IMT 574 Final Group Project

Topic 6: Rainfall Forecasting for Weekend Planning

Kevin Ko, Yongxu Sun, Ke Yu, Bella Wei, Melody Chang, Haocheng Bao

1. Topic Introduction

This report explores the complex domain of weather forecasting, specifically honing in on predicting rainfall for an upcoming day. Our focus is on harnessing a variety of machine learning models to forecast rainfall occurrence, drawing upon historical weather data for analysis. The primary objective is to offer actionable insights into whether the specified day will encounter rainfall or remain dry. We aim to furnish decision-makers with accurate information crucial for scheduling activities reliant on favorable weather conditions. Ultimately, our goal is to furnish the necessary tools for weather forecasting and efficient contingency planning, thereby enhancing preparedness and adaptability in the face of changing weather patterns.

2. Significance Overview

Weather forecasting using machine learning (ML) is becoming increasingly crucial in today's society due to its profound impact on various sectors and everyday activities. Accurate weather predictions are essential for public safety, disaster preparedness, and resource allocation. ML-based forecasting models can analyze vast amounts of data, including historical weather patterns, satellite imagery, and atmospheric conditions, to generate more precise and reliable forecasts. This technology is particularly vital in regions prone to extreme weather events, where early warnings can save lives and minimize property damage.

Moreover, ML-powered weather forecasting has significant implications for

businesses across industries. In agriculture, where weather conditions directly affect crop growth and yield, accurate forecasts help farmers make informed decisions about planting, irrigation, and pest control. In the transportation sector, airlines, shipping companies, and logistics firms rely on weather forecasts to optimize routes, minimize delays, and ensure passenger and cargo safety. Additionally, retail businesses can use weather predictions to adjust inventory, plan promotions, and optimize staffing levels based on expected consumer demand influenced by weather conditions.

Furthermore, ML-based weather forecasting contributes to technological advancement by pushing the boundaries of data analysis and predictive modeling. The continuous refinement of ML algorithms enables more accurate and granular forecasts, leading to improved decision-making and risk management capabilities across sectors. By harnessing the power of ML in weather forecasting, society can enhance resilience to climate-related challenges, optimize resource utilization, and foster sustainable development in an increasingly dynamic and interconnected world.

3. Data Collection and Cleaning

3.1 Data Collection

We have four potential sources of data, namely: U.S. National Oceanic and Atmospheric Administration (NOAA), Climate.gov, OpenWeatherMap, and Kaggle dataset. After exploration, we found that OpenWeatherMap is the highest-quality data source for our prediction task. It allows us to access all weather data for Seattle from 1979 to the present through an API, with standardized data formats and minimal missing or dirty data. Therefore, we have chosen OpenWeatherMap as our data source.

3.2 Data Cleaning

To clean our dataset, we decided only to use relevant columns including 'dt_iso', 'temp', ‘feels_like', 'temp_min', 'temp_max', 'pressure', 'humidity', 'wind_speed', 'wind_deg', 'clouds_all', 'weather_main'. These columns provide us with information about the time of the record, the weather of the time, and features we want to use to predict future weather. Our decision to focus on these features stemmed from a comprehensive analysis of their relevance to weather forecasting. Through careful consideration, we recognized these variables as pivotal factors influencing weather patterns, thus warranting inclusion in our predictive model. We dropped all other columns such as "sea_level," "latitude," and "longitude," because they are less relevant to the prediction of the outcome. We also dropped rows with missing values to ensure the integrity of the data. By streamlining the feature set in this manner, we aimed to enhance the model's efficacy by focusing solely on the most influential predictors.

4. Model Selection

With the dataset containing a series of weather indicators and a weather label column (rain or no rain), we define our project as a supervised learning classification problem. Therefore, we have selected the following five classification models for prediction: Decision Tree, Random Forest, Support Vector Machine, Neural Network, and K-Nearest Neighbors.

4.1 Decision Tree

We chose the decision tree as one of our selected models because it can provide insight into feature importance by quantifying the impact of each feature on the prediction outcome. This information can be valuable for feature selection, identifying

the most influential variables, and understanding the underlying mechanisms driving predictions.

4.2 Random Forest

We also selected random forests to increase the resistance to overfitting. Compared to using a single decision tree, a random forest combines the outputs of many trees trained on different subsets of the data and features to achieve better generalization performance and are less susceptible to noise and variance in the dataset. It is also more effective in handling high-dimensional datasets with a large number of features.

4.3 Support Vector Machine (SVM)

Another machine learning model that is well suited for our weather prediction problem is the Support Vector Machine (SVM). SVMs are efficient in handling both linear and nonlinear relationships within the dataset. SVMs excel in identifying complex patterns and decision boundaries by maximizing the margin between different classes. In the context of weather forecasting, where the prediction of rain or no rain requires subtle distinctions in variables such as temperature, humidity, and wind speed, SVMs offer a robust framework for capturing these intricate relationships.

4.4 Neural Network

We also think neural networks can be useful in predicting the weather because their ability to learn hierarchical representations makes them well-suited for capturing intricate relationships and patterns with the complex features we have. By leveraging the flexibility and learning capacity of neural networks, we aim to develop a robust and adaptable model capable of accurately predicting weather outcomes.

4.5 K-Nearest Neighbors (KNN)

Our decision to use the k-Nearest Neighbors (KNN) is based on its simplicity, intuitive approach, and ability to capture local patterns within the data. KNN operates under the principle that similar instances in the feature space tend to belong to the same class. In weather forecasting, KNN offers a straightforward mechanism for leveraging historical data from neighboring instances to predict future weather conditions.

5. Data Pre-Processing

To train the suggested machine learning models mentioned in the section above with better predictive results and performance, several data preprocessing steps were performed on the original data provided by OpenWeatherMap. Several features were created and extracted from the original dataset to capture potential patterns and relationships in weather data.

5.1 Feature Engineering

The original dataset contained records of data from 1979-2024. Each row of records indicated that day's specific date and hour, meaning each day contained 24 rows of data. The dataset was grouped to create predictions at a daily interval instead of an hourly interval, resulting in a one-day, one-row structure. For each grouping of one day, the minimum, maximum, and mean of the following features was calculated: feels_like, temp_min, temp_max, pressure, humidity, wind_speed, wind_deg, and clouds_all. This increased the number of features, as we took the single value weather of that day and the range of that specific day.

5.2 Data Transformation

In this phase, the daily weather statistics were retroactively rolled back by three days for every data entry. Thus, each row of data encapsulated not only the weather status for the current day (such as whether it rained or not) but also the feature details from the preceding three days. This methodology allowed the supervised machine learning models to leverage insights from the weather data observed over the past three days to predict weather conditions. It's essential to highlight that while this experiment utilized a three-day rollback window, the window size can be adjusted for different scenarios or requirements.

6. Results

6.1 Performance Metrics

In this project, we have implemented a 70-30 split for dividing our dataset into training and testing sets. As the project aimed to conduct a binary classification task, we computed various performance metrics, including accuracy, precision, recall, and F1 measure, to evaluate and compare the models' performance, which are further explained below:

- Accuracy: The ratio of correctly predicted observations to the total observations.
- Precision: The ratio of correctly predicted positive observations to the total predicted positives.
- Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in actual class.huo
- F1 Measure: The weighted average of Precision and Recall.

For this project, we have prioritized the Recall metric in evaluating our model's performance. In weather prediction, particularly concerning rain, the consequences of false negatives (predicting no rain when it rains) are significantly higher than those of false positives (predicting rain that does not occur). A false negative could lead to inadequate preparation for rain-related conditions, potentially resulting in negative impacts on agricultural activities, urban infrastructure, and the general well-being of individuals unprepared for inclement weather.

6.2 Model Parameters

The parameters of the models is reported below for the purpose of future replication and testification:

- **Decision Tree**
    - Parameters: Default
- **Random Forest**
    - n_estimators: 200 (indicating 200 trees)
    - min_samples_leaf: 1
    - min_samples_split: 2
- **SVM (Support Vector Machine)**
    - Kernel: Linear
- **Neural Network (Multi-layer Perceptron)**
    - Hidden Layers: 2 layers with 2000 and 1000 nodes, respectively
    - max_iter: 14 (maximum number of iterations)
    - learning_rate: 0.01
- **K-Nearest-Neighbor (KNN)**

- K: 49 (number of neighbors)

6.3 Model Comparison

The results of the performance of the models is synthesized into the following table:

	Decision Tree	Random Forest	SVM	Neural Network	K-Nearest Neighbor
Accuracy	0.6421	**0.7513**	0.7008	0.7473	0.7206
Precision	0.6225	0.7415	0.6826	**0.7672**	0.7073
Recall	0.6081	**0.7237**	0.6803	0.6643	0.6927
F1 Score	0.6152	**0.7325**	0.7325	0.7121	0.7001

Table 1: Comparative Performance Metrics of the Model on Accuracy, Precision, Recall, and F1 Score

In this case, we will choose Random Forest as the best-performing model for its highest Recall Rate. The model also has the highest Accuracy and F1 score, which further confirms our choice.

7. Conclusion and Prediction

To further engage with our model, we've gathered data spanning from March 4th to March 6th, 2024. This dataset was used to predict weather conditions for our upcoming presentation day on March 7th, 2024. The prediction suggests that there will be no rain during the daytime (8 am to 8 pm) of the presentation day.

8. Limitations & Future Work

8.1 Limitations

Data Quality and Missing Information: The model's performance depends heavily on the quality and completeness of the historical weather data. Missing or inaccurate

data can impact the model's accuracy and reliability.

Short-Term Predictions: We only use three days of historical data for the fourth-day prediction, which may limit the model's ability to capture long-term weather patterns. While it performs well for short-term predictions, it may struggle with forecasting events over a more extended period.

Complexity of Weather Systems: Weather is a complex system influenced by various factors. We only use features provided in the OpenWeatherMap dataset while not considering features outside our dataset. Thus our model might not capture the full complexity of these interactions.

Changes in Climate Patterns: Climate change is an ongoing global trend. We use historical data from 1979 to the present, which may not accurately represent current climate conditions and is limited in making long-term predictions for future weather.

8.2 Future Work

Expand Model Prediction Scope: Collect data from more regions to generalize the model's prediction scope globally, making it more practically applicable.

Increase Lookback Days: Attempt to extend the data used for predictions from the current three days to a greater number, providing the model with more inputs to assess whether it can enhance accuracy.

Add Features: Explore additional datasets to introduce more dimensions of data, better addressing the complexity of weather systems.

9. Further Insights

Our weather prediction model can be further utilized in assisting with numerous outdoor research experiments as follows:

Experiment Planning: Our weather prediction model can be utilized to plan outdoor experiments. Knowing the expected weather conditions in advance can help researchers schedule experiments more effectively, considering factors like temperature, precipitation, and wind speed.

Agricultural Research: If the outdoor experiments involve agriculture or plant studies, our model can be used to plan planting and harvesting schedules, taking into account weather conditions that may affect crop growth.

Long-Term Trend Analysis: Our future work includes extending the model's predictions over a more extended period to analyze long-term weather trends. Researchers can study how seasonal variations impact their experiments and adapt protocols accordingly.

In summary, our rain prediction model not only facilitates efficient experiment planning and agricultural research but also offers insights into long-term weather trends. By incorporating our model into their research methodologies, students and scientists can optimize their experiments and adapt to changing environmental conditions, ultimately advancing their understanding in various fields.