**Employee Attrition Prediction**

**Team Members:**

1. Krishna Vamsi Koppula (700745021)
2. Sravani Lankala (700746285)
3. Sudhakumari Kothapalli (700744316)

**Goals and Objectives:**

**Objectives:**

* The primary goal of the project is to build the machine learning classification models to accurately predict the potential attrition in the organizations. By implementing this project, we can help the organizations to plan the measures to avoid high attrition rates. Here are the project specific objectives to achieve good accuracy of the models
* The relationships of the variables, distributions and diversity is explored in EDA. Exploratory Data Analysis gives the idea of which models should be implemented for better results.
* Data gathering: In data gathering step in addition to the ready to use parameters new parameters also used to check the model performance and their importance in prediction of attrition the features are: previous work experience, reason for leaving and previous company salary. Data preparation: Collected raw data is cleaned and prepared to match the requirements of the predictive models.
* Feature selection: Selecting most useful parameters in the raw data using Univariate feature selection method. Feature selection methods improves the results of the models.
* Building ensemble machine learning models to predict the probability of the employee attrition.
* Deploying the model into web application.

**Features:**

Main features of the project are:

1. Building the web application to display the potential attrition in the organization
2. Creating a report for high-risk employees.

**Significance:**

Organizations profit from the use of attrition prediction in the following ways:

1. Increased employee involvement: By knowing the reason for employee attrition organizations can plan the measures to retain them by creating positive work environment and opportunities
2. Manpower planning: Early prediction of attrition helps to plan the successor of the exit employees.
3. Retaining highly valued: Replacement of highly valued employees is a difficult task for organizations. Clustering the attention into highly valued and other helps to retain them.

**Motivation:**

In terms of revenue planning and other financial considerations, employee attrition is crucial. Though organizations can replace employees, the typical duration to fill a post is 42 days, and it typically takes 6 to 8 months to reach maximum productivity. To a certain extent employees leaving the organization is normal but losing highly valued employees is a risk factor. The motivation of the project originates from the need to solve the issues of organizations with the help of data driven methods to manage workforce, cost cuttings and retaining highly valued employees.

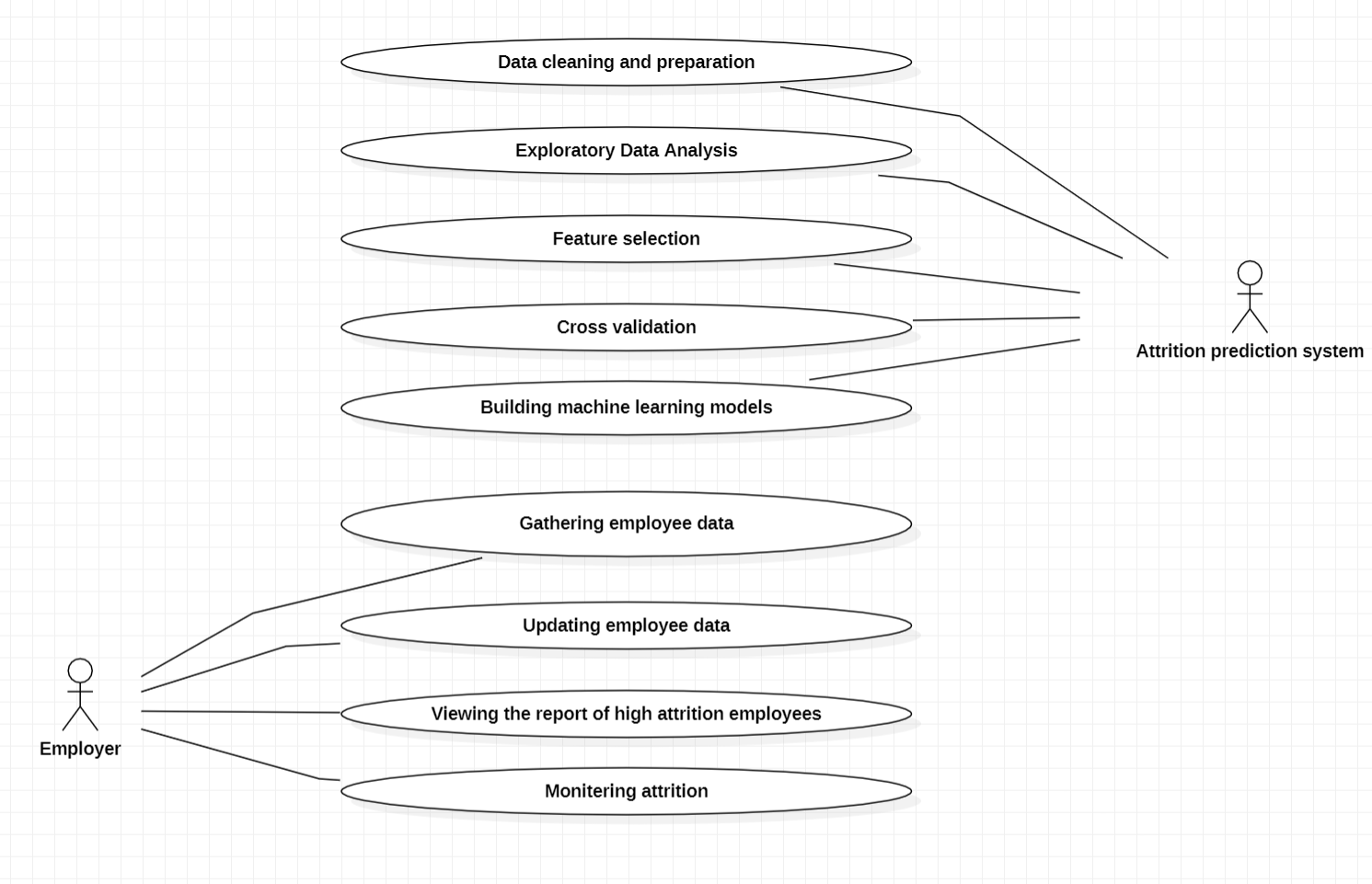
**Dataset:**

The dataset contains 1028 employee records and 31 features. Out of 31 features 15 features are adopted from IBM HR analytics dataset and other 16 parameters like job roles in previous company, salary in previous company, reason for leaving, promotion(y/n), designation after promotion is manually added.

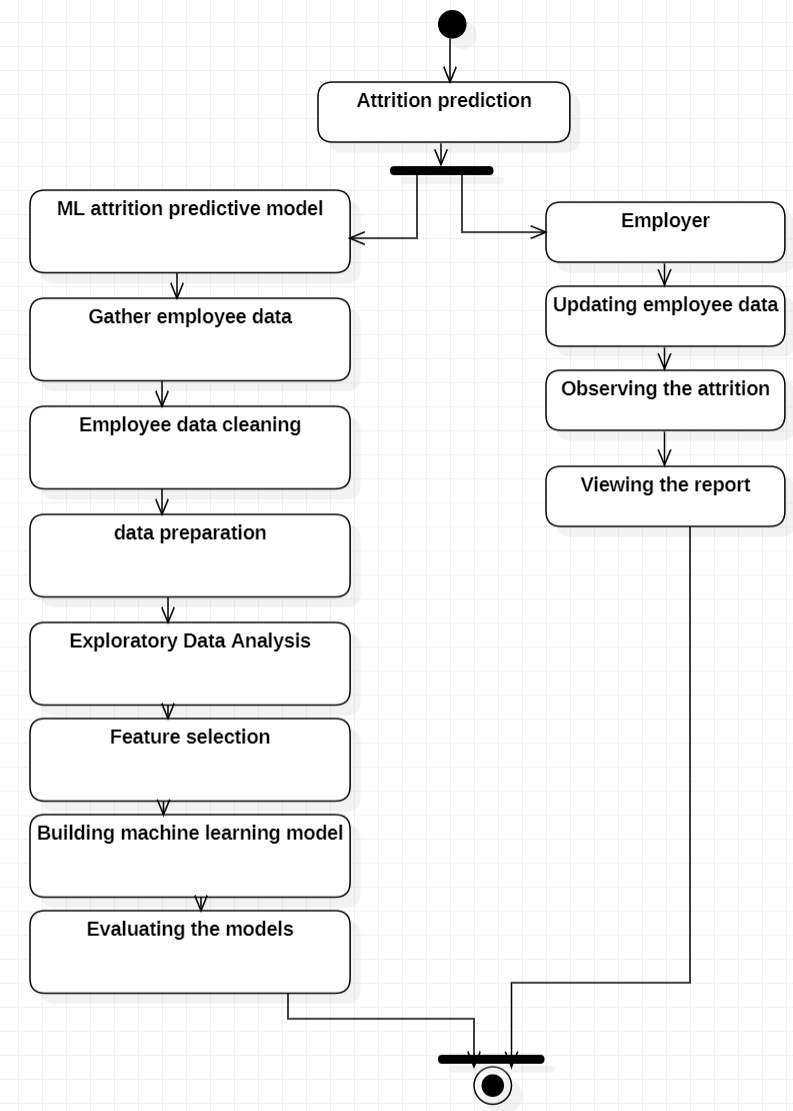
| Feature | Description |
| --- | --- |
| Employee details | Employee name, employee number, age, marital status, gender and education level, department, job role |
| Work experience | Number of companies worked, previous work experience, job roles in previous company, salary in previous company, reason for leaving, promotion(y/n), designation after promotion, job level |
| Employee expectation | Salary expectation, % hike expectation |
| Work engagement | Job involvement, job satisfaction, monthly income, overtime, performance rating, total working experience, Years at company, years with current manager, Environment satisfaction |
| other | Business travel, stock option level |

*Table1: Dataset description*

**Detail design of Features:**

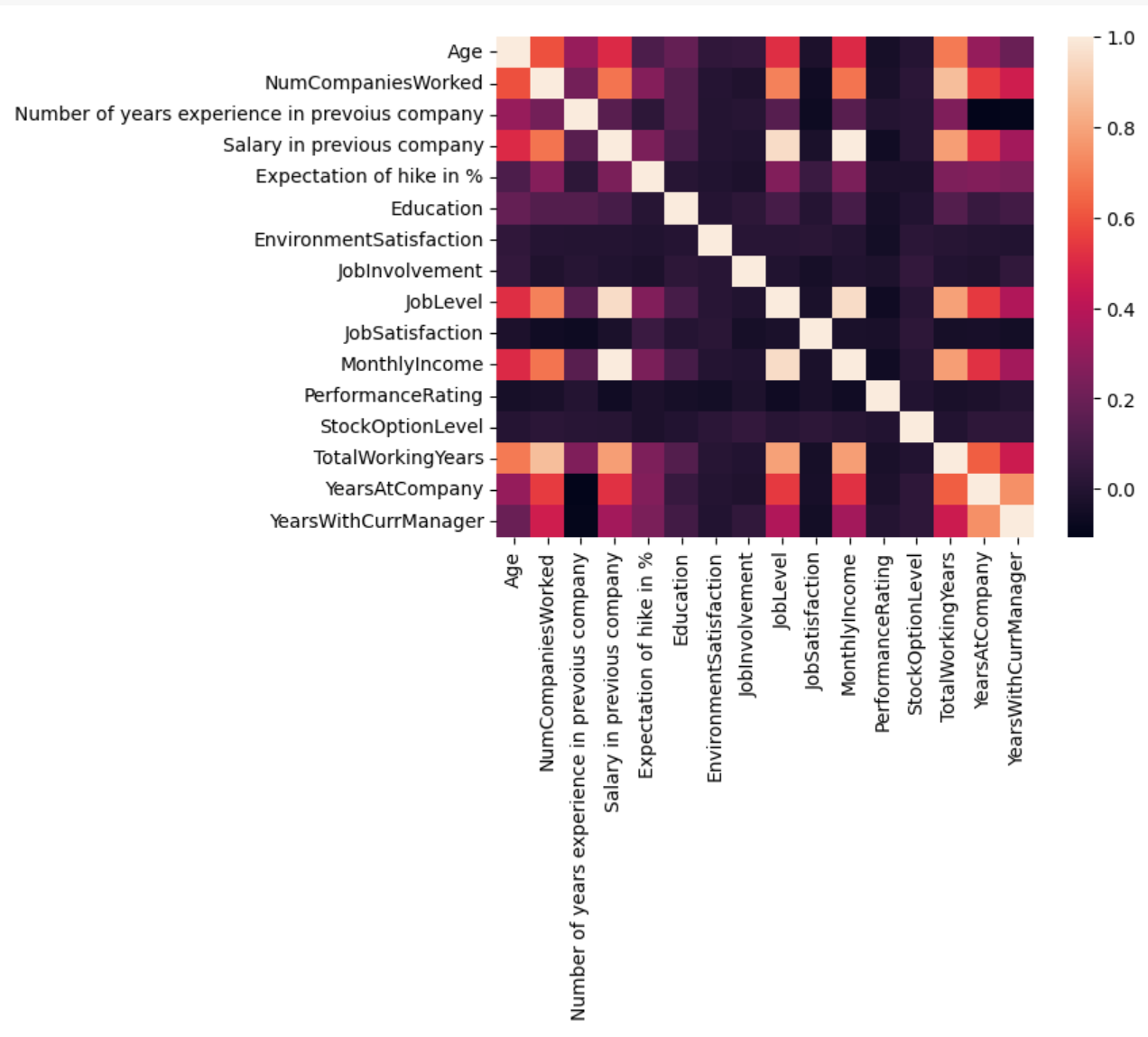


*Fig.1: Use case diagram*

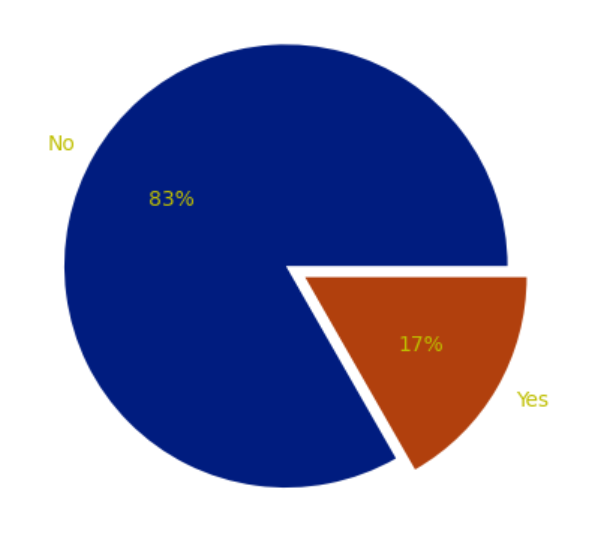
**

*Fig2: Activity diagram*

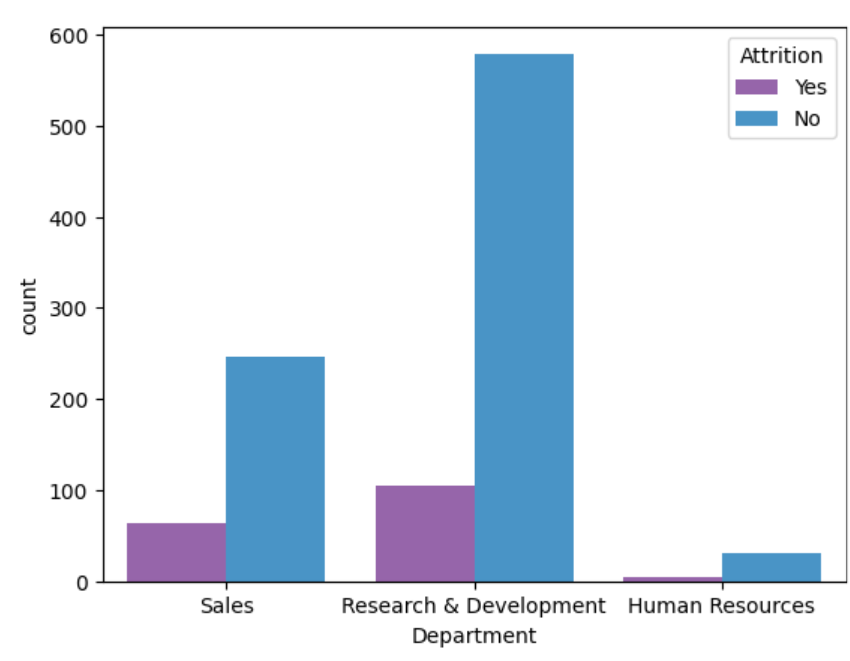
**Preliminary Results:**

**

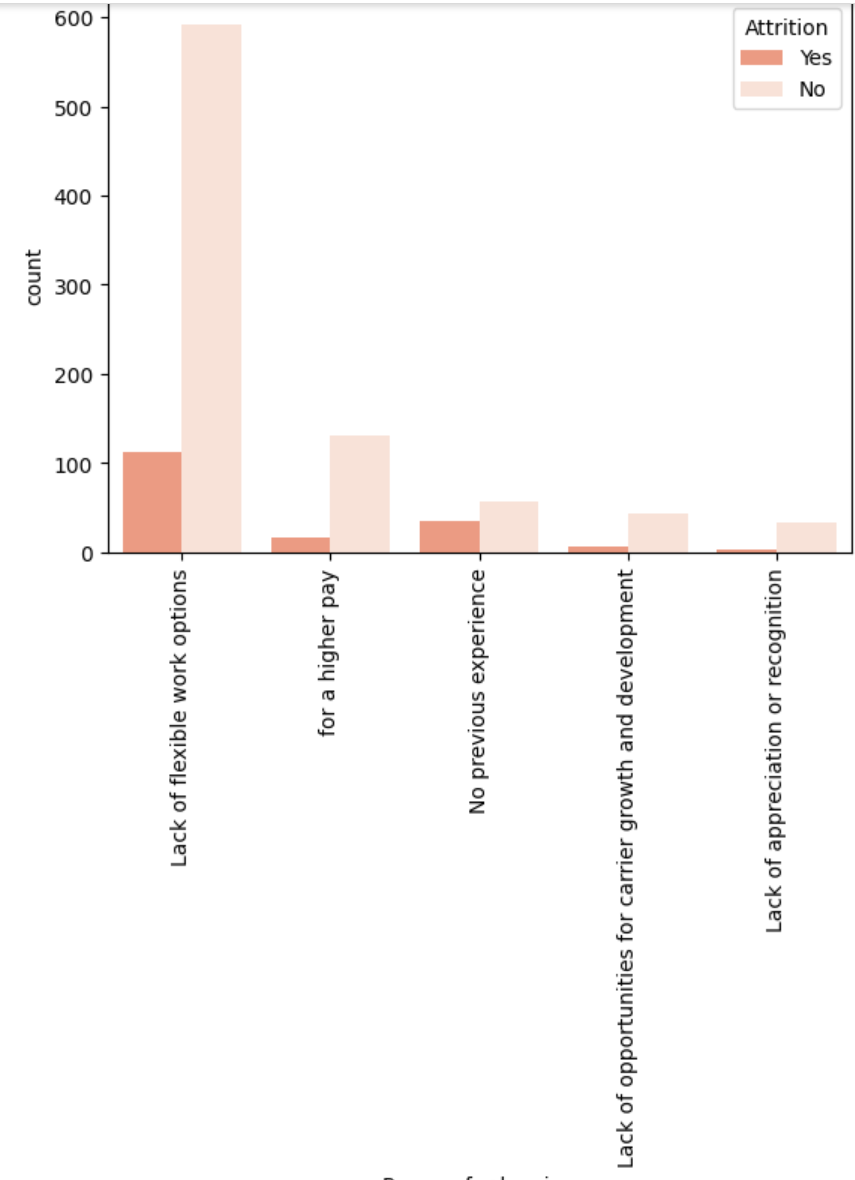
*Fig3: Correlation matrix*

**

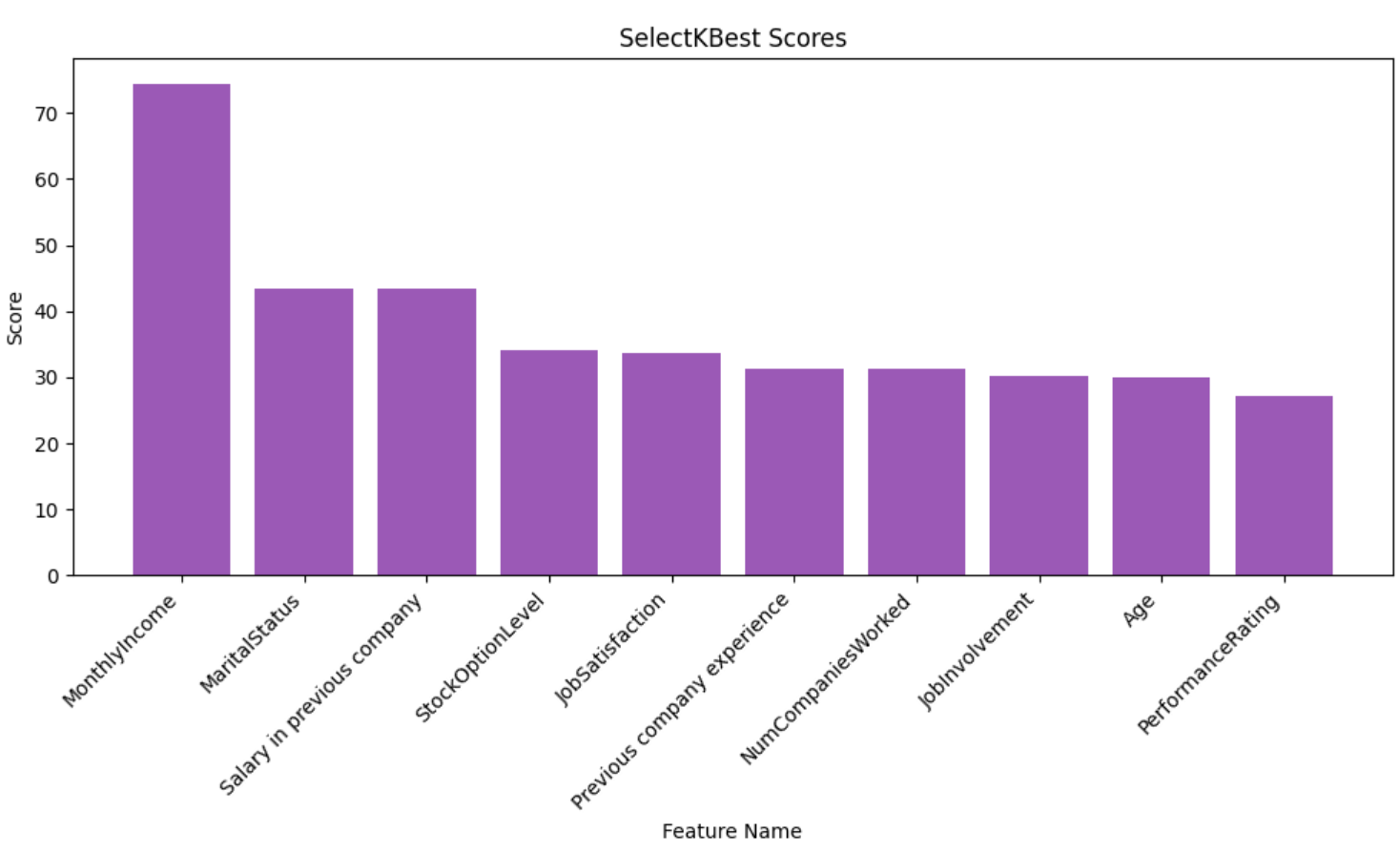
*Fig4: Targets distribution*



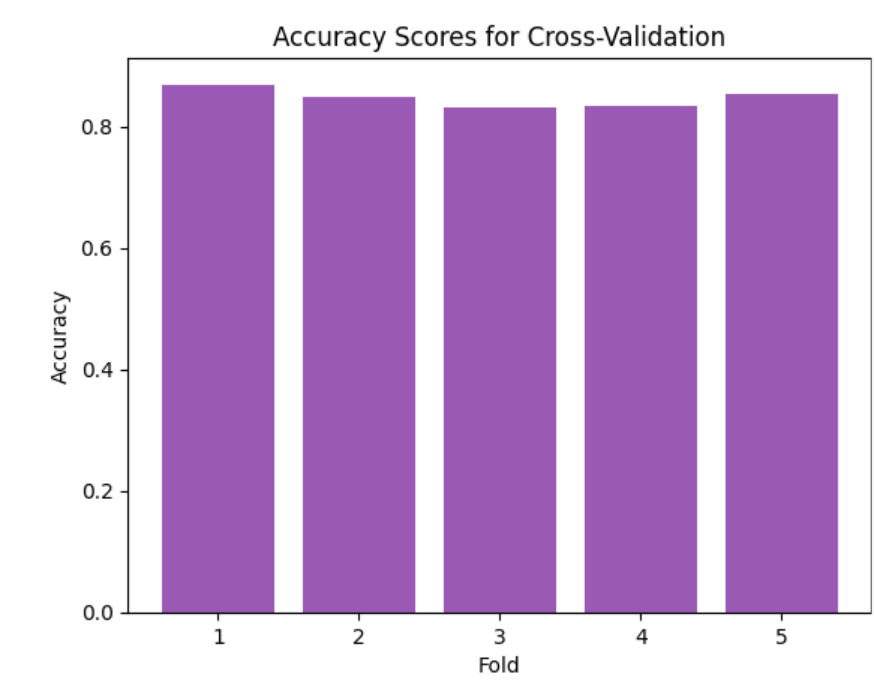
*Fig5: Attrition in various departments*



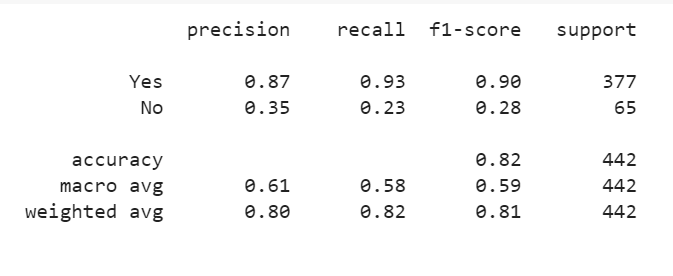
*Fig6: Cause of attrition*



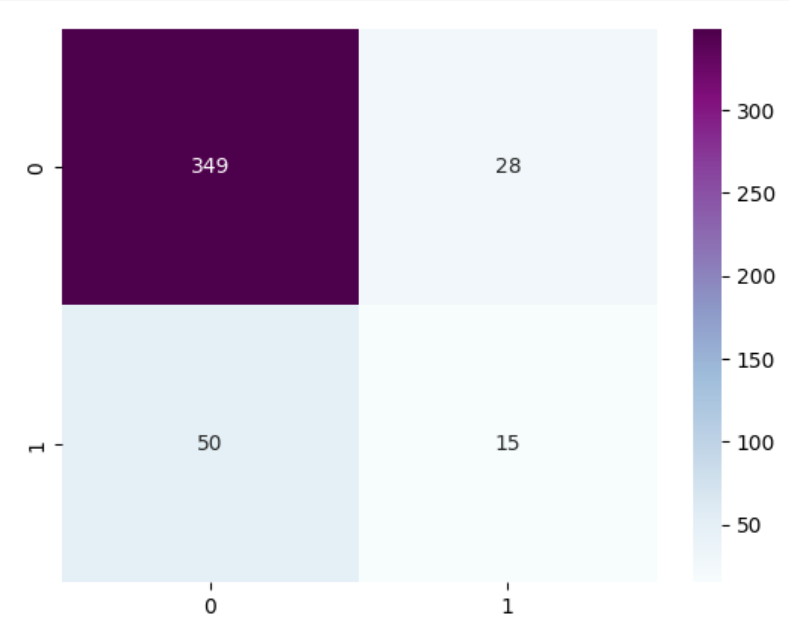
*Fig7: Feature selection scores*

****

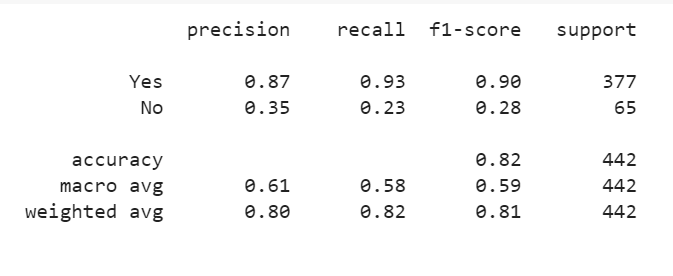
*Fig8: Cross validation scores*

****

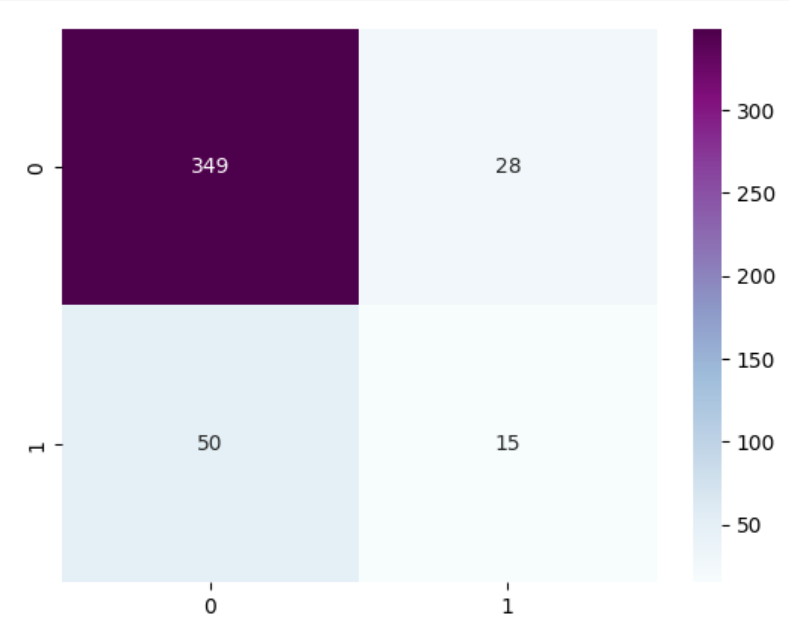
*Fig9: Classification report of Gradient boosting classifier*

**

*Fig10: Confusion matrix of gradient boosting classifier*

**

*Fig11: Classification report of AdaBoost classifier*

****

*Fig12: Confusion matrix of AdaBoost classifier*

**Analysis:**

Because of the increased interest in machine learning among business leaders and decision-makers, researchers must investigate its use within commercial organizations. One of the most pressing concerns for business owners is the loss of brilliant people. This study uses machine learning models to investigate employee attrition. Three key tests were done to forecast employee attrition using synthetic data generated by IBM Watson [1].

Professionals in science, technology, engineering, and mathematics (STEM) play an important role in economic development. STEM experts are in high demand, but there is a human resource deficit in these professions. One approach to addressing this issue is to identify students who are at risk of dropping out and then intervene with targeted methods to guarantee that these students complete the program. Therefore, this research aims to use a data mining classification technique to identify students who are at-risk of dropping out [2].

**Implementation:**

Implementation of attrition prediction follows these key steps:

**1. Data collection and preparation:**

In data gathering steps to the existing dataset additional features are added like reason for leaving the previous company, previous work experience, salary hike expectation, promotion factors. After data collection null values are observed and removed. Categorical values are encoded using label encoder. Removed constant features using variance threshold and quasi constant columns.

**2. Exploratory Data Analysis:**

In the exploratory data analysis, the relationships of the features are observed:

* Dataset is imbalanced but it is a natural case where the no attrition employees are higher than the attrition employees
* Highest attrition is observed in the department of research and development, sales. Lowest attrition is recorded in the Human Resources department.
* The reason for high attrition is lack of flexible work options.
* Most of the employees are 30-40 years of age
* Distribution of monthly income is skewed towards the right.
* Low salary group employees are the potential attrition employees
* Attrition in female and male group is similar

**3. Feature selection:**

In feature selection step 10 most important features are selected using univariate feature selection. To select the features, select K Best method is used and the underlying statistical method is f\_classification. The 10 most important features are: monthly Income, marital status, salary in previous company, stock option level, job satisfaction, previous company experience, number of companies worked, job involvement, age, performance rating

**4.Cross validation:**

To analyze the model performance prior to the training cross validation is implemented for 10 folds and accuracy of each fold is recorded and the average accuracy is 84%.

**5.Building machine learning models:**

Ensemble machine learning models, gradient boosting and AdaBoost classifiers are implemented. Each model is trained on the training data. Models are evaluated using classification report, confusion matrix and ROC-AUC curves.

**Project management:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PROJECT PROGRESS REPORT** | | | | |
| EMPLOYEE ATTRITION PREDICTION | | | | |
| **STATUS** | **CONTRIBUTION** | **TASK** | **DESCRIPTION** | **ASSIGNED TO** |
|  |  |  |  |  |
| Completed | 100% | Employee data collection | Collecting employee details and updating |  |
| Completed | 100% | Data cleaning and preparation | Preprocessing the data |  |
| Completed | 100% | Exploratory Data Analysis | Analyzing the behavior of the data |  |
| Completed | 100% | Feature selection | Selecting the features using select K Best method |  |
| Completed | 100% | Cross validation | Cross validation of data for 5 folds |  |
| Completed | 100% | ensemble classifiers | Building machine learning models |  |
| Completed | 100% | Model evaluation | Evaluating the machine learning models |  |

**REFERENCES:**

[1].S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.

[2].M. Naseem, K. Chaudhary, B. Sharma and A. G. Lal, "Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science," 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Melbourne, VIC, Australia, 2019, pp. 1-8, doi: 10.1109/CSDE48274.2019.9162389.