Big Data Analytics for 5G Self-Healing

Ibrahim Naffaa, Electrical Engineering

Mentor: Dr. Tamer Omar

Kellogg Honors College Capstone Project

Abstract

In modern times, various types of structured and unstructured data are produced and collected in massive quantities due to the increased connectivity of our society. With such a large amount of data being collected, it is both beneficial and necessary to create infrastructure to analyze "Big Data" and finding meaningful results. In the case of a 5G self-healing simulator developed by Cal Poly Pomona's Wireless Network Security Lab (WNSL), data about the simulated 5G networks are generated in the order of gigabytes with the potential to rapidly increase over time. Since it is essential to analyze the data produced by the 5G simulator, Amazon Web Services (AWS) is utilized to rapidly create a big data system where big data analytics can be performed. Due to the complexity of 5G, various parameters can affect the performance of the network including data rates, antenna sectors, and transceiver angles, which results in massive data sets, but not all data is necessary for our analytics. For this project, an algorithm was developed with Amazon Athena to reduce our data set for analysis using other tools on AWS.

Background

Our decision tree is generated using data from a 5G Self-Healing Simulator that is designed to simulate a 5G network under various conditions that ultimately affect the network's ability to provide the demanded data rates to its users. Depending on the simulation parameters of the network and the empirical values obtained from 5G literature, the simulator can model a network over time.

5G Terminology

- **Base Station (eNodeB):** A cellular tower with 5G capabilities.
- **Antenna:** A component within a base station that provides coverage for users within a given sector of the base station.
- **User Equipment:** Any device connected to the 5G network.
- **Transceiver:** A transmitter or receiver that is present on an antenna or on user equipment. It is used to both transmit and receive data between devices.
- **Data Rate:** Defined as the amount of data delivered in bits per second.

Results

Generating 5G Data Set

The data set generated from the 5G simulator is modeled as seen in Fig. 1 where hexagons representing base stations are arranged in a honeycomb pattern as shown. Base station 0 was set to "failed", meaning that the base station is providing no data, while base stations 1 and 4 are set to "congested", meaning that some data (but not all data) is provided.

Combining Data Sets using Amazon Web Services

In order to consolidate the data generated from the 5G simulator, Amazon Web Services (AWS) was used. Specifically, Amazon Athena combined all data into a single csv file that could be used to analyze our data using Python.

Creating Decision Tree Using Python

The decision tree seen in Fig. 2 was generated using Python as well as the sklearn, pandas, and pydotplus machine learning libraries. The decision tree shown had 90% accuracy with the data set that was used for testing.

Fig. 1: 5G Network Model for Training Decision Tree

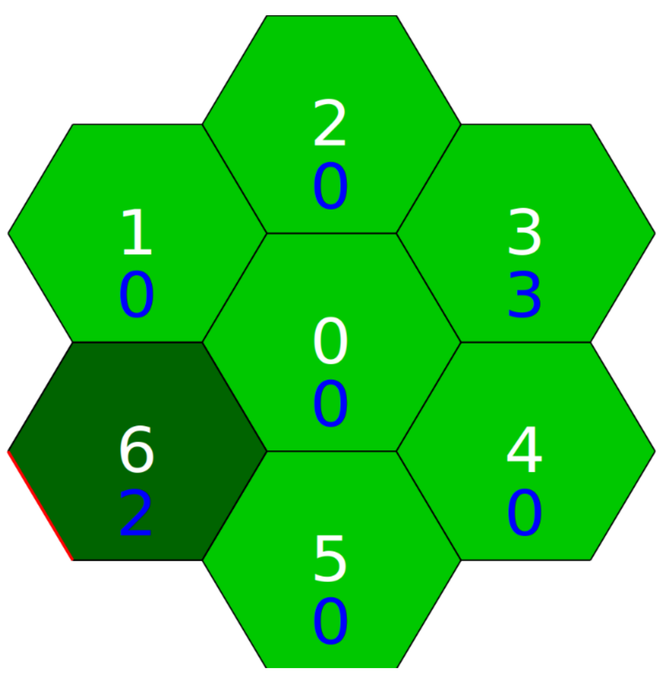
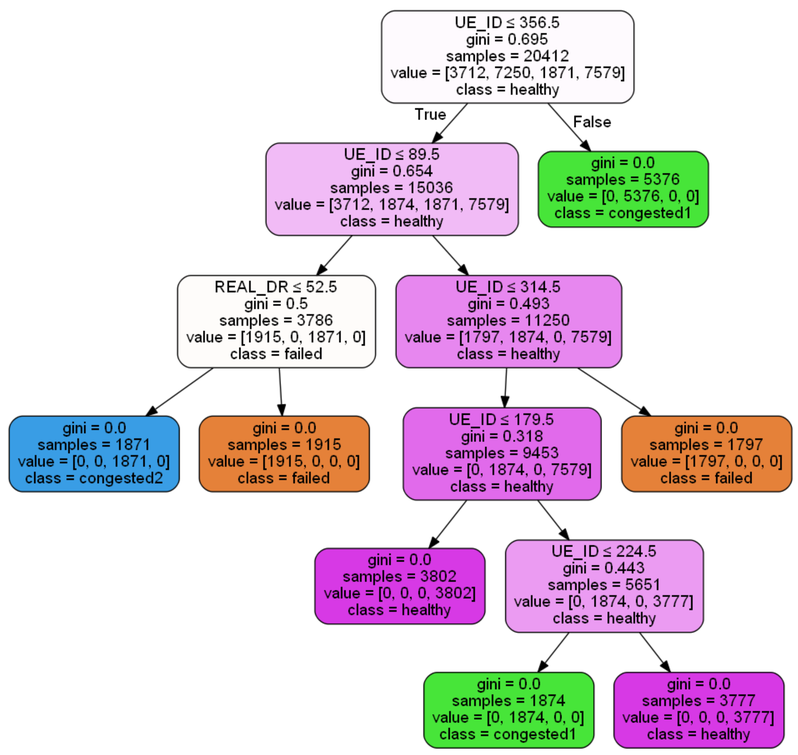


Fig. 2: Decision Tree Output



Conclusion

In this project, a decision tree algorithm was successfully implemented using Python's machine learning libraries. Additionally, the data used for training the decision tree was successfully combined on AWS to create one csv file containing all the data. While compiling our data was successful and AWS proved to be economical, many columns of data from the 5G simulator were not useful for training our decision tree and should be removed in the future. Finally, the decision tree's accuracy was dependent on the variable used to train it, which meant that developing an accurate decision tree required some trial and error.

Future Work

- When combining all simulation data on AWS, extra costs are incurred due to the large amount of data being processed. In the future it will become both beneficial and necessary to reduce the data set and only select columns that are determined to be necessary for training our machine learning models.
- The data generated from the simulator was for one scenario where base station 0 was failed, and base stations 1 and 4 were congested. A richer data set where scenarios are varied will be included to better train our model.
- The machine learning model developed in this project will be integrated in a big data system on AWS to fully utilize big data capabilities.

Acknowledgements

I would like to thank my advisor Dr. Tamer Omar for his support and guidance throughout this project. I would also like to thank the Wireless Network Security Lab (WNSL) at Cal Poly Pomona as well as the Kellogg Honors College for their support in completing this project.

References

[1] T. Rao, P. Mitra, R. Bhatt and A. Goswami, "The Big Data System, Components, Tools, and Technologies: A Survey," *Knowledge and Information Systems,* vol. 60, no. 3, pp. 1165-1245, 2019.

[2] L. Qiao, Y. Li, S. Takiar, Z. Liu, N. Veeramreddy and M. Tu, "Gobblin: Unifying Data Ingestion for Hadoop," *Proceedings of the 41st International Conference on Very Large Data Bases,* vol. 8, no. 12, pp. 1764-1769, 2015.

[3] A. Matacuta and C. Popa, "Big Data Analytics: Analysis of Features and Performance of Big Data Ingestion Tools," *Informatica Economica,* vol. 22, no. 2, pp. 25-34, 2018.

[4] S. K. a. G. T. T. Deshpande, Hadoop: Data Processing And Modelling: unlock the power of your data with Hadoop 2.x ecosystem and its data warehouse techniques across large data sets., Birmngham: Packt, 2016.

[5] K. Matsuzaki, "Functional Models of Hadoop MapReduce with Application to Scan," *International Journal of Parallel Programming,* 2017.