**Project Title:** Spatial-Temporal Machine Learning Model for Weekly Temperature Prediction

**Project Description:** This project proposes the development of a machine learning and neural network-based model to predict weekly average temperatures across the state of Kentucky. The model will leverage 50 years of historical station data from the National Oceanic and Atmospheric Administration (NOAA), incorporating spatial features (latitude, longitude, elevation) of temperature stations and the recorded weekly average temperatures. The primary objective is to accurately predict weekly average temperatures for the year 2024, serving as a baseline validation of the model's predictive capabilities. The project will address challenges such as missing data imputation and the assessment of temperature reading normalcy to determine appropriate modeling and data normalization strategies.

**Problem Statement:** Accurate and localized temperature prediction is crucial for various sectors, including agriculture, energy consumption forecasting, public health preparedness, and climate change impact assessment. Existing temperature prediction models often operate at broader scales or rely on complex atmospheric physics, which can be computationally intensive and may not always provide the granular, localized insights needed for specific regional applications. The problem this project addresses is the lack of a robust, data-driven model specifically tailored to predict weekly average temperatures at a localized level within a given state or geographic region, utilizing readily available historical station data and spatial features. Such a model would empower stakeholders with more precise and timely temperature forecasts, enabling proactive decision-making and resource allocation.

**Background:** Traditional temperature forecasting has relied heavily on numerical weather prediction (NWP) models, which solve complex atmospheric equations. While highly sophisticated, these models require significant computational resources and often provide forecasts at a resolution that may not fully capture localized variations.

In recent years, machine learning (ML) and neural networks (NN) have emerged as powerful tools for time series forecasting and spatial prediction. Studies have demonstrated the effectiveness of ML/NN in meteorological applications, including temperature prediction. These approaches can capture complex non-linear relationships within the data without requiring explicit physical process modeling.

Previous works have explored various ML algorithms for temperature prediction, including:

* **Regression models**: Linear Regression, Support Vector Regression (SVR).
* **Tree-based models**: Random Forest, Gradient Boosting Machines (GBM).
* **Neural Networks**: Multilayer Perceptrons (MLP), Recurrent Neural Networks (RNNs) like LSTMs (Long Short-Term Memory) for handling sequential data.
* **Hybrid models**: Combining statistical methods with ML for improved performance.

Challenges commonly encountered in such projects include:

* **Missing data**: Gaps in historical records are common and require robust imputation techniques (e.g., interpolation, mean imputation, K-Nearest Neighbors imputation).
* **Non-stationarity**: Temperature data often exhibits trends and seasonality, requiring detrending or differencing techniques, or models capable of handling non-stationary time series.
* **Spatio-temporal dependencies**: Temperatures at nearby locations and previous time steps are highly correlated, necessitating models that can capture these dependencies.

This project aims to build upon these existing methodologies by specifically focusing on the Kentucky region and integrating spatial (latitude, longitude, elevation) and temporal (weekly average temperature) features to create a comprehensive predictive model.

**Methodology:** The project will follow a structured methodology encompassing data acquisition, preprocessing, model development, training, and validation.

**Data Acquisition:**

* Obtain 50 years of historical weekly average temperature station data for Kentucky from NOAA archives.
* Collect latitude, longitude, and elevation data for each identified temperature station.

**Data Preprocessing:**

* **Missing Data Handling**: Implement imputation techniques (e.g., linear interpolation, mean imputation, or more advanced methods like K-Nearest Neighbors imputation based on surrounding stations or temporal patterns) to address gaps in the historical temperature records.
* **Outlier Detection and Handling**: Identify and address anomalous temperature readings that could skew model training.
* **Feature Engineering**:
  + Extract cyclical features from the weekly data.
  + Potentially create lagged temperature features to capture temporal dependencies.
* **Data Normalization/Standardization**: Determine the normalcy of temperature readings (e.g., using statistical tests like Shapiro-Wilk) and apply appropriate normalization techniques (e.g., Min-Max scaling, Z-score standardization) to ensure optimal model performance.

**Model Selection and Development:**

* **Baseline Models**: Start with simpler regression models (e.g., Linear Regression, Ridge/Lasso Regression) and tree-based models (e.g., Random Forest Regressor, Gradient Boosting Regressor) as initial benchmarks.
* **Neural Networks**: Explore various neural network architectures:
  + **Multilayer Perceptron (MLP)**: For capturing non-linear relationships between spatial features, engineered temporal features, and temperature.
  + **Recurrent Neural Networks (RNNs) like LSTMs or GRUs**: For their ability to model sequential dependencies in the weekly temperature data.
  + **Convolutional Neural Networks (CNNs)**: Potentially explore 1D CNNs for extracting features from the time series data or 2D CNNs if spatial grids are created from the station data.
* **Ensemble Methods**: Consider combining predictions from multiple models to improve robustness and accuracy.

**Training and Validation:**

* **Data Splitting**: Divide the preprocessed historical data into training, validation, and test sets. The 2024 data will specifically be used for the final baseline validation.
* **Cross-Validation**: Employ techniques like K-fold cross-validation on the training data to tune model hyperparameters and assess generalization performance.
* **Evaluation Metrics**: Use relevant regression metrics such as:
  + Mean Absolute Error (MAE)
  + Root Mean Squared Error (RMSE)
  + R-squared (R2)
  + Mean Absolute Percentage Error (MAPE)

**Tools and Libraries:**

* **Programming Language**: Python
* **Data Manipulation**: Pandas, NumPy
* **Machine Learning Frameworks**: Scikit-learn, TensorFlow, Keras, PyTorch
* **Data Visualization**: Matplotlib, Seaborn
* **Geospatial Libraries**: Geopandas, Shapely

**Expected Outcomes:** The successful completion of this project is expected to yield the following results:

* **A robust and accurate machine learning model**: Capable of predicting weekly average temperatures across Kentucky using historical station data, latitude, longitude, and elevation.
* **Validated performance for 2024**: The model's ability to accurately predict weekly average temperatures for the year 2024 will serve as a strong baseline validation of its predictive capabilities.
* **Insights into influential factors**: The model's architecture and feature importance analysis may provide insights into which spatial and temporal factors most significantly influence weekly temperatures in Kentucky.
* **A reusable framework**: The developed methodology and code will provide a reusable framework for similar localized temperature prediction tasks in other regions or for different time scales.
* **Enhanced localized temperature forecasting**: The model's predictions can contribute to more precise and timely temperature forecasts for various applications, including:
  + **Agricultural planning**: Optimizing planting, irrigation, and harvesting schedules.
  + **Energy demand forecasting**: Assisting utility companies in managing energy production and distribution.
  + **Public health initiatives**: Preparing for extreme weather events and implementing public safety measures.
  + **Environmental monitoring**: Contributing to a better understanding of local climate patterns and changes.

This project will demonstrate the practical application of machine learning and neural networks in addressing real-world environmental challenges, providing a valuable tool for decision-makers and stakeholders in Kentucky.

**Team Roles:** The responsibilities of each team member will be determined by individual strengths and familiarity with the data. Some defined roles and descriptions may include:

* **Project Lead and Data Architect**: This role is responsible for the overall strategic direction, project management, and ensuring the technical architecture of the data pipeline and model is sound.
* **Data Scientist and Modeler**: This role will focus on the core machine learning aspects, including model selection, training, and evaluation.
* **Data Engineer and Preprocessing Specialist**: This role is crucial for acquiring, cleaning, and preparing the large historical dataset, ensuring data quality and availability for the modeling phase.
* **Research and Documentation Specialist**: This role will focus on the theoretical underpinnings, literature review, and comprehensive documentation of the project.