Brett Tully

Final Project Paper

Dr. Kelley

DS Capstone

24 April 2023

DS Capstone

Executive Summary: The aim of this project is to predict classroom occupancy based on temperature, humidity, and the day of the week, using machine learning algorithms. A Raspberry Pi was used to collect temperature data every 5 minutes, and the data was stored on an SQL database. The data was preprocessed and cleaned to create a predictive model using three classification algorithms, Logistic Regression, Decision Tree, and Random Forest. The best-performing algorithm was Random Forest, with an accuracy score of 0.9487. The project has significant potential to be applied in a wide range of industries and settings, including energy efficiency, security, healthcare, retail analytics, and smart homes

Introduction:In today's world, technology is evolving at a rapid pace, and the Internet of Things (IoT) is becoming increasingly popular. The IoT provides new opportunities to collect and analyze data in real-time, which can be used to improve efficiency, safety, and quality of life. One of the applications of IoT is occupancy detection, which is essential for energy management, security systems, and healthcare facilities. The aim of this project is to predict classroom occupancy based on temperature, humidity, and the day of the week using machine learning algorithms. The project uses a Raspberry Pi to collect temperature data every 5 minutes, which is stored on an SQL database. The data is preprocessed and cleaned to create a predictive model using three classification algorithms, Logistic Regression, Decision Tree, and Random Forest.

This project aimed to predict classroom usage based on temperature, humidity, and the day of the week. To accomplish this, the project was divided into several sub-projects:

1. Setting up the Raspberry Pi and Sensors: The first sub-project involved setting up the Raspberry Pi and DHT22 temperature sensors. The Raspberry Pi was connected to a monitor, keyboard, and mouse, and the Raspbian operating system was installed. The sensors were connected to the Raspberry Pi using jumper wires.
2. Collecting Data: Once the Raspberry Pi was set up, it was programmed to collect temperature and humidity data from the DHT22 sensors every 5 minutes using the Python programming language. The data was stored on an SD card for later analysis.

Data Preprocessing: The collected data was preprocessed using Pandas library to clean,

transform and aggregate it, in preparation for analysis.

1. Labeling Classes: The labeling of classes sub-project involved identifying whether a class was actually in session or not. This was done by creating a loop that iterated through the main data frame and compared it to another data frame containing all the times when a class was in session, then marked a 1 or 0 in a new column to identify if it was actually in session.
2. Feature Selection: The feature selection sub-project involved identifying the most important features to include in the machine learning models. Temperature and humidity were the most important features, followed by the day of the week.
3. Model Training: Three classification models were trained to predict classroom usage: Logistic Regression, Decision Tree, and Random Forest. A grid search was used to tune the hyperparameters for each model. The Random Forest model had the best performance with an accuracy score of 0.9487. The Decision Tree model also performed well with an accuracy score of 0.9305, and the Logistic Regression model had an accuracy score of 0.9275.
4. Real-time Data Visualization: A data visualization dashboard was created using Microsoft Power BI that displayed the live occupancy in the classrooms. The dashboard was connected to the Raspberry Pi and configured to update every 5 minutes.

During this project, I faced several challenges and successes. One of the successes of the project was the data collection system. Once I was able to set up the Raspberry Pi sensors and configure them to send data to the SQL database, the data collection process was smooth and efficient. I was able to collect data on temperature, humidity, and classroom usage for multiple classrooms over several weeks. This allowed me to build a robust dataset for the project.

However, setting up the Raspberry Pi sensors and configuring them to send data to the SQL server was a challenging process. I had to install several libraries and packages on the Raspberry Pi to enable it to push data to the SQL server. This process took a significant amount of time and required troubleshooting to get it right. Additionally, making one of the Raspberry Pis the main node for the SQL server was another challenge. This required knowledge of networking and setting up a secure connection between the Raspberry Pi and the SQL server.

Another issue that didn't go well was creating the loop to assign a 1 when classes were in session. Identifying when a class was actually in session was a challenging task. I had to create a loop that iterated through the main data frame and compared it to another data frame containing all the times when a class was in session. This was a time-consuming and challenging process, but eventually, I was able to create a loop that marked a 1 or 0 in a new column to identify if a classroom was in session.

Despite the challenges, the project provided me with valuable experience in data preprocessing, feature engineering, and machine learning. It highlighted the importance of collecting accurate data and preprocessing it effectively, as well as the value of exploring different models and tuning their hyperparameters. I learned how to use various classification models, such as Logistic Regression, Decision Tree, and Random Forest, and was able to evaluate their performance using metrics such as accuracy score.

In conclusion, this project was challenging but rewarding. The data collection system and classification models were successful, but there were several challenges along the way. If I were to do this project again, there are a few things I would change to make the process smoother. Firstly, I would spend more time on the initial setup of the Raspberry Pi devices, including ensuring that all necessary libraries and packages are installed correctly. This would save time and frustration down the line, as setting up the devices and getting them to communicate with the SQL server was one of the most challenging parts of the project.

Another change I would make is to include more lines of code in the script that the Raspberry Pi devices use when sending data to the SQL server. Specifically, I would pre-train one of the machine learning models on a large portion of the data and include that same model in the script so that it outputs a prediction when it sends the data to the SQL server. This would allow for "real-time" predictions, which would be very useful in creating more dynamic visualizations and monitoring classroom usage in real-time.

Lastly, I would have allocated more time for the project, as creating the models was done basically the day of the celebration of research. While I had some knowledge of machine learning and model creation from the Machine Learning in the Cloud class, spending more time on the project would have allowed for more in-depth analysis and exploration of different models and techniques.

In conclusion, this project was challenging, but quite a memorable experience. There certainly is significant potential for real-world applications, particularly in building automation, security, healthcare, retail analytics, and smart homes. The results have shown that it is possible to accurately predict classroom usage based on temperature, humidity, and day of the week, using machine learning algorithms. Future work could include expanding the dataset to include more features, such as sound and light levels, to further improve the accuracy of the model. It may also be useful to explore more advanced machine learning techniques, such as neural networks, to improve the predictive power of the model. Another area for future work is the application of the model to other settings, such as office buildings, hospitals, and retail stores. This could involve adapting the model to consider the unique features of each setting and exploring different methods for data collection and preprocessing. As the Internet of Things continues to evolve, there will be even more opportunities to apply these techniques to a wide range of industries and settings, with the potential to improve efficiency, safety, and quality of life.