Separating dataset into x & y, where y is the target variable

4.2 Checking the skewness

A screenshot of a computer screen

Description automatically generated

Checking for skewness is an essential part of the **data preprocessing pipeline** in machine learning. Addressing skewness ensures that your models perform better by improving the stability, interpretability, and predictive power of your models.

It can be observed that some of the variable are Highly Positively Skewed and some are strongly negatively skewed. So, in order to fix this power transform is used. The power\_transform function is a powerful tool in preprocessing data for machine learning. It helps transform skewed data into a more normal-like distribution, stabilizes variance, and reduces the effect of outliers. This improves the performance of machine learning algorithms that assume normally distributed data, particularly in linear models and gradient-based optimizers.

4.3 Performing VIF

A screenshot of a computer code

Description automatically generated

A screenshot of a table

Description automatically generated

A table with numbers and text

Description automatically generated

Before make the data ready for modelling, Variance Inflation Factor (VIF) for each feature in the dataset is carried out. VIF is a measure of multicollinearity in a regression model, and it tells us how much the variance of a regression coefficient is inflated due to collinearity with other features. Multicollinearity occurs when two or more features are highly correlated with each other, which can cause issues in linear regression models and affect the stability of the estimates. The criteria to check which variable do not show significant multicollinearity and are safe to keep in the model are as follow:

1. Low VIF (1-5): Features with VIF values below 5 generally do not show significant multicollinearity and are safe to keep in the model.
2. Moderate VIF (5-10): Features with VIF values between 5 and 10 indicate moderate multicollinearity. These features should be reviewed, and if necessary, transformations (like feature scaling, dimensionality reduction) or removal may be considered.
3. High VIF (> 10): Features with VIF values above 10 show serious multicollinearity issues. You should consider removing one of the correlated features, as they may add redundancy to the model and lead to unstable estimates.

On the bases of above criteria all the features are good to go ahead to carry out modelling procedure.

4.4 Splitting the data in Train test split using industry standard 70:30

A screenshot of a computer program

Description automatically generated

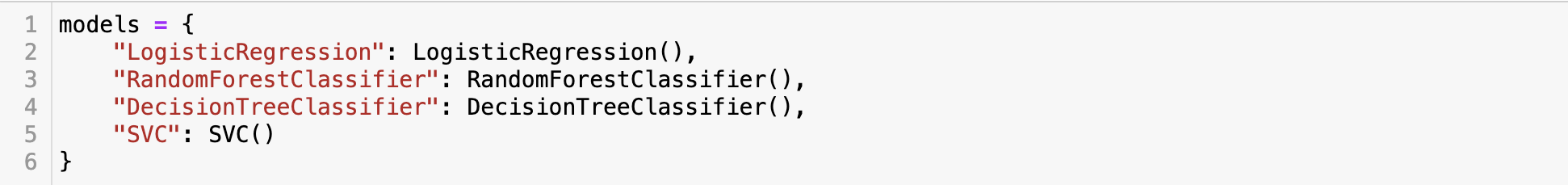
5. Modelling

we can find many types of methods for model building in skLearn Library. There are two types of models which are present skLearn Library.

1. Regression
2. Classification

In the case of our dataset, we have to make machine learning model to predict whether HR Analytics help in attrition of the employees or not. Since, we have only values, so we will use classification models to build machine learning models.

We can build as many model as we want to compare the accuracy given by these models and select the best model amongst them.



5.1 Logistic Regression

Logistic Regression is a supervised learning classification algorithm used to predict the probability of target variable. The nature of target or dependent variable is binary, which means there would be only two possible classes 1(stands for yes) or 0(stands for No).Mathematically, a Logistic Regression model predict P(Y=1)as a function of X. It is one of the simplest ML algorithm that can be used for various classification problems.

A screenshot of a graph

Description automatically generated

5.2 Random Forest Classifier

As we all know forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision tree on data sample and then prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better thana single decision tree because it reduce overfitting by averaging the result.

A screenshot of a number of numbers

Description automatically generated

5.3 Decision Tree Classifier

Decision Tree Classifier is class capable of performing multiclass classification on a dataset. As with other classifier, Decision Tree Classifier takes as input two arrays: an array X, spares or dense ,of shape(n\_smaples, n\_features) holding the training samples and an array of Y integer values, shape(n\_samples),holding class labels for training samples. It is capable of both binary (where the labels are[-1,1]) classification and multi-class (where the labels are[0,…,k-1])classification.

A screenshot of a number of numbers

Description automatically generated

5.4 SVC(Support Vector Classifier)

SVC is non parametric clustering algorithm that does not make any assumption on the number or shape of the cluster in data. In SVC data points are mapped grom data space to high dimensional features space using a kernel function. In kernel’s feature space the algorithm for the smallest sphere the enclosed the image of data using Support Vector Domain Description algorithm. This sphere when mapped back to data space, forms a set of contours which enclosed the data point. Those contours are then interpreted as cluster boundaries, and points enclosed by each contours are associated by SVC to the same cluster.

A screenshot of a computer

Description automatically generated

6. Cross Validation of the models

After comparing the model it is found that we are getting maximum accuracy score Logistic Regression. But we cannot decide the best model on the bases of accuracy score only ,since this might be possible that the data is over-fitted .So, to decide the best fit model we will check the cross validation score of each model. We can import cross\_val\_scorce from skLearn Library.

A screenshot of a computer code

Description automatically generated

The model having lowest difference between accuracy score and cross validation score will be the best model for our machine learning algorithm. So, from the above logistic regression is the best model for HR Analytics dataset. As, it shows 87% accuracy score and f1 score of approx. 88%.

7. Hyper parameter tuning

Hyper parameter optimization in machine learning intends to find hyper parameter of given machine learning algorithm the deliver the best performance as measured on validation set. Hyper parameter, in contrast to model parameter are set by machine learning engineer before training. The number of trees in random forest is a hyper parameter while weights in neural network are model parameter learned during training. I like to think of hyper parameter as a model settings to be tuned so that model can optimally solved the machine learning problem. We will use GridSearchCV for the hyper parameter tuning.

7.1 GridSearchCV

In GridSearchCV approach, machine learning model is evaluated for a range of hyper parameter values. This approach is called GridSearchCV because, it search for best set of hyper parameter from grid of hyper parameter values.

A screenshot of a computer

Description automatically generated

A white rectangular object with a black border

Description automatically generated

**Conclusion**

In this HR Analytics project, we successfully identified key factors contributing to employee attrition using a combination of exploratory data analysis, machine learning models, and hyperparameter tuning. After testing multiple models, logistic regression emerged as the best predictor of employee turnover with an accuracy of 87%. Future steps could involve gathering more data on job satisfaction, management policies, and external factors that influence attrition. Additionally, incorporating advanced models like neural networks or deep learning could further enhance predictive accuracy.

References

1. <https://scikit-learn.org>
2. <https://towardsdatascience.com/>