GIFT UNIVERSITY
Converting Knowledge into Practical Experience

22/3/2023

Assignment 03

Instructor: Umer Ramzan

Group Members

Sufyan Ahmad:	**201980013**
Muhammad Abdullah:	**201980007**
Abdul Basit:	**201980014**
Hira Naseer:	**201980036**

Problem statement
Predict a customer buy the thing or not.

Introduction
The purpose of this report is to document the process of generating a synthetic dataset, labeling it based on a numerical attribute, splitting it into training and test sets, and training and evaluating two classifiers - logistic regression and decision tree - on the dataset.

Dataset Generation
The synthetic dataset is generated using the Faker library, which is installed using the !pip install command. The dataset consists of 1000 instances and 10 attributes, including a binary label and a numerical attribute. The attributes generated for each instance include name, address, email, job, phone number, date of birth, company, and credit card number.

Dataset Labeling
The data is labeled based on the numerical attribute, assuming that the domain problem is binary classification. The label is assigned a value of 1 if the numerical attribute is greater than 0, and 0 otherwise.

Dataset Splitting
The dataset is split into training and test sets using the train_test_split function from the scikit-learn library. The test size is set to 0.2, and the random state is set to 42.

Model Training and Evaluation
Two classifiers are trained on the dataset - logistic regression and decision tree. The logistic regression model is implemented using the LogisticRegression class from the scikit-learn library, while the decision tree model is implemented using the DecisionTreeClassifier class.

Both models are trained on the training set and evaluated on the test set using four metrics - accuracy score, precision score, recall score, and F1 score. The results of the evaluation are printed to the console.

Conclusion

In conclusion, this report documents the process of generating a synthetic dataset, labeling it based on a numerical attribute, splitting it into training and test sets, and training and evaluating two classifiers - logistic regression and decision tree - on the dataset. The results of the evaluation show the performance of the models on the test set, and can be used to inform the selection of the most appropriate model for the domain problem.