Project Plan: Snowpack Prediction Challenge

Phase 1: Understanding the Problem and Data Exploration

1.1 Read and Understand the Challenge Requirements

- Review the problem statement and objectives.
- Identify key deliverables, including SWE predictions and performance metrics.

1.2 Data Collection and Understanding

- Access the dataset from the given directories:
  - **Meteorological Data:** Contains various weather parameters.
  - **SWE Data:** Provides daily SWE values for training and testing.
  - **Additional Test Locations:** For model generalization testing.
- Perform an initial exploratory data analysis (EDA):
  - Check for missing values.
  - Understand the distribution of variables.
  - Identify correlations between features.

Phase 2: Data Preprocessing

2.1 Handling Missing Values

- Identify missing values across datasets.
- Apply appropriate imputation techniques:
  - Mean/median for numerical variables.
  - Forward or backward fill for time-series data.

2.2 Spatial Association of SNOTEL Locations

- Associate each **SNOTEL** station to the nearest meteorological grid point.
- Use spatial distance metrics (e.g., Euclidean distance or k-NN) for mapping.

2.3 Combining Data Sources

- Merge meteorological, SWE, and static feature datasets.
- Generate a **final dataset**
- Normalize or standardize variables if required.

Phase 3: Feature Engineering

3.1 Time-Series Feature Engineering

- Create **lagged features** to capture historical trends.
- Compute **rolling averages** or moving windows for smoothing.

3.2 Derived Features

- Compute additional features such as **temperature variations** or **cumulative precipitation**.
- Explore incorporating external datasets (e.g., climate indices).

3.3 Dimensionality Reduction (Optional)

- Apply **Principal Component Analysis (PCA)** or feature selection techniques.

Phase 4: Model Development

4.1 Baseline Model

- Train a simple **Linear Regression or Random Forest** model as a benchmark.

4.2 Advanced Models

- Train and compare different machine learning and deep learning models:
    - **Gradient Boosting Models (XGBoost, LightGBM, CatBoost)**
    - **Recurrent Neural Networks (RNN, LSTM, GRU) for time-series forecasting**
    - **Convolutional Neural Networks (CNN) for spatial dependencies**
    - **Hybrid models combining ML and DL approaches**
- Tune hyperparameters using **Grid Search** or **Bayesian Optimization**.

Phase 5: Model Evaluation and Validation

5.1 Performance Metrics

- Compute key evaluation metrics:
  - **Nash Sutcliffe Efficiency (NSE)**
  - **Relative Bias (%)**
  - **Root Mean Square Error (RMSE)**
  - **Mean Absolute Error (MAE)**

5.2 Cross-Validation

- Split data into **train, validation, and test sets**.
- Use **k-fold cross-validation** for model robustness.

5.3 Visualizations

- Generate plots to compare **predicted vs. actual SWE values**.
- Visualize **SWE trends over time**.

Phase 6: Model Deployment & Testing

6.1 Prediction for Test Locations

- Apply the best-performing model to **additional test locations**.
- Generate time-series plots for SWE predictions.

6.2 Output Submission

- Save predictions in CSV format

Phase 7: Documentation and Report

7.1 Final Report

- Document all steps, methods, and decisions.
- Include:
  - Data preprocessing summary.
  - Feature engineering techniques.
  - Model selection process.

    - Performance evaluation results.

7.2 Code and Execution

- Prepare **a shell script (.sh) file** to run on **Kamiak**.
- Include necessary commands for model execution.

Phase 8: Project Review and Improvement

- Analyze model limitations and areas for improvement.
- Test additional **feature selection techniques or ensemble models**.
- Explore external datasets to enhance predictions.