Project title: Automated model selection and hyperparameter optimization using Bayesian optimization: Enhancing Machine learning models

Team members:

1. **Name:** JAGMOHAN PRAJAPAT

   CAN ID Number: 33511238

2. **Name:** B GEETHANJALI

   CAN ID Number: 33508437

3. **Name:** MAHABOOB PASHA

   CAN ID Number: 33507359

4. **Name:** JUNAID MEHRAJ PANDITH

   CAN ID Number: 33515666

Institution Name: HKBK College of Engineering

Phase 1: Problem Analysis

1) Problem statement
2) Understanding the problem
3) Dataset understanding
4) Tools and technology selection
5) Constraints and assumptions

1. Problem Statement

This project proposes the automation of the model selection and tuning process using Bayesian Optimization. The general scope of the project is to improve the performance of machine learning models in time series classification, image classification, clustering, and regression tasks on diverse datasets.

This solution is working towards the following objectives:

- Improve accuracy and reduce manual labour in hyperparameter tuning.
- Ensure scalability and versatility to the optimization framework across diverse datasets.
- Minimize computational overhead while ensuring good model performance.

The success of the project will be measured with respect to metrics that include accuracy, precision, recall, and efficiency using different datasets.

2. Understanding the Problem

The automation of model selection and tuning hyperparameters has a tremendous influence on machine learning model precision and efficiency. Approaches that rely extensively on trial-and-error are tedious, especially when it comes to datasets with different levels of complexity. Bayesian Optimization, which is a probabilistic model-based optimization technique, allows far more systematic exploration of the hyperparameter space and guarantees finding the optimal setting in a relatively efficient manner.

This would translate l data to solve the details defined below:

- Treatment of a variety of data structures (sequential data, images, structured table).
- Availability to searching classification, regression, and clustering algorithms.
- Balance between the exploited computational resource and the depth of the hyperparameter search.

3. Dataset Understanding

The datasets selected for the project include:

1. **Time Series Classification**
   - Dataset Description: Temporal datasets with features like trends, seasonality, and irregular patterns (e.g., daily temperature, birth rates).
   - Challenges: Sequential dependencies, missing values, and potential non-stationarity.
   - Preprocessing: Time series decomposition, missing value imputation, and scaling.
2. **Image Classification and Clustering**
   - Dataset Description: Skin cancer segmented images with labeled categories for classification.
   - Challenges: High-resolution image processing, class imbalance, and potential noise in data.
   - Preprocessing: Data augmentation, resizing, normalization, and oversampling for minority classes.
3. **Classification and Regression**

- Dataset Description: Includes structured datasets like car prices, house prices, insurance costs, salaries, and customer data.
- Challenges: Mixed feature types (categorical and numerical), missing values, and multicollinearity.
- Preprocessing: Feature encoding, outlier detection, and normalization.

4. Tools and Technology Selection

The tools and libraries selected for the project are chosen based on their efficiency, compatibility, and ease of use:

1) **Bayesian Optimization Framework**:
    - Optuna or Scikit-Optimize for systematic hyperparameter tuning.
2) **Machine Learning Frameworks**:
    - Scikit-learn for traditional ML models, TensorFlow/Keras, or PyTorch for deep learning tasks.
3) **Visualization and Analysis**:
    - Pandas for data manipulation, Matplotlib and Seaborn for exploratory data analysis.
4) **Experiment Tracking**:
    - MLflow to manage and track hyperparameter configurations and results.
5) **Image Processing**:
    - OpenCV and PIL for preprocessing image datasets.

5. Constraints and Assumptions

Constraints:

- Computational Limitations: Efficient use of resources to avoid prolonged training times.

- Dataset Variability: Addressing the unique challenges posed by time series, images, and structured data.
- Project Timeline: Adherence to deadlines for each phase of the project.

Assumptions:

- Datasets are pre-verified for quality and are representative of their respective domains.
- Hardware resources (e.g., GPUs) are available for high-performance computation.
- Tools and libraries used are well-documented and provide sufficient community support for troubleshooting.