Wheat Yield Prediction in Germany (1950-2023)

Introduction

This project explores the relationship between climate factors and wheat yield in Germany from 1950 to 2023. The goal is to build a predictive model that estimates wheat yield (tons/hectare) based on climate data. By understanding how factors like temperature, precipitation, and sunshine influence yield, this project aims to provide insights that could guide agricultural planning, especially in the face of climate change. The final model, developed using a Random Forest algorithm, aims to predict wheat yield for any location and period in Germany.

Column Descriptions & Relevance

- Year: The year when the data was recorded. Relevant to identify time trends in wheat yield and climate changes over decades.
  Reference for normalization: https://onlinelibrary.wiley.com/doi/full/10.1002/fes3.372
- Wheat yield: Average wheat yield (tons per hectare), normalized to account for improvements in agricultural practices. The primary target variable for the prediction model. *Reference*: Placeholder for reference on agricultural yield trends.
- days_above_30degrees: Number of days with temperatures exceeding 30°C during the wheat growing cycle (March-September). Relevant for understanding the impact of extreme heat on crop stress and yield. *Reference*: https://link.springer.com/article/10.1007/s13593-017-0443-9
- evapotranspiration average: Average evapotranspiration rate (mm/day) during the growing cycle. Important as it represents water loss through evaporation and plant transpiration, affecting soil moisture and crop growth. *Reference*: https://www.frontiersin.org/journals/sustainable-foodsystems/articles/10.3389/fsufs.2023.1203721/full
- precipitation total: Total precipitation (mm) during the growing cycle. Essential for assessing water availability for crops, influencing yield. *Reference*: https://www.frontiersin.org/journals/sustainable-foodsystems/articles/10.3389/fsufs.2023.1203721/full
- sunshine: Average daily sunshine hours during the growing cycle. Sunshine is a critical factor for photosynthesis, which directly affects plant growth and yield. *Reference*: https://link.springer.com/article/10.1007/s13593-017-0443-9
- sunshine total: Total sunshine hours during the growing cycle. A cumulative measure that helps understand the total energy available for crop growth. *Reference*: https://www.frontiersin.org/journals/sustainable-foodsystems/articles/10.3389/fsufs.2023.1203721/full
- Growing degree days: Cumulative growing degree days (°C) during the wheat growing cycle. This variable tracks the accumulation of heat required for crop development, crucial for understanding growth stages. *Reference*: https://link.springer.com/article/10.1007/s13593-017-0443-9
- SPI: Standardized Precipitation Index (SPI) during the growing cycle, used to indicate drought conditions. Drought can severely affect crop yields, making this an important drought indicator. *Reference*: https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-index-spi

Dataset Description Table

Column Name	Description	Source
Year	Year of data recording	
Wheat_yield	Average wheat yield (tons per hectare) that year in Germany.	https://wwwgenesis.destatis.de/genesis/online?operation=statistic&levelindex=0&levelid=1724407549377&code=41241#abreadcrumb
days_above_30degrees	Total number of days with temperatures above 30°C during the wheat growing cycle	https://open-meteo.com/en/docs/historical-weather-api
evapotranspiration_average	Average evapotranspiration rate (mm/day) during the growing cycle	https://open-meteo.com/en/docs/historical-weather-api
precipitation_total	Average daily sunshine hours during the wheat growing cycle	https://open-meteo.com/en/docs/historical-weather-api
sunshine_avg	Total sunshine hours during the growing cycle	https://open-meteo.com/en/docs/historical-weather-api
growing_degree_days	Cumulative growing degree days (°C) during the wheat growing cycle	
SPI	Standardized Precipitation Index (SPI) indicating drought levels during the growing cycle	

Approach

1. Data Collection: Climate data was obtained from the German Meteorological Data API, and wheat yield data was sourced from German agricultural statistics. The dataset includes daily weather data aggregated to wheat growing cycles.
2. First Jupyter Notebook:
    - Data Cleaning: Missing values were handled, and outliers were removed.
    - Aggregation: Daily weather data was aggregated into growing season metrics (e.g., total precipitation, average sunshine) and combined with wheat yield data.
    - Normalization: Wheat yield data was normalized to account for improvements in agricultural practices over time.
3. Second Jupyter Notebook:
    - Data Analysis: Basic statistics and data visualizations were created to explore relationships between climate variables and wheat yield.
    - Model Building: A Random Forest model was built to predict wheat yield using climate data.
    - Model Evaluation: The model was evaluated using metrics like Mean Absolute Error (MAE) and Accuracy.

Conclusion

Thanks to the normalization of the wheat yield data, the Random Forest model achieved a decent result with an MAE of 2.94 and an accuracy of 90.66%. Although the model was trained on a relatively small dataset, it provides a solid starting point for predicting wheat yield based on climate factors. The model may be overfitted due to the limited data, but it lays the groundwork for further refinement and the development of a tool to calculate wheat yield per hectare given climate conditions. The project's scope has been successfully attained.

Improvements & Future Directions

To enhance the predictive capabilities of this model and make it more applicable in real-world agricultural planning, several improvements and future directions can be pursued:

Finer-Scale Data Collection:

Current Limitation: The dataset used in this project relies on averages from 6 randomly selected locations, which limits regional specificity and forces normalization to account for technological improvements over time.

Improvement: Gathering climate and yield data at the level of individual hectares, not just in Germany but from multiple countries, could eliminate the need for normalization. By capturing granular data across diverse geographic locations, the model could better account for regional differences in soil quality, farming practices, and microclimates, leading to more accurate and localized predictions.

Integration of Technological Advances:

Current Limitation: The current model normalizes yield data to account for advancements in agricultural technology (e.g., better fertilizers, improved seed varieties, precision farming).

Improvement: By directly incorporating variables related to technology (e.g., types of fertilizers used, irrigation methods, machinery employed), the model can distinguish between yield improvements due to climate factors and those resulting from technological advancements. This would allow for more precise modeling of the relationship between climate and yield without relying on normalization.