Environmental Data Mining

Quinton Randall, Skye Gerald

Grand Valley State University

Environmental Data Mining

Our project centers on the advanced domain of Environmental Data Mining, and we have examined the weather patterns between Grand Rapids and Allendale in particular. We did this to gather relevant data for our investigation, we carefully looked for information on high and low temperatures, midday temperatures, humidity levels, wind direction and speed, and precipitation cumulation events. This investigation was conducted as a tactical move to get important information on the changing weather patterns, and conditions encountered in this location area. This is to show how different weather patterns can be, even when in close proximity. Through this extensive analysis, we want to clarify the small differences in weather between Grand Rapids and Allendale and offer a deeper understanding of the atmospheric characteristics that define the relationship between spaces. Our project is more than just gathering data, it's a complicated motion with environmental details that tells a story about climate trends. The search that we undertake captures the essence of Environmental Data Mining, from the ongoing movement of wind and precipitation to the fluctuating pattern of high and low temperatures. By doing this, we want to provide insightful information that will benefit future generations of students, and others not familiar with northern weather, as we are both from places closer to the equator.

Our primary goal is to successfully complete this assignment, and our project's topic doesn't align with our career paths or personal interests. It's essentially a class requirement, and while we were allowed to choose the topic, there isn't a strong intrinsic motivation beyond fulfilling the course requirements.

In this Environmental Data Mining project, we are taking a methodical and thorough approach to investigating weather trends between Grand Rapids and Allendale. In order to gather important meteorological data for our study, we first carried out a comprehensive search. We paid close attention to important factors including high and low temperatures, midday temperatures, humidity levels, wind direction and speed, and precipitation. By cross-referencing data from many reliable meteorological sources, we used a strict testing procedure to guarantee the correctness and dependability of our dataset. This solid approach is the basis for the quality of our results. By utilizing correlation data-mining techniques, and cluster data, we were able to find different patterns, and trends within the data that has been collected. The observation that Grand Rapids exhibits a higher high temperature and a lower low temperature than Allendale suggests a climatic distinction between these two locations. A higher high temperature in Grand Rapids implies that, on average, the maximum daily temperatures tend to be greater in comparison to Allendale. Conversely, a lower low temperature indicates that Grand Rapids experiences, on average, colder minimum daily temperatures than Allendale.

Prior research in the field of environmental data mining has explored many strategies and uses that provide information relevant to our topic.  Researchers have looked at a variety of approaches, including decision trees, support vector machines, and clustering algorithms for the study of environmental data. These different strategies offer a wide range of techniques for identifying patterns and trends in environmental information. Furthermore, related research in urban planning and environmental impact assessments has applied data-driven methods, highlighting the usefulness of mining environmental data for decision-making procedures.

We choose to collect data from many online weather sources as part of our data gathering strategy. The accessibility and quantity of data provided by these platforms which provide a thorough picture of weather conditions were the primary factors in our selection. Utilizing these resources allowed us to use data mining techniques such as clustering, and correlation strategies. Our selection of online weather sources gave us an easy way to get both historical and current weather information, enabling a dynamic and current study. These websites usually include a variety of data, such as changes in humidity, precipitation amounts, wind patterns, and other weather aspects that are essential to comprehending the complexity of the environment. Although there are useful benefits to this data gathering strategy, it's vital to be aware of any potential drawbacks, such as differences in data quality and dependability between platforms. Ensuring the quality and integrity of our dataset will need highly organized verification and cross-referencing of the gathered data. Despite these factors, using internet weather portals is a wise decision as they offer a wealth of data for our thorough investigation and prediction undertakings.

Our data mining pipeline for extracting and analyzing weather data between Grand Rapids and Allendale is designed for precision, and comparison. The following steps outline our comprehensive approach:

* **Data Collection:**

We started the process by gathering information from several dependable online weather resources. This involves gathering data on humidity levels, wind direction and speed, precipitation, high and low temperatures, and noontime temperatures.

* **Data Cleaning and Preprocessing:**

Strict cleaning techniques are used to deal with inconsistent or missing data. To guarantee the quality and integrity of the data, calculation, normalization, and validation are required. For consistency, we additionally transform dates into a uniform format.

* **Verification and Cross Referencing:**

We cross-reference data from several weather sources in order to undertake validation tests and improve the dependability of our dataset. Inconsistencies are carefully resolved, and information from sources that exhibit errors is marked for additional examination.

* **Exploratory Data Analysis (EDA):**

EDA is used to acquire early dataset insights. In order to identify trends and possible outliers, statistical summaries, visualizations, and correlation studies are useful tools for informing later modeling choices.

* **Interpretation and Insight Generation:**

The final step involved interpreting the results and generating meaningful insights. We explore correlations between weather statistics and identify trends that contribute to a deeper understanding of the environmental dynamics between Grand Rapids and Allendale.

Our data mining pipeline guarantees transparency, repeatability, and dependability in the extraction, processing, and analysis of weather data for the Grand Rapids - Allendale area by carefully recording and reproducing these stages.

The performance and dependability of the selected forecasting models were evaluated in large part thanks to the model assessment analyses carried out for our project. Important assessment metrics were calculated to express the degree of accuracy and consistency of the forecasts. These metrics included statistical measurements like mean, standard deviation, correlation, and clustering. Cross-validation techniques were applied to validate the models' robustness over the month.

We prioritized straightforward communication with the use of excellent visuals that provide insights into our forecasting models' performance. The models' ability to accurately represent the dynamics of environmental data between Grand Rapids and Allendale may be easily understood through periods of time visualizations of projected vs actual values. By incorporating visual elements such as line charts, cluster graphs, and coloration charts. Our goal is to improve the findings' understanding and aid in well-informed decision-making. These visualizations serve as a powerful means to communicate the strengths and limitations of the forecasting models, contributing to a comprehensive understanding of the project's outcomes.

The project's accuracy and dependability are significantly relying on the quality of the information gathered from various internet weather sources. Effects might be introduced, and the accuracy of the results could be impacted by inconsistencies, errors, or missing data in the sources. Due to differences in reporting standards, equipment, or measuring techniques, different weather websites may offer inconsistent data. This range of data may make it more difficult to build a trustworthy dataset.

A graph showing the temperature of a person

Description automatically generated with medium confidenceFuture work may incorporate other environmental factors, such as air quality measures, and investigate localized changes in the weather using a spatial study utilizing GIS data. Comprehensive insights might be obtained by expanding the study to include a long-term trend analysis and working with urban planners to integrate data into municipal planning projects. Accuracy might be improved by investigating cutting-edge machine learning methods, refining forecasting models, and taking collaboration data integration into account. A more dynamic environmental monitoring system may benefit from the development of user-friendly visualization tools and the comparison of current weather patterns with past events. Implementing real-time weather monitoring and working with climate experts would guarantee the project's continued relevance.

A screenshot of a graph

Description automatically generatedA diagram of different colors

Description automatically generatedA diagram of different colored dots

Description automatically generatedA graph showing different colored dots

Description automatically generatedA graph showing different types of weather

Description automatically generatedA graph of a graph

Description automatically generated with medium confidenceA graph and chart with text

Description automatically generated with medium confidence

The software used was what was presented in class, “Google Colab”.

https://github.com/GVSU-CIS635/gvsu-cis635-term-project-weather

*Conclusion*

Our project primarily focuses on examining weather patterns between Allendale and Grand Rapids. The project findings reveal a climatic distinction between Grand Rapids and Allendale, with Grand Rapids experiencing higher high temperatures and lower low temperatures on average.  The project's accuracy and dependability are significantly relying on the quality of the information gathered from various internet weather sources. Lots of data regarding weather can often be misleading, or inaccurate when analyzing many different sources. The goal was to provide insight into the differences of weather patterns that can happen between cities that are close together. There was a decent amount of data collection, filtering, analyzing, and analysis. The chosen weather sources offer accessibility and quantity information. However, potential drawbacks, such as varying data quality, require lots of verification. Performance evaluation involves model assessment analyses, considering metrics like mean, standard deviation, correlation, and clustering. Visualizations, including line charts and cluster graphs, aid in communicating forecasting models' strengths and limitations.

*References*

C. Kunjumon, S. S. Nair, D. Rajan S., P. Suresh and S. L. Preetha, "Survey on Weather Forecasting Using Data Mining," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, India, 2018, pp. 262-264, doi: 10.1109/ICEDSS.2018.8544326.

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? An assessment of data mining methods in Environmental Data Science. Environmental modelling & software, 110, 3-27.

Hannachi, A. (2021). Patterns identification and data mining in weather and climate (Vol. 10). Cham: Springer.

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., & Canu, S. (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environmental Modelling & Software, 19(9), 845-855.