**Phase 1: Problem Analysis** (self research)

**1. Understand the Problem**

**Objectives:**

* Define the purpose of the project clearly: Automating model selection and hyperparameter optimization using Bayesian Optimization to enhance ML model performance.

**Key Steps:**

1. **What Problem are You Solving?**
   * Automate the tedious and error-prone manual processes of selecting models and tuning hyperparameters.
   * Provide a scalable and efficient solution for both novice and expert data scientists.
2. **Set Clear Goals**:
   * Minimize computational costs while maximizing model performance.
   * Ensure the system works for a variety of ML tasks (classification, regression, etc.).
3. **Define Success Metrics**:
   * For example: Accuracy, F1-score (for imbalanced data), Mean Squared Error (MSE), or domain-specific metrics.
   * Success includes not just high accuracy but also reduced time and effort in experimentation.

**2. Dataset Understanding**

**Objectives:**

* Analyze the dataset(s) you’ll be working with to anticipate challenges and preprocessing needs.

**Key Steps:**

1. **What Type of Data Do You Have?**
   * Is it tabular (CSV files), text, images, or time series?
   * Identify features (independent variables) and target labels (dependent variables).
2. **Check for Data Characteristics**:
   * **Size of the dataset**: Large datasets may require sampling; small datasets may need cross-validation.
   * **Dimensionality**: High-dimensional data may require dimensionality reduction techniques like PCA.
   * **Missing values**: Determine whether they are random or systematic.
3. **Preliminary Data Exploration**:
   * Conduct Exploratory Data Analysis (EDA) to understand distributions, outliers, and feature importance.
   * Tools: pandas, seaborn, matplotlib.

**3. Tool and Technology Selection**

**Objectives:**

* Identify tools, libraries, and frameworks suitable for automating model selection and hyperparameter tuning.

**Key Steps:**

1. **Tools for Bayesian Optimization**:
   * Optuna, GPyOpt, scikit-optimize, hyperopt.
2. **Libraries for Model Training and Comparison**:
   * scikit-learn, XGBoost, LightGBM, TensorFlow, or PyTorch.
3. **Experiment Tracking**:
   * Use tools like MLflow or Weights & Biases to log and compare model performances.
4. **Automation Frameworks**:
   * Consider libraries like auto-sklearn, H2O.ai, or TPOT for automating parts of the workflow.
5. **Hardware and Computing Resources**:
   * For large-scale datasets, leverage GPUs or cloud-based services like AWS EC2 or Google Colab.

**4. Define Constraints and Assumptions**

**Objectives:**

* Identify limitations in terms of resources, time, and scope.

**Key Steps:**

1. **Resource Constraints**:
   * Computing power: Decide whether you’ll use local machines or cloud services.
   * Budget: Consider free-tier cloud services or open-source tools.
2. **Assumptions**:
   * Dataset size is manageable within the computational resources available.
   * Bayesian Optimization is appropriate for the problem (e.g., not overly complex hyperparameter space).
3. **Scalability Requirements**:
   * Will the pipeline need to scale for larger datasets or additional models in the future?

**5. Research and Literature Review**

**Objectives:**

* Explore existing solutions, methods, and best practices in automated model selection and Bayesian Optimization.

**Key Steps:**

1. **Study Bayesian Optimization**:
   * Understand concepts like acquisition functions (Expected Improvement, Upper Confidence Bound).
   * Learn about Gaussian Processes (GP) for modeling the objective function.
2. **Review Existing Work**:
   * Check research papers and case studies for insights on automating ML workflows.
   * Tools like Google Scholar, ArXiv, and Medium blogs can help.
3. **Identify Gaps**:
   * Highlight how your project addresses limitations in existing solutions.

**6. Scope Definition**

**Objectives:**

* Clearly outline the boundaries of what the project will achieve in Phase 1 and beyond.

**Key Steps:**

1. **What is In Scope?**
   * Automating model selection and hyperparameter optimization for common ML tasks.
   * Implementation of Bayesian Optimization.
2. **What is Out of Scope?**
   * Custom model development (like deep learning architectures).
   * Real-time optimization for streaming data.

**7. Deliverables for Phase 1**

**Expected Outcomes:**

1. **Problem Statement Document**:
   * Define objectives, constraints, and success metrics.
2. **Dataset Analysis Report**:
   * Summarize key characteristics, challenges, and preprocessing needs.
3. **Tool Selection Justification**:
   * List tools and libraries to be used, along with reasons for selection.
4. **Research Summary**:
   * Include findings from literature review and existing solutions.

**A VARIETY OF DATASETS**

**1. Time Series Datasets**

* **Time Series Classification Data**: Includes time-series data for various classification tasks. Useful for exploring sequential dependencies in data while testing model optimization techniques.  
  [Explore here](https://www.kaggle.com/datasets/modlee/time-series-classification-data)​

**2. Classification and Regression Datasets**

* **Collection of Classification & Regression Datasets**: A comprehensive dataset covering multiple use cases such as linear regression, logistic regression, and advanced classification tasks. Suitable for benchmarking different models.  
  [Explore here](https://www.kaggle.com/datasets/balakrishcodes/others)​

**3. Beginner-Friendly and Diverse Datasets**

* A curated list of datasets spanning classification, regression, visualization, CNN, NLP, and more. Great for practicing on various ML tasks and understanding optimization impact across problem types.  
  [Explore here](https://www.kaggle.com/discussions/getting-started/196714)​

**4. Image Data for Classification**

* **Image Classification & Clustering Dataset**: Useful for testing image classification and clustering models with visual data. Offers a step-by-step guide for leveraging advanced deep learning models.  
  [Explore here](https://www.kaggle.com/code/naim99/image-classification-clustering-step-by-step)​

**5. Pre-Trained Models with Associated Datasets**

* Pre-trained models and datasets available for classification, segmentation, and other tasks. These datasets can be integrated with pre-built models to evaluate hyperparameter optimization impact.  
  [Explore here](https://www.kaggle.com/models)​