Methodology

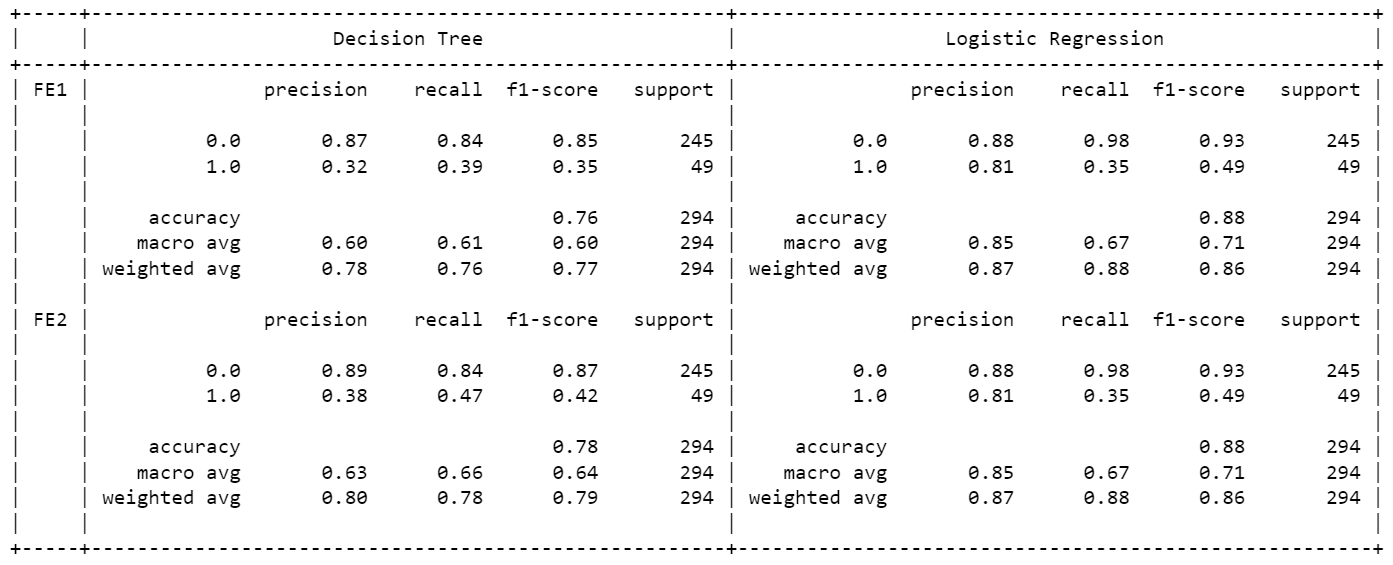
* **The two machine learning techniques used**
* Decision tree Classifier [Sec. (8 c i) of ipynb]
* Logistic Regression Classifier [Sec. (8 c ii) of ipynb]

Overview

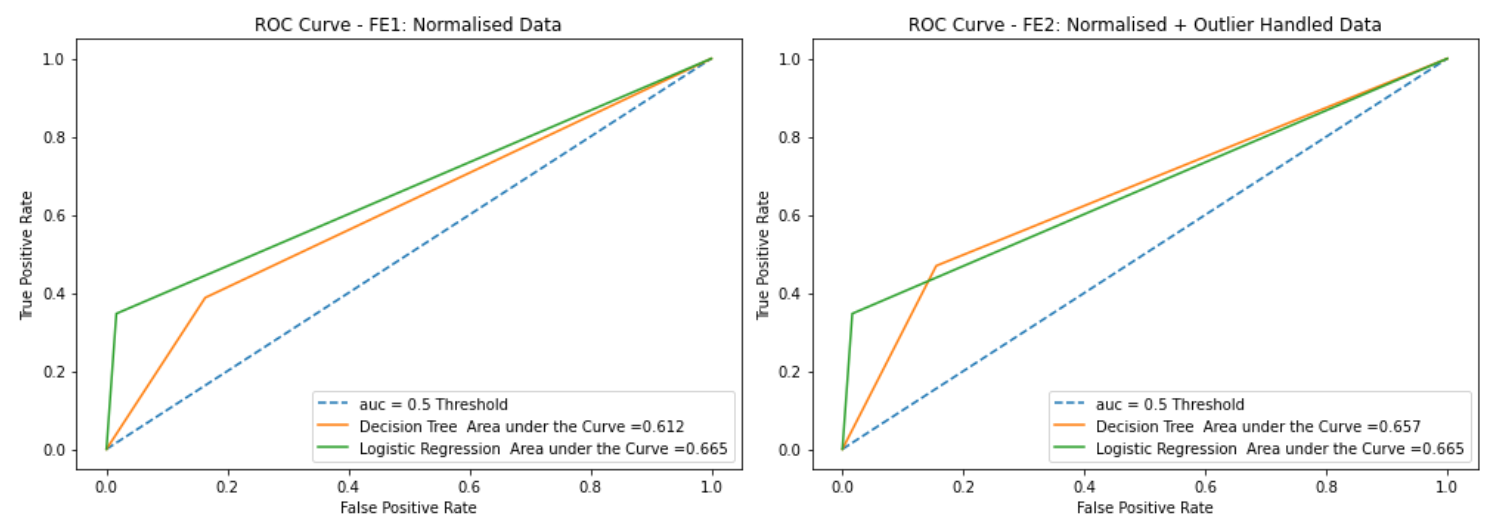
* **Objective:** Analyze and Build models for Predicting (Employee) Attrition. [Sec. 1 a i of ipynb]
* **Methodology:** Understand data, reduce features by feature selection, feature engineering, building model, evaluating model.

Results

* **Table for the evaluation metric for each ML technique used** [Sec. (10 a) of ipynb]



* **Plot of the curves** [Sec. (10 b) of ipynb]



* **Conclusion**

Logistic Regression Training Algorithm proved better while dealing with dataset having more categorical variables. It Proved better that Decision Tree Algorithm for normalized, outlier smoothened data [Refer Sec (11, 12) ipynb file]

Dataset

* **Size of the dataset**Total records: 1470, Total Attributes: 33 [Sec. 3 c of ipynb]
* **Variable type– Categorical / Interval ….**

Data types can be found in the excel object [Sec. (3 c) ipynb)



* **Data Distribution and handling imbalanced data**

Analyzed spread of data with histogram plots for continuous variables and categorical variables [Sec. (3 e) and (4 b) of ipynb]

* **Feature Wrangling**

Filled NANs with mean, median, mode values [Sec (5 a i) ipynb]. Checked for duplicates in the dataset.

* **EDA outcomes and discussion**

Analysed data by drawing count plots for categorical data, histogram plots for numeric data. [Sec. (4 b) ipynb]

Feature Engineering Techniques

* **Features removed:**

Two features were removed manually, the ‘EmployeeNumber’ and ‘Over18’ fields and using Kbest ChiSquare approach selected top 20 fields [Sec. (6 b, 6d) ipynb]

* **Feature ranking:**

KBest Chisquare and Info Gain techniques are used to determine for feature selection [Sec. (6 c, 6d) ipynb]

* **Any other:**

Using Normalisation using min-max scaler and outlier smoothening by capping the outlier at 0.95 quantile levels

[Sec. (7 c) ipynb]