CODSOFT Data Science Internship - Project Report

This report presents the work completed during the CODSOFT Data Science Internship. Three tasks were successfully implemented, each covering different aspects of data science and machine learning. The tasks are as follows: 1. Titanic Survival Prediction 2. Iris Flower Classification 3. Credit Card Fraud Detection The objective of this internship was to apply machine learning techniques to solve real-world classification and prediction problems using Python.

Task 1: Titanic Survival Prediction

Dataset: Kaggle Titanic Dataset (https://www.kaggle.com/datasets/yasserh/titanic-dataset)
The Titanic Survival Prediction project involved building a classification model to predict whether a passenger survived the Titanic disaster based on features like age, gender, ticket class, fare, and cabin. Steps followed: - Data loading and cleaning - Exploratory data analysis (EDA) with visualizations - Feature encoding and handling missing values - Model training using Logistic Regression - Evaluation using accuracy, confusion matrix, and classification report Outcome: The model achieved good accuracy in predicting passenger survival with clear feature importance insights.

Task 2: Iris Flower Classification

Dataset: Iris Dataset (https://www.kaggle.com/datasets/arshid/iris-flower-dataset)
The Iris Flower Classification task aimed to classify iris flowers into three species: Setosa, Versicolor, and Virginica based on sepal and petal measurements. Steps followed: - Dataset exploration and visualization using pair plots - Splitting data into training and test sets - Standardizing features - Model training using Logistic Regression - Evaluation using accuracy, confusion matrix, and classification report Outcome: The model achieved very high accuracy and successfully classified all three iris species.

Task 3: Credit Card Fraud Detection

Dataset: Credit Card Fraud Dataset (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)
The Credit Card Fraud Detection project focused on identifying fraudulent transactions from an imbalanced dataset. Steps followed: - Loading and understanding the highly imbalanced dataset - Feature scaling - Train-test split with stratification - Model training using Random Forest Classifier with class balancing - Evaluation using accuracy, ROC-AUC score, confusion matrix, and classification report Outcome: The model was able to detect fraudulent transactions effectively with high ROC-AUC score, demonstrating strong predictive performance even with class imbalance.

Conclusion

Overall, the internship tasks helped in understanding the full machine learning workflow including data preprocessing, visualization, model training, and evaluation. The skills learned are applicable to a wide range of real-world data science problems.