Report of the project “Predicting weather data in Australia”

** Link to repository: https://github.com/mq38/Aus_weather_forecast*

The title of our project is “Predicting weather data in Australia”. The team members are Mihkel Paal and Laura Heleene Tirkkonen, geoinformatics Master’s students at Tartu University. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is used in this report to define the process framework for mining weather data, primarily from a business perspective.

1. Background

Weather forecasting is a practical application of meteorology, critical for many citizens, and even more so weather-sensitive industries like wind/solar farms and agriculture which rely heavily on predictions for effective operation. Farmers, for example, use forecasts for tactical decision-making to increase productivity while reducing risks caused by climate variability (Evans et al. 2019). Governmental meteorology institutions are primarily responsible for gathering weather intelligence from sources such as radar and satellite data to provide forecasts.

However, inaccuracies in observations and predictions can lead to late warnings or misjudged extreme weather events. As this information is used for decision-making and strategic planning, this can result in significant economic losses, at the worst in whole industries collapsing which is what has happened recently in Australia (The Nightly 2024). Bureau of Meteorology (BoM), the agency responsible for providing weather services to Australia, has faced criticism for issuing incorrect warnings, negatively impacting local farmers (AFR 2024). Which tools could the BoM utilize to make more accurate predictions?

Solution and measurement of success

A potential solution to improve forecasting accuracy would be to implement deep learning models using historical weather data, particularly for rainfall predictions. Standard statistical methods and existing atmospheric models often fall short in consistently achieving high levels of accuracy. Thus, the goal would be to improve country-wide and location-specific forecasts, enabling timely, reliable warnings and better guidance for industries and government planning. Any improvement counts as a win as forecasting, given the increasing unpredictability caused by climate change. With the implementation of these models, we expect to reduce yearly infrastructure losses and boost the profitability of weather-dependent industries over time. The weather agency might also be able to explore more cooperation opportunities from companies wanting to use their services.

Constraints of data and using machine learning tools

Challenges include the uneven distribution of weather stations which are more concentrated in certain areas, particularly the significantly colder northern part while there are significant gaps in the hotter central area of Australia (overview of weather stations of BoM). Predicting long-term trends might prove to be difficult as weather measurements from earlier times are not as reliable as nowadays, reducing forecast confidence over extended time frames, and effectively not being as useful for weather institutions.

2. Assessment of situation

Team and resources needed

The project requires data scientists with expertise in business processes involving machine learning and predictive modeling of weather-related data. They also need to be able to evaluate the results and understand existing methods of weather modelling. Additionally, meteorologists should be involved to provide domain-specific insights on weather patterns, and inform the data scientists on issues of climate change. Lastly, a project manager is needed to coordinate the team and project timeline.

The primary dataset includes historical weather observations (daily temperature and rainfall), supplemented by climate patterns. High-performance computing servers are essential in keeping computing times at a minimum, given that weather forecasts need to be issued every day and often with hourly accuracy. Python is used as a universal and easy to understand programming language. The most important libraries used are TensorFlow, scikit-learn, and (geo)pandas. Jupyter Notebooks helps visualize data and make maps to show the geographical variance and extent of weather phenomena.

Assumptions and risks

The project assumes consistent access to Australian weather data, limited to a specific time period to preserve comparison. Erroneous data points and gaps must be addressed in that we need to decide to leave them out or rather preserve real-life conditions with potentially false measurements. The data scientists are responsible for implementing robust data cleaning and imputation methods. Maintaining detailed documentation is critical to ensure transparency, reproducibility and to find shortcomings in the model. The timeline, requirements for the finished work as well as financial and legal restrictions are omitted from this report as these are dependent on agreements with our partner.

Costs and benefits

Major costs of the project include personnel (salaries for each team member), hardware (purchasing and maintaining servers) as well as software (licenses). However, the expected benefits far outweigh the expenses as improved weather prediction accuracy leads to better disaster preparedness which ultimately has the potential to deliver substantial long-term savings in reduced infrastructure losses and more efficient planning in the first place. Additionally, in terms of agriculture and water management, resources can be allocated more economically.

Terminology

It is of importance to define a few terms which will be used when publishing the results of the project.

1. **Climate and weather**, the former referring to the weather of a specific region averaged over a long period of time (typically 30 years) and the latter to short term atmospheric conditions which can be measured daily or even hourly.

2. **Data accuracy** which we define as a collection of different metrics that help evaluate how often a machine learning model correctly predicts the outcome - in this case the specific measurements and their change in time
3. The **sliding windows method** which involves dividing the data into overlapping windows of a fixed size, and processing each window independently (Data Overload 2022). This helps to preserve the context or relationship between adjacent data points which is undoubtedly relevant in weather data as the conditions of one day affect the conditions on the next day, and so on.

3. Defining data-mining goals

The primary objective of this project is to build a machine learning model capable of accurately predicting critical weather parameters in Australia, focusing on rainfall and temperature variability.

Key goals for the project

1. Prediction of rainfall and temperature trends

The most important parameters for the model to learn to predict are in Australia's climate rainfall and other connected parameters - there are extended periods of dry weather and temperature variability as well as a high frequency of weather phenomena such as bushfires (article on 2019/20 bushfire season). Thus, we aim to develop a machine learning model (or multiple) to forecast rainfall and potentially even daily rainfall amounts, accounting for temporal and spatial variability. We also want to predict daily minimum and maximum temperatures with a high accuracy.

2. Filling spatial gaps with data

We want to ensure consistent model performance across different areas of Australia with diverse climate zones, including under-monitored areas like central regions where there are fewer weather stations.

3. Model performance metrics

We want to achieve at least 80% accuracy (as well as precision and recall) for key predictive tasks, evaluated using metrics like RMSE as well as precision and recall. To achieve these goals, we will deliver trained machine learning models for rainfall and temperature prediction, cleaned and preprocessed datasets ready for analysis and modelling, geospatial prediction maps and trend analysis graphs, and lastly, comprehensive documentation and results summaries for stakeholders, highlighting model performance and recommendations for future work.

Success criteria

We want to achieve model accuracy of at least 80% for predictive tasks and reduce RMSE for rainfall forecasting by at least 5% compared to existing methods. We wish to demonstrate spatial consistency by testing on diverse Australian climate regions with acceptable performance variance. The developed models should additionally integrate seamlessly with existing operational workflows of BoM and could be used together with statistical modelling of weather.

4. Gathering data

Outlining data requirements
The objective of the project is to perform data mining on Australian weather data to identify patterns, trends, or predictive models related to weather observations. This requires a dataset with features such as temperature, precipitation, wind speed, and weather conditions over time and across diverse locations in Australia.

Verifying data availability
The dataset sourced from Kaggle comprises around 10 years of daily weather observations from over 40 locations in Australia. It includes key meteorological variables like minimum and maximum temperatures, rainfall, humidity, wind speed, and weather conditions (e.g., sunny, rainy). These variables align with our data requirements for predictive modeling and exploratory analysis.

Defining selection criteria

- **Timeframe**: Data spanning 10 years is sufficient to capture seasonal variations, long-term trends and short term weather fluctuations.
- **Geographic scope**: The inclusion of data from over 40 locations ensures spatial diversity, capturing different climatic zones within Australia (e.g., coastal, desert, tropical).
- **Variables of interest**: Initial selection includes:
    - Minimum and maximum temperature
    - Rainfall (today)
    - Date and location

5. Describing data

The dataset contains approximately 142,193 rows and 24 columns. Each row represents a daily observation for a specific location, with the following notable fields:

- **Date**: Records the day of observation.
- **Location**: Specifies the observation station or city.
- **MinTemp** and **MaxTemp**: Daily minimum and maximum temperatures in degrees Celsius.
- **Rainfall**: Measured in millimeters.
- **WindGustSpeed**: Maximum wind gust speed recorded in km/h.

- **Humidity9am** and **Humidity3pm**: Relative humidity at 9 a.m. and 3 p.m. as percentages.
- **RainToday** and **RainTomorrow**: Binary indicators for rain occurrence.
- **Weather Conditions**: Includes cloud cover, pressure, and wind direction data.

Initial exploration shows a mix of numerical, categorical, and time-series data types.

6. Exploring data

Preliminary exploration reveals the following insights:

- **Missing values**: A substantial number of missing values exist in fields like `Rainfall`, `WindGustSpeed`, and `Humidity3pm`. Missing data is not uniformly distributed and may reflect incomplete observations for specific stations or days.
- **Data Distribution**:
  - Temperature values follow a roughly normal distribution but vary significantly by location.
  - Rainfall is positively skewed, with many zero-rainfall days.
  - Wind speed and humidity values show notable seasonal patterns.
- **Temporal trends**: Clear seasonal variations are observed, particularly for temperature and rainfall, reflecting Australia's diverse climate zones.
- **Outliers**: Some extreme values (e.g., unusually high rainfall or wind speeds) warrant further investigation.
- **Correlations**: Early checks indicate a strong negative correlation between temperature and humidity, as expected. Rainfall shows weak to moderate correlations with humidity and cloud cover.

7. Verifying data quality

- **Accuracy**: The dataset originates from reputable meteorological sources (Australian Bureau of Meteorology via Kaggle), but individual anomalies, such as outliers or unrealistic values, will need review.
- **Completeness**: Several variables have missing values:
  - `Rainfall`: ~10% missing
  - `WindGustSpeed`: ~9% missing
  - Other variables like humidity and cloud cover also exhibit sporadic gaps.
  - Missingness patterns need analysis to determine if they are random or systematic.

- **Consistency**: Units are consistent across all locations (e.g., temperatures in Celsius, rainfall in mm). However, weather condition labels may need standardization.
- **Timeliness**: Data spans from 2007 to 2017, covering a relevant and recent timeframe for analyzing current weather patterns.

Next steps

Based on this understanding:

1. **Useable data**: Fields like `Date`, `Location`, `MinTemp`, `MaxTemp`, `Rainfall`, and `RainTomorrow` are key candidates for modeling and analysis.
2. **Field understanding**: Each selected variable has been reviewed for its role in analysis, ensuring clarity of meaning.
3. **Data cleansing**: While extensive cleaning belongs to the data preparation stage, initial steps include:
    - Addressing missing values using imputation or exclusion strategies.
    - Investigating and handling outliers.
    - Standardizing categorical data.

8. Project plan

We have defined five main tasks which make up our project plan.

1. Ensuring the dataset is accurate, complete, and ready for analysis by cleaning it
    a. **Inputs**: Raw weather data from all weather stations.
    b. **Outputs**: Cleaned and preprocessed dataset.
    c. **Description of process**: We remove duplicates and handle missing values, scale the data and handle outliers.
    d. **Requirements**: Accessing raw data from Kaggle and having access to tools such as Python libraries.
    e. **Timeline**: One week.
2. Using sliding window prediction while training models to predict rainfall
    a. **Inputs**: Cleaned dataset from previous task.
    b. **Outputs**: Machine learning models with optimized parameters.
    c. **Description of process**: First, we have to define the sliding window features (the past 7 days), after which we can train models (i.e. logistic regression), calibrate hyperparameters and iterate until we are satisfied with the results.

    d. **Requirements**: Completion of data cleansing.

    e. **Timeline**: Two weeks.

3. Creating maps based on predicted values using geospatial tools

    a. **Inputs**: Predicted values from the machine learning model and spatial data (Australian territory, weather stations).

    b. **Outputs**: Prediction maps for rainfall and temperature.

    c. **Description of process**: We have to first geocode the weather station data for the predicted values to be location-specific. Then we apply interpolation techniques for the areas in between the stations to also have values, even if there are none there. Lastly, we generate maps with the help of geospatial plotting libraries.

    d. **Requirements**: Completion of model training.

    e. **Timeline**: 2 days.

4. Creating an overview of the climate during the observation period by analyzing and summarizing historical climate trends

    a. **Inputs**: Cleaned dataset and predictions, additional research from external sources.

    b. **Outputs**: Summary of trends and descriptive statistics of weather parameters.

    c. **Description of process**: We calculate the mean, median, and standard deviation for climate variables during the observations period and visualize the trends with graphs. We additionally highlight anomalies or significant changes over the period.

    d. **Requirements**: Completion of data preprocessing from step 1.

    e. **Timeline**: One week.

5. Creating a report and evaluating the results and model performance

    a. **Inputs**: All project outputs, including maps and models.

    b. **Outputs**: Final project report and presentation.

    c. **Description of process**: Comparing the accuracies of different models with metrics like RMSE and precision. We additionally assess the strengths, weaknesses, and areas for improvement for the models.

    d. **Requirements**: Completion of all previous steps.

    e. **Timeline**: 5 days.