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

**Phase 1: Problem Analysis**

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

**Objectives:**

1. Automate model selection and hyperparameter optimization for diverse datasets across different machine learning tasks using Bayesian Optimization.
2. Evaluate the performance gains and computational efficiency of the automated pipeline compared to manual tuning.

**Constraints:**

* **Computational Limitations**: Optimization should balance computational cost with performance improvements.
* **Scalability**: Ensure the pipeline handles datasets of varying sizes and complexities.
* **Tool Compatibility**: Selected tools and libraries must support Bayesian Optimization and the chosen ML tasks.

**Success Metrics:**

* **Accuracy/Performance**: Achieve >90% accuracy for classification tasks, reduce regression errors by a predefined margin.
* **Optimization Efficiency**: Reduce hyperparameter tuning time by at least 30%.
* **Model Versatility**: Validate the pipeline across tasks like time series forecasting, image classification, and regression.

**Dataset Analysis Report**

**1. Time Series Classification Data**

* **Key Characteristics**: Temporal data with sequential dependencies.
* **Challenges**: Requires preprocessing steps like resampling, handling missing timestamps, and feature engineering for time-lags.
* **Preprocessing Needs**: Normalize data, impute missing values, and split into training/testing sequences.

**2. Image Classification and Clustering: Skin Cancer Dataset**

* **Key Characteristics**: Segmented image data with pixel-level annotations for skin cancer types.
* **Challenges**: Class imbalance, high dimensionality, and the need for augmentation techniques.
* **Preprocessing Needs**: Resize images, apply data augmentation (e.g., rotation, flipping), and normalize pixel values.

**3. Classification and Regression Datasets**

* **Examples**: House prices, salaries, car prices, etc.
* **Key Characteristics**: Structured/tabular data with numerical and categorical variables.
* **Challenges**: Handling missing data, encoding categorical variables, and addressing multicollinearity.
* **Preprocessing Needs**: Feature scaling (e.g., MinMaxScaler), imputation of missing values, and one-hot encoding for categorical features.

**Tool Selection Justification**

**1. Bayesian Optimization**

* **Library**: Optuna or scikit-optimize
* **Reason**: Provides efficient hyperparameter optimization with advanced acquisition functions.

**2. Model Training and Evaluation**

* **Libraries**:
  + scikit-learn: Versatile for tabular datasets (regression, classification).
  + TensorFlow/Keras: Ideal for image classification tasks.
* **Reason**: Wide support for ML models and compatibility with optimization libraries.

**3. Experiment Tracking**

* **Tool**: MLflow
* **Reason**: Simplifies tracking model performance and comparing different experiments.

**4. Data Visualization and Exploration**

* **Libraries**: pandas, matplotlib, seaborn.
* **Reason**: Facilitates exploratory data analysis and insight generation.

**Research Summary**

**Findings from Literature Review:**

1. **Bayesian Optimization**: Proven effective in reducing computational overhead in hyperparameter tuning by focusing on promising regions of the search space.
2. **Time Series Modeling**: Studies highlight the importance of capturing temporal dependencies and feature engineering for better forecasting accuracy.
3. **Image Classification**: Preprocessing like augmentation and balanced sampling significantly improves model generalization, especially in medical imaging datasets.
4. **Regression Challenges**: Handling missing data and ensuring model interpretability are critical for success in structured data regression tasks.