Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
Department of Computer Science

Course Information

Course Code: CS 351L
Course Title: Artificial Intelligence Lab
Instructor: Mr. Usama Arshad, PhD CS
Program: BS Cybersecurity
Semester: 5th
Reference for Lab Resources:
[CS 351L - AI Lab GitHub Repository]
(https://github.com/usamajanjua9/CS-351L---AI-Lab-)

Lab Task Details

Lab Task: 04

Lab Title: Supervised Learning - Classification with k-NN and Decision Trees

Assigned Date: 25th September 2024

Submission Deadline: 30th September 2024

Task Type: Individual

Submission Instructions

- Make a public repository on GitHub with following name:
  CS 351L - AI Lab GitHub Repository_Your_reg_no.
- Submit each completed lab task on repository and share the link to my email with screenshots of output.
  usama.arshad@giki.edu.pk
- File Naming Convention: [YourName]_CS351L_Lab02.ipynb

Late Submissions: Will incur a deduction of marks unless approved in advance by the instructor.

Task Overview

Scenario:

You are hired as a data scientist for a university. The university wants to predict whether passengers survived the Titanic disaster based on various factors such as their age, gender, ticket class, and fare paid. You will use the k-NN and Decision Tree algorithms to build models that predict whether a passenger survived.

Part 1: Data Exploration and Preprocessing

1. Explore the Dataset:

- Load the dataset and display the first few rows.
- Visualize the distribution of key features (like `Pclass`, `Age`, `Sex`, etc.).
- Check for any missing values or outliers.

2. Data Preprocessing:

- Handle missing values by either filling them (e.g., with median) or removing records with missing data.
- Encode categorical variables like `Sex` and `Embarked` into numerical values.
- Standardize or normalize the numerical features like `Age` and `Fare`.

Part 2: Implementing k-NN and Decision Trees

1. Model Training:

- Split the dataset into training and testing sets (70% training, 30% testing).
- Implement the k-Nearest Neighbors (k-NN) algorithm and train the model using the training set.
- Implement a Decision Tree algorithm and train it using the same training set.

2. Model Evaluation:

- Use the test set to make predictions for both models.
- Evaluate the performance of each model using accuracy, precision, recall, and F1-score.
- Compare the results and discuss which model performed better.

Part 3: Visualization

1. Decision Boundaries:

- Create visualizations to display the decision boundaries of both models (k-NN and Decision Tree) using two features from the dataset.
- Plot the data points along with the decision boundaries to show how each model classifies the data.

2. Performance Visualization:

- Plot a bar chart showing the performance metrics (accuracy, precision, recall, F1-score) of both models for easy comparison.

Dataset Source:

For this lab, you will use the publicly available Titanic dataset from Kaggle. Download it from the following link:
https://www.kaggle.com/c/titanic/data

How to Load the Dataset in Python:

Use the following code to load the dataset:

```python
import pandas as pd

# Load the dataset
url = 'https://www.kaggle.com/c/titanic/data'
titanic_data = pd.read_csv('train.csv')

print(titanic_data.head())
```

----------*to err is human*----------