****

**Assessment Report**

on

**Employee Attrition Prediction Using**

**Machine Learning**

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Group Members- Deepak Kumar (202401100300093)

Ashutosh Chauhan (202401100300077)

Aviral (20240110030080)

Himalaya Vashistha (202401100300121)

Section: B

**Under the supervision of**

“Shivansh Prashad”

**KIET Group of Institutions, Ghaziabad**

**1. Introduction**

With the rise of digital lending platforms, automating credit risk assessment using data-driven techniques is becoming increasingly important. This project focuses on predicting loan defaults using supervised machine learning methods. By analyzing borrower information such as credit scores, income, and loan history, the model aims to assist financial institutions in making informed lending decisions, thereby reducing the risk of loan defaults.

**2. Problem Statement**

The objective is to develop a classification model that predicts whether a borrower will default on a loan based on their financial and credit history data. Accurate prediction helps lenders identify high-risk applicants and mitigate financial losses.

**3. Objectives**

* Preprocess the dataset to prepare it for machine learning model training.
* Train a Logistic Regression classifier to predict loan default status.
* Evaluate the model's performance using standard classification metrics.
* Visualize the confusion matrix using a heatmap for better interpretability.

**4. Methodology**

**Data Collection**

The dataset is provided by the user through a CSV file upload, containing relevant borrower and loan information.

**Data Preprocessing**

* Handling missing numerical values by imputing with mean values.
* Categorical variables are converted into numerical format using one-hot encoding.
* Features are scaled using StandardScaler to normalize the range of values.

**Model Building**

* Splitting the dataset into training (80%) and testing (20%) sets.
* Training a Logistic Regression classifier on the training data.

**Model Evaluation**

* Calculating accuracy, precision, recall, and F1-score to assess performance.
* Creating a confusion matrix and visualizing it using a heatmap for intuitive understanding of prediction outcomes.

**5. Data Preprocessing Details**

* Missing numerical values are replaced with the mean of their respective columns.
* One-hot encoding is applied to transform categorical features into binary indicator variables.
* Feature scaling normalizes the dataset for better model convergence and performance.
* The dataset is split into training and testing subsets to validate model generalizability.

**6. Model Implementation**

Logistic Regression is chosen for its simplicity and effectiveness in binary classification problems. The model is trained on the preprocessed dataset to classify loan default status and then tested on unseen data.

**7. Evaluation Metrics**

* **Accuracy:** Measures the overall correctness of the model.
* **Precision:** Indicates the proportion of predicted defaults that were actual defaults, reflecting false alarm rates.
* **Recall:** Represents the proportion of actual defaults correctly identified, showing the model's sensitivity.
* **F1 Score:** The harmonic mean of precision and recall, balancing both metrics.
* **Confusion Matrix:** Visualized via a heatmap to identify true positives, false positives, true negatives, and false negatives.

**8. Results and Analysis**

The logistic regression model demonstrated reasonable predictive performance on the test dataset.  
The confusion matrix heatmap helped in understanding the distribution of correct and incorrect predictions.  
Precision and recall metrics provided insights into the trade-off between detecting loan defaults and minimizing false alarms.

**9. Conclusion**

The project successfully developed a Logistic Regression model to classify loan default risk with satisfactory performance. This approach shows the potential for automating credit risk assessment and aiding lending decisions. Future work could explore advanced machine learning models and address imbalanced data to further enhance prediction accuracy.

**10. References**

* scikit-learn documentation: <https://scikit-learn.org>
* pandas documentation: https://pandas.pydata.org
* Seaborn visualization library: https://seaborn.pydata.org

To access the website - <https://v0-attrition-prediction-website.vercel.app/>

**Screenshots-**



