Data Science Internship Project Report

Data Science Intern May 2025

**Role:** Model Development, Data Cleaning, Feature Engineering, Evaluation & Reporting

# Task 1: Predict Employee Attrition

## Objective

To build a classification model that predicts whether an employee will leave a company and to derive actionable insights for retention strategies.

## Dataset

IBM HR Analytics Employee Attrition Dataset

## Preprocessing

* + - Handled missing values (none found)
    - Encoded categorical variables using One-Hot Encoding
    - Scaled numerical features using StandardScaler

## Models Implemented

* + - Logistic Regression
    - Random Forest Classifier

## Performance Evaluation

* + - Accuracy: 85%
    - F1 Score: 0.74
    - AUC-ROC: 0.80

## Explainability

* + - SHAP used to identify key drivers of attrition:
      * Overtime
      * Job Satisfaction
      * Monthly Income

## Insights

* + - Employees working overtime and with low job satisfaction are more likely to leave.
    - Increase engagement activities and compensation reviews.

# Task 2: Text Summarization

## Objective

To create a system that generates concise summaries from long articles.

## Dataset

CNN/Daily Mail Dataset

## Preprocessing

* + - Cleaned HTML tags, special characters
    - Tokenized and removed stopwords

## Methods

* + - **Extractive Summarization:** Implemented using spaCy and TextRank

## Abstractive Summarization:

* + - * Used HuggingFace’s BART model (facebook/bart-large-cnn)
      * Truncated input to 1024 tokens max

## Evaluation

* + - ROUGE Scores:

**–** ROUGE-1: 45%

**–** ROUGE-2: 22%

* + - * ROUGE-L: 40%

## Insights

* + - BART performed better for coherence and fluency
    - Extractive summaries were more factual but less concise

# Task 3: Disease Diagnosis Prediction

## Objective

To build a predictive model for early diagnosis of diseases like diabetes.

## Dataset

PIMA Indians Diabetes Dataset

## Preprocessing

* + - Imputed zero values for glucose, BMI, etc., using medians
    - Scaled features using MinMaxScaler
    - Addressed class imbalance using SMOTE

## Models Implemented

* + - SVM
    - Gradient Boosting (XGBoost)
    - Neural Network (MLPClassifier)

## Performance Evaluation

* + - Best Model: Gradient Boosting
      * F1 Score: 0.79
      * AUC-ROC: 0.84

## Insights

* + - Glucose levels and BMI were major predictors
    - Useful for community health screening tools

# Task 4: Loan Default Prediction

## Objective

To classify whether a loan applicant will default using financial data.

## Dataset

Lending Club Loan Dataset

## Preprocessing

* + - Converted Excel to DataFrame
    - Encoded categorical features using One-Hot Encoding
    - Handled class imbalance using SMOTE

## Models Implemented

* + - LightGBM
    - SVM (RBF Kernel)

## Evaluation Metrics

* + - Precision: 0.72
    - Recall: 0.68
    - F1 Score: 0.70

## Insights

* + - Annual income, credit history length, and installment rate were significant features
    - Model can assist lenders in minimizing high-risk loans

# Conclusion

Each task demonstrated practical implementation of data science methodologies across various domains. Models were developed, evaluated, and explained effectively, providing valuable insights and tools for real-world applications.

**Prepared by:** Aitsam Zulfiqar