Predicting Classroom Occupancy Based on Ambient Temperature Changes

Brett Tully

[btully@bellarmine.edu](mailto:btully@bellarmine.edu)

26 April 2023

**Executive Summary**

This project aims to predict classroom occupancy based on Temperature (F), Humidity, and the Weekday, using machine learning. Three Raspberry Pis were placed in three separate classrooms and were used to collect temperature data every 2 seconds, and the data was stored on an SQL cloud database. Luckily, I did not have to do very much data cleansing, as every data point was recorded in the proper format. I did have to do some preprocessing that involved changing the data types of the date column, and within my Jupyter notebook, I separated the Date Time column into a column for just the time (in military time), Day of the Week, and the date by itself. After that, I had to create a dictionary that contained all the possible times that the rooms which contained the Raspberry Pi would be in session. I then turned the dictionary into a Data Frame and created a loop that iterated through the data from the database to mark a 1 if a class was in session based on the room number, time, and day of the week. Then, I was able to create a predictive model using three classification algorithms in a Grid Search CV using 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 project uses a Raspberry Pi to collect temperature data every 2 seconds, to be able to use the Temperature, Humidity, and Day of the Week to be able to predict whether a classroom is occupied or empty.

**Project Details**

This project was divided into several sub-projects which have their own level of difficulty:

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 then connected to the Raspberry Pis.
2. Collecting Data: Once the Raspberry Pi was set up, it was programmed to collect temperature and humidity data from the DHT22 sensors every 2 seconds using the Python programming language.
3. Data Preprocessing: The collected data was preprocessed using the Pandas library to clean, transform and aggregate it, in preparation for analysis.
4. Labeling Classes: The labeling of classes sub-project involved identifying whether a class was 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 possible 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 per the time stamp on the data.
5. Feature Selection: The feature selection sub-project involved identifying the most important features to include in the machine learning models. I did not have many features to work with, but temperature and humidity were the most important features, followed by the day of the week. I had to scale my temperature and humidity columns while also using a label encoder for the day of the week.
6. Model Training: Three classification models, Logistic Regression, Decision Tree, and Random Forest were combined into a Grid Search CV and then the data had to be taken by samples to run through the model (due to hardware limitations, the model is unable to process over 4 million rows promptly). A grid search was used to tune the hyperparameters for each model as you can add as many parameters as you would like. I chose this method because I figured it would help me select the best model with the best parameters, rather than manually testing each model with different parameters. 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.
7. 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 or whenever I refreshed the query.

**Reflection**

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, although there was an instance where two of my devices were disconnected (likely by someone bumping into them or an internet connection issue, which resulted in only one classroom providing data to my SQL server). Aside from the disconnection errors, 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, which eventually amassed over 4.5 million rows of data.

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, which is why we opted out of this method and just used an SQL server in the cloud. 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 in session was a challenging task because I originally was using just a nested dictionary that listed each classroom with the possible class times for each day of the week. The problems were difficult to troubleshoot due to it being a nested loop, so I then opted to turn the dictionary into a data frame with each row being a possible class time. After converting it to a data frame, I eventually figured out the loop that iterated through the main data frame and compared it to the other data frame containing all the times when a class 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. 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 that I had to deal with.

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 to the project, as creating the models was done on the day of the celebration of research. Luckily, I had great knowledge and experience in 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.

**Conclusion and Future Work**

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.