Final Exam
Course: Machine Learning
Topic: Supervised Learning and Model Deployment

Instructions:
Work on the following projects in a clear structure and organized manner.
Include all relevant code either as an attachment or as part of your answer.
Explain each step in detail.
Time allotted: The time to collect answer document is on the scheduled MPML exam date

Example of datasets that can be used in this project:

- UPI Payment Transactions Dataset
- Predict Restaurant Menu Items Profitability
- Customer Engagement in Online Food Service
- Uber Fares Dataset
- Shoe Dataset

You can also use any other suitable dataset available on Kaggle.

Exam Questions:
You are tasked with performing a supervised learning task using a dataset from Kaggle. The project involves comparing several supervised learning methods and deploying the best model using Heroku.

Stage 1: Dataset Selection and Exploration (10 Points)

- Select a dataset from Kaggle suitable for a supervised learning task. Provide a brief description of the dataset and its features. Explain why you chose this dataset.
- Perform an exploratory data analysis (EDA) to understand the dataset. Include visualizations to illustrate key insights.

Stage 2: Data Preprocessing (20 Points)

- Describe the steps you would take to preprocess the dataset, including handling missing values, encoding categorical variables, and scaling numerical features.
- Write Python code to preprocess the dataset based on your described steps.

Stage 3: Model Training and Comparison (30 Points)

- Choose at least three supervised learning algorithms (e.g., Regression, Decision Tree, SVM, KNN, Naive Bayes, ANN). Explain the rationale behind choosing these algorithms.
- Train each model on the preprocessed dataset and evaluate their performance using appropriate metrics.
- For classification tasks, use metrics such as accuracy, precision, recall, and F1-score.
- For regression tasks, use metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2).

- Provide visualizations to compare the performance of the models.
- Write Python code to train and evaluate each model, including any necessary hyperparameter tuning.
- Implement cross-validation to validate the performance of each model and ensure robustness.

Stage 4: Model Selection and Deployment (20 Points)

- Based on the performance metrics, select the best-performing model. Explain why you chose this model over the others.
- Provide a step-by-step guide on how to deploy the chosen model using Heroku. Include the necessary files and configurations.
- Write Python code to deploy the model on Heroku. Use your personal name in the Heroku domain name. Provide a link to the deployed model.

Stage 5: Documentation and Interpretation (10 Points)

- Document the entire process, including dataset selection, preprocessing, model training, and deployment. Ensure clarity and completeness so that others can reproduce your work.
- Interpret the results of your analysis and model performance. Discuss any patterns or insights you discovered during your analysis.

Stage 6: Evaluation (10 Points)

- How you could further improve the model's performance. Implement one of the proposed improvements and evaluate its impact on the model's performance.

Kaggle dataset link:

https://www.kaggle.com/datasets/devildyno/upi-payment-transactions-dataset

https://www.kaggle.com/datasets/rabieelkharoua/predict-restaurant-menu-items-profitability

https://www.kaggle.com/datasets/hamzasajjad321/customer-engagement-in-online-food-service

https://www.kaggle.com/datasets/yasserh/uber-fares-dataset

https://www.kaggle.com/datasets/mdwaquarazam/shoe-dataset

or you can use any other suitable dataset on Kaggle