Capstone Project Guidelines

June 08, 2025

Capstone Project Guidelines

Overview

This document outlines the guidelines for the Capstone Project Exam scheduled on Day 6/7. The project aims to reinforce Machine Learning concepts learned over the previous 5 days. All 7 teams are required to complete one regression task and one classification task using the assigned datasets.

Instructions

- Develop a simple ML model for each task using the techniques covered (e.g., supervised learning, model evaluation).
- On Day 6/7, each team will present their models, explaining the process and outcomes, including inference using Streamlit.
- Ensure all code and documentation are prepared in advance for a smooth presentation.

Team Dataset Assignments

- Team 1: Regression dataset: COVID-19 Dataset, Classification dataset: Iris Dataset
- Team 2: Regression dataset: California Housing Dataset, Classification dataset: Fish Market Dataset
- Team 3: Regression dataset: Diabetes Progression Dataset, Classification dataset: Breast Cancer Dataset
- Team 4: Regression dataset: Auto MPG Dataset, Classification dataset: Email Spam Dataset
- Team 5: Regression dataset: Wine Quality Dataset, Classification dataset: Titanic Dataset
- Team 6: Regression dataset: Concrete Strength Dataset, Classification dataset: Mushroom Dataset
- Team 7: Regression dataset: Energy Efficiency Dataset, Classification dataset: Credit Card Fraud Dataset

Dataset Links

- COVID-19 Dataset: https://www.kaggle.com/datasets/meirnizri/covid19-dataset
- Iris Dataset: https://www.kaggle.com/datasets/uciml/iris
- California Housing Dataset: https://www.kaggle.com/datasets/camnugent/california-housing-prices
- Fish Market Dataset: https://www.kaggle.com/datasets/vipullrathod/fish-market
- Diabetes Progression Dataset: https://www.kaggle.com/competitions/diabetese-progression/data?se
- Breast Cancer Dataset: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
- Auto MPG Dataset: https://www.kaggle.com/datasets/uciml/autompg-dataset
- Email Spam Dataset: https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset
- Wine Quality Dataset: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset
- Titanic Dataset: https://www.kaggle.com/competitions/titanic/data
- Concrete Strength Dataset: https://www.kaggle.com/competitions/concrete-strength-regression/data
- Mushroom Dataset: https://archive.ics.uci.edu/dataset/73/mushroom
- Energy Efficiency Dataset: https://www.kaggle.com/datasets/elikplim/energy-efficiency
- Credit Card Fraud Dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Complete ML Process

- **Problem Definition**: Identify the regression (e.g., predicting continuous values like housing prices) and classification (e.g., predicting categories like spam vs. not spam) tasks.
- **Data Collection**: Download the assigned datasets from the provided links.
- **Data Preprocessing**: Clean data by handling missing values, encoding categorical variables, and normalizing features.
- **Exploratory Data Analysis (EDA)**: Visualize data distributions, correlations, and key trends using plots.
- **Model Selection**: Choose appropriate algorithms (e.g., Linear Regression for regression, Logistic Regression or Decision Trees for classification).
- **Training**: Split data into training and testing sets, then train the model on the training data.
- **Evaluation**: Assess model performance using metrics (e.g., Mean Squared Error for regression, Accuracy for classification) on the test set.

- **Inference**: Implement the trained model in a Streamlit app for real-time predictions.
- **Documentation**: Record the process, including code, visualizations, and results for the presentation.

Submission and Presentation

- Submit your models and Streamlit app code by the end of Day 6.
- Each team will have 15 minutes for presentation on Day 7.
- Focus on explaining the dataset, model development, and inference results.