College of Computers and Information Technology

Department of Computer Engineering

Taif University

Machine Learning Project Report

"Saudi Arabia Weather History"

Name : Renad Zaid Alosaimi

ID:44107056

supervised by: Dr. Nada Al-Tuwairiqi

Introduction:

Accurate weather prediction plays an important role in sectors like energy, farming, transportation, and public safety. In Saudi Arabia, where solar energy is a key part of Vision 2030, forecasting weather conditions like clear or cloudy skies is especially important for managing solar power and making informed agricultural decisions. In this project, we use historical weather data from 2017 to 2019 to train machine learning models that can classify general weather conditions. The goal is to explore how well machine learning can help automate weather condition prediction and support smarter decision-making. The dataset (Kaggle)

Dataset description:

We utilize the Saudi Arabia Weather History 2017 – 2019 dataset, which contains approximately 249,000 hourly weather records from January 2017 to April 2019 across 13 major cities. Each record includes the location (city), date and time, the observed general weather description, and several meteorological measurements. Key features provided are: City, Date, Time of each observation, Temperature, Wind Speed, Humidity, Barometer (atmospheric pressure in mbar), Visibility, and a textual Weather description. The data spans all seasons and covers cities from various regions. The Weather description is a categorical label describing general conditions. For this study, we focus on the most common broad categories of weather, we filter the data to include only Sunny, Clear, or Cloudy conditions as our target classes.

Data preprocessing:

The dataset required minimal cleaning since it was already in a structured format, we removed a small number of duplicate records and handled the humidity values, which were originally stored as strings with a "%" sign, the humidity strings were converted to numeric by stripping the percent symbol and casting to float, a few records had missing humidity, and these were dropped to ensure complete cases. After filtering to the three target weather classes, we also checked class balances across cities, to prevent location bias in our models, we applied a controlled down sampling: for each city, we randomly sampled at most 500 observations. After that we applied label encoding to convert the categorical city names and weather labels into numeric values for modeling.
In term of model training, and evaluation we split the dataset into training and testing sets (75% for training, 25% for testing), using a stratified sampling strategy to preserve the proportion of the three weather classes in each split, this ensures that each weather category (Sunny, Clear, Cloudy) is well-represented in both training and testing data. We then performed feature scaling on the input features using StandardScaler, Standardization transforms features like temperature, wind speed, to have mean 0 and unit variance. The scaler was fit on the training set features and applied to both training and test sets. The final feature set used for modeling includes: City, Month, Hour, Temperature, Wind Speed, Humidity, Barometer, and Visibility. (Date, year, day, and minute were not used as features since month and hour capture the essential temporal context for daily and seasonal patterns).

Modeling phase:

We trained seven different classification models to predict the weather condition (Sunny, Clear, Cloudy) from the features.
All models were implemented using scikit-learn with mostly default hyperparameters, in a few cases we adjusted settings to prevent training issues for instance, increasing the maximum iterations for the neural network and logistic regression to guarantee convergence, the goal was to use reasonable settings for each algorithm and then compare their performance on the classification task.

For each trained model, we generated predictions on the test set and evaluated the results using standard classification metrics, we computed the accuracy (overall percentage of correct predictions) as well as the precision, recall, and F1-score for each class.

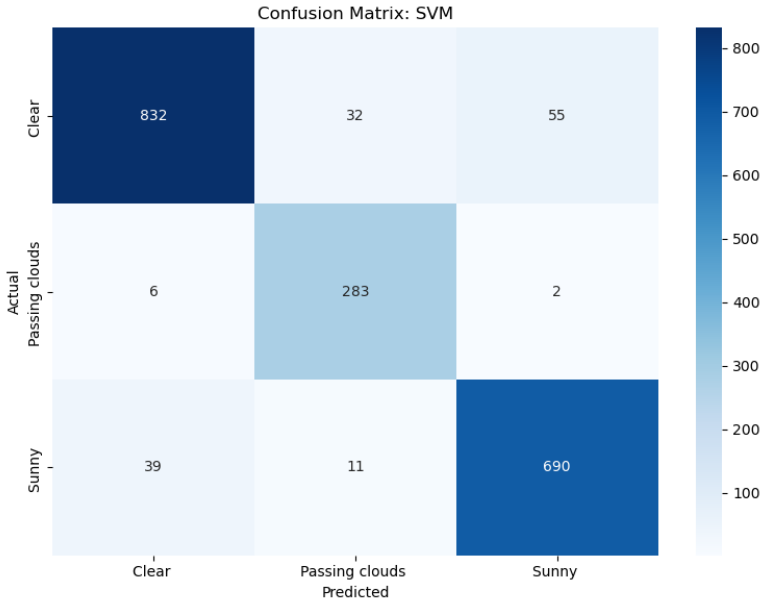
Models performance table

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
Decision Tree	94%	93%	93%	93%
Random Forest	96%	96%	95%	96%
KNN	89%	89%	89%	89%
SVM	93%	91%	94%	92%
Naive Bayes	84%	85%	86%	85%
Logistic Regression	73%	75%	78%	76%
ANN	95%	94%	94%	94%

After analyzing this table, we can see that the best-performing model is the Random Forest, achieving 96% accuracy and matching high macro-averaged scores (Precision, Recall, F1-score) all at 96% or 95%. This demonstrates its ability to generalize well and handle the multi-class classification effectively, leveraging its ensemble nature to capture complex feature interactions.
The Artificial Neural Network (ANN) scored 95% accuracy, with equally impressive macro metrics (94% across the board), indicating consistent strength across all classes. On the other hand, the worst-performing model is Logistic Regression, with only 73% accuracy, 75% macro precision, and a macro F1-score of 76%. This suggests that its linear decision boundary is too simplistic to capture the non-linear relationships between weather conditions and environmental features such as temperature, humidity, and time of day.
Models like SVM and Decision Tree also performed well (93% and 94% accuracy, respectively), while KNN and Naïve Bayes lag slightly behind due to their sensitivity to local variation and distribution assumptions.

Data visualization:

For the last section of the report, I want to mention that I used confusion matrix to better visualize how each model performed across the different weather classes, we used confusion matrices. These matrices offer a clear breakdown of correct and incorrect predictions for each class. All confusion matrices generated during the evaluation process are available in the /Results folder. Here is one example, **the confusion matrix for the SVM model**, which illustrates how well the model classified each weather condition.



Conclusion:

I selected this weather dataset because weather prediction is crucial in real-world planning, especially in countries like Saudi Arabia where solar energy is vital. Accurate classification of general weather conditions supports better management of solar power and helps farmers and authorities make timely decisions. The dataset spans multiple cities and years, offering a diverse and realistic view of environmental conditions. It includes key meteorological features such as temperature, humidity, wind speed, and barometric pressure, which allowed for rich analysis. Focusing on common weather states like Sunny, Clear, and Cloudy helped simplify the classification task while keeping it practical. The data required minimal cleaning, which made the modeling process more efficient. I performed label encoding and applied StandardScaler to make the feature values suitable for training. The models were then evaluated using precision, recall, F1-score, and accuracy to determine their ability to classify weather conditions correctly. Among all models, Random Forest performed the best, achieving 96% accuracy and strong balanced metrics across all classes. Its ensemble approach captured complex interactions between variables and avoided overfitting. The Artificial Neural Network came second with 95% accuracy and similarly strong results. Logistic Regression, on the other hand, performed the worst with only 73% accuracy, showing its limitations in handling non-linear patterns in the data. Visualization through confusion matrices helped identify where misclassifications occurred. Most errors were between Sunny and Clear labels, which often overlap in real conditions. One insight I gained is that humidity and the hour of the day are highly influential features in predicting weather. Down sampling cities also prevented any location bias from dominating model decisions. Overall, this project showed how machine learning can support weather forecasting and contribute to smarter, data-driven decisions in energy and agriculture sectors.

Project link on GitHub

https://github.com/renadzaid/Saudi-Arabia-Weather-History---Renad-Zaid-Alosaimi--44107056.git