EXL Foundation Track Training

Final Project

Team 1
Kovida Samir Kshatri
MTech - CSE
IIIT - Delhi

Contents

- Summary
- Objective/Problem statement
- Project Flow
- Tools Used
- My Contribution
- Results
- Conclusion
- Further Enhancements

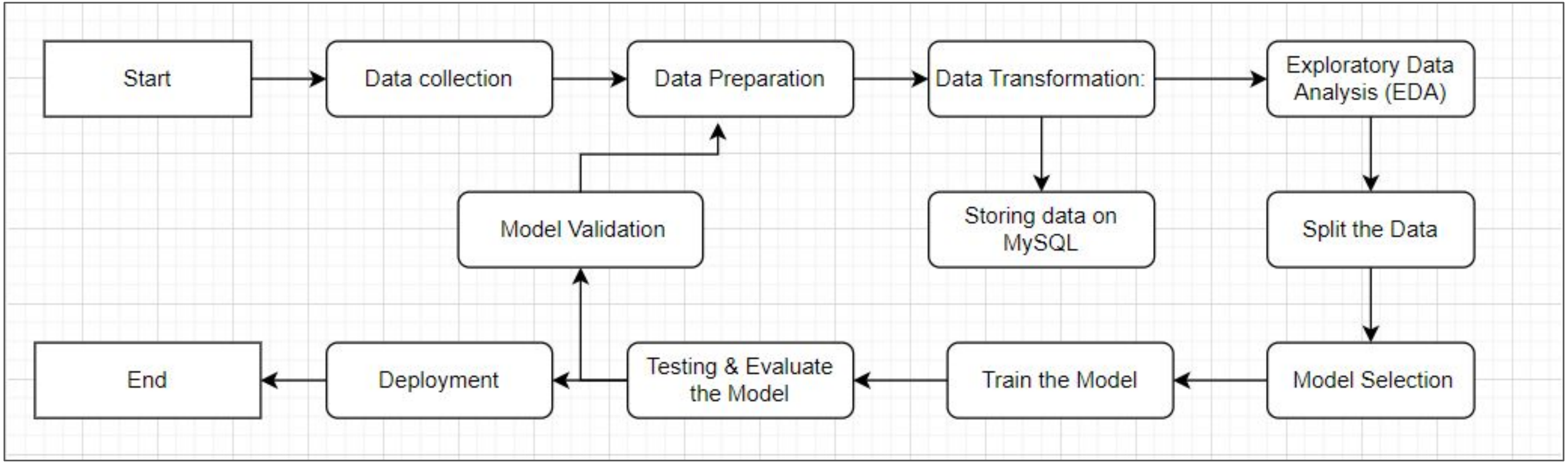
Summary

This project is an implementation of various machine learning models that is trained on a dataset that is based on employee attrition. This data is transformed and preprocessed to achieve better results. Further docker and jenkins are used for deployment.

Objective

The core objective of the project is to build a machine learning model that can predict the employee attrition, i.e., whether an employee is expected to leave an organization or not. Docker is used for deployment.

Project Flow



Tools Used

- Google Colab
- MySQL
- VS Code
- Git & Github
- Docker
- Docker Hub
- Kubernetes
- Jenkins

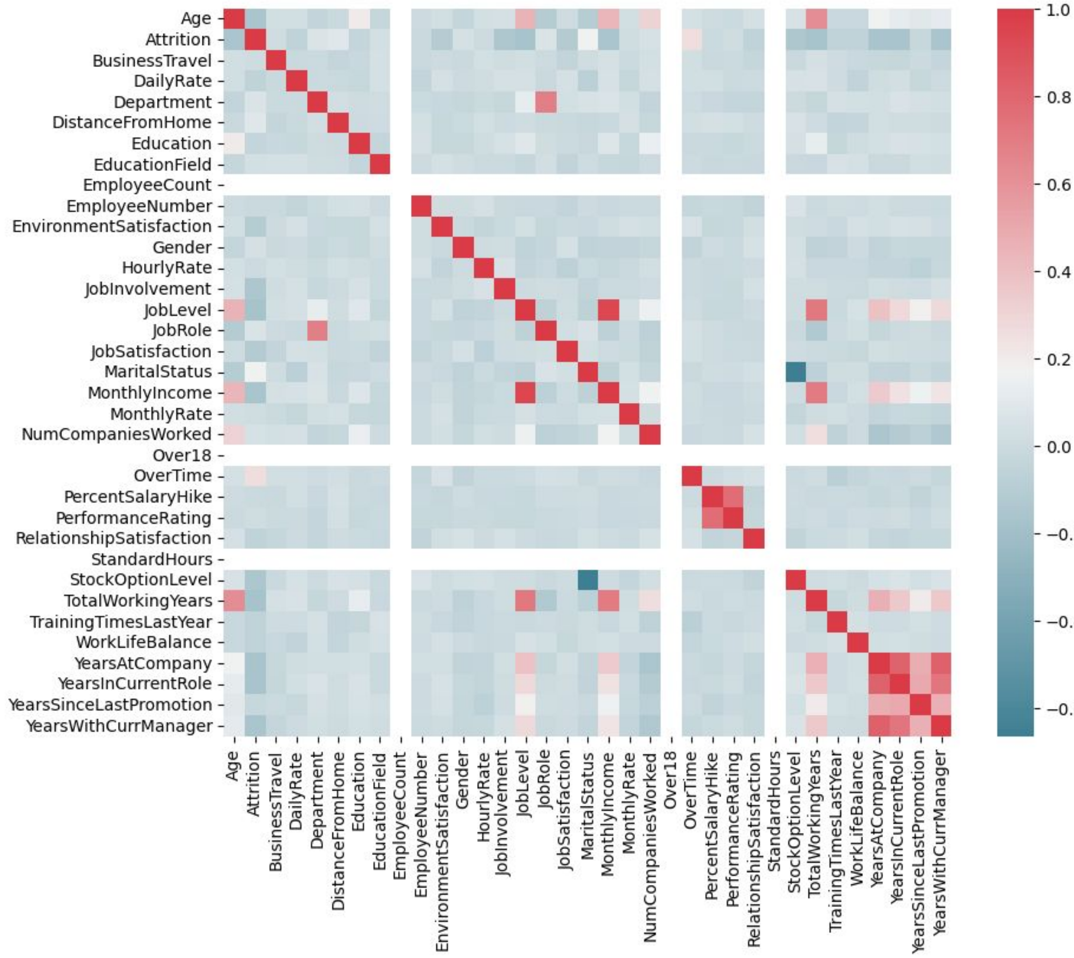
My Contribution

1. Label Encoding and Standardization
2. Visualization or Univariate analysis of all 35 features.
3. Heatmap for correlation for all 35 features.
4. Standard dataset Model Training(SVM & Logistic Regression)
5. Evaluation of this model.
6. Visualization of test results.
7. Balancing the data.
8. Again training the data on the balanced data(LinReg, LogReg, SVM, DT & RF).
9. Evaluation of all the models.
10. Visualization of test results.
11. Feature Selection on the data.
12. Again balancing the data.
13. Again training the models on the obtained data.
14. Evaluation of all above models.
15. Visualization of test results and pytest.

Exploratory Data Analysis

Data Analysis has been performed by performing univariate analysis using visualization on the data, and drawing conclusions from correlation matrices.

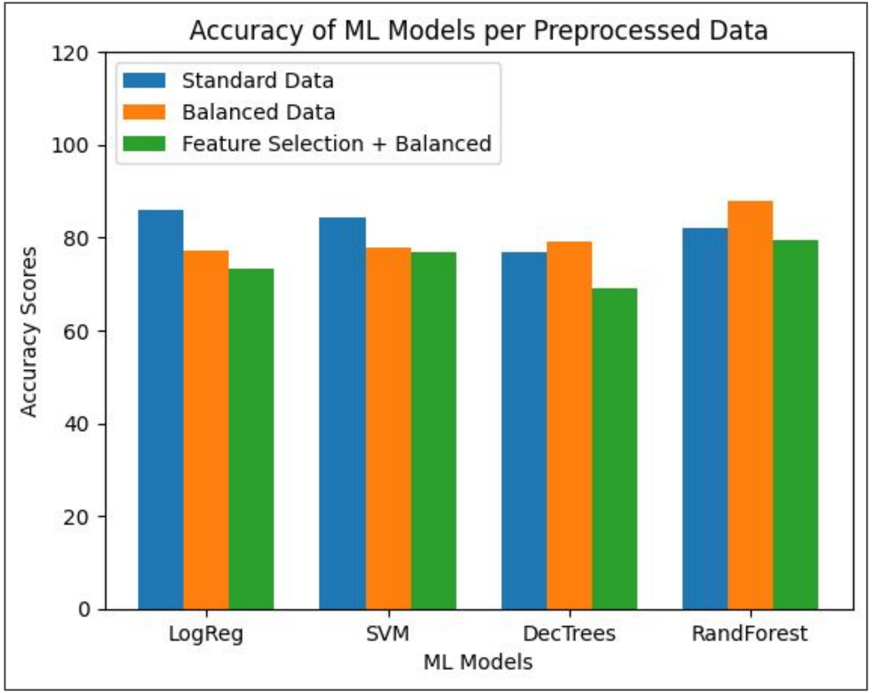
Correlation Matrix



Univariate Analysis

Refer Colab Notebook for Univariate Analysis as there are 35 graphs for 35 features. Link

Results



The plot shows the accuracies of 4 models viz., Logistic Regression, SVM, Decision Trees and Random Forest on distinctly processed data.

Also, the datasets were trained on the linear regression model but since the model is not a classifier, accuracy scores could not be obtained.

Precision and Recall have also been considered while evaluating the models.

Conclusion

Random Forest Classifier gave the best accuracy score along with acceptable precision and recall scores.

Balanced Data gives better results.

Feature Selection does not aid in improving the numbers. Poor scores are obtained upon performing feature selection.

Further Enhancements

- To reduce the dimensions of the dataset, principal component analysis can be applied.
- Further fine tuning the model can be implemented.
- Also, to improve model performance on unseen data, cross validation can be carried out.