Analysis of Australian weather data using Apache Storm

O. Grimes[1], Q.T. Trinh[1], V. Vaid[1], S. Menon[1], R. Watson[2] and P. Ryan[2]

[1]Swinburne University of Technology, Hawthorn VIC 3122, Australia
[2]Ryan Watson Consulting Pty Ltd, Vermont VIC 3133, Australia
Email: richard@ryanwatsonconsulting.com.au

Abstract: The science of weather forecasting has come a long way since the first use of scientific instruments to measure barometric pressure and other parameters in the mid-19th century. In those days reasonably accurate forecasts could be made of the next-day's weather in a relatively small region like the British Isles. In the years since then, meteorological data collection and analysis have become more sophisticated and weather forecasting has become a multi-national collaboration able to accurately forecast the weather over much of the planet up to 10 days ahead, and make longer-term predictions of wet or dry seasons. This has made a big difference in sectors like agriculture, tourism, water resources, aviation and fire prevention.

Weather data includes humidity, temperature, rainfall, radiation, wind speed, air pressure, sunlight strength, etc. and is produced in real-time. Big data analytics is the science which has been developed to handle such data and extract value and insights from it (Fathi et al 2021). Modern weather forecasting produces petabytes of data every day and professional weather forecasting utilises high performance computing systems with distributed storage. However smaller forecasting systems which can run on desktop computers can still produce valuable information. This paper describes a two-semester project carried out by a team of Swinburne University students to forecast whether fog or haze is expected in Australian capital cities based on various parameters including temperature, humidity, wind speed and air pressure.

The project used open-source software, and the Swinburne team, in consultation with the project sponsor, Ryan Watson Consulting Pty Ltd, chose to use the Apache Storm framework (Kumawat 2020). This can ingest real-time data from a variety of sources such as the Apache Kafka publish-subscribe messaging system. The Apache Cassandra No-SQL database system was also used. Apache Storm is highly scalable, fault tolerant and user-friendly, and can be used in small as well as large organisations.

The input weather data was analysed using the python machine learning library scikit-learn, with the Support Vector Machine classification algorithm the chosen option. This calculates a decision boundary, which is a hyperplane between classes which can be linear or non-linear. The classes for outputting predictions were positive haze, positive fog, or both negative haze and negative fog.

The students were required to produce the full range of project documentation including a software design and research report, quality assurance plan and test plan. The sponsor provided a virtual cloud server to enable all members of the team to access and work on project files and documents. The project met all its objectives, including the production of an app to predict the probability of fog or haze in Australian cities selected from a pull-down menu.

REFERENCES

Fathi, M., Haghi Kashani, M., Jameii, S.M., Mahdipour, E., 2021. Big Data Analytics in Weather Forecasting: A Systematic Review. Arch Computat Methods Eng . https://doi.org/10.1007/s11831-021-09616-4

Kumawat, D.C., 2020. Apache Storm architecture: Real-time Big data analysis engine for streaming data. https://medium.com/analytics-steps/apache-storm-architecture-real-time-big-data-analysis-engine-for-streaming-data-4fc34ce0adae

***Keywords**: Weather forecasting, Big data, Apache Storm, Apache Kafka, Publish-Subscribe, Apache Cassandra*