In this project, I aimed to predict classroom usage based on temperature, humidity, and the day of the week. To do so, I used a dataset that was collected by Raspberry Pi sensors installed in classrooms. However, setting up the Raspberry Pi sensors and sending the data to an SQL database was complex and time-consuming. Additionally, identifying whether a class was actually in session or not was challenging, but I was able 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, then marked a 1 or 0 in a new column to identify if it was actually in session.

Next, I used three classification models: Logistic Regression, Decision Tree, and Random Forest, to predict classroom usage. I used a grid search to tune the hyperparameters for each model and found that the Random Forest model had the best performance with an accuracy score of 0.9487. The best parameters for the Random Forest model were: {'clf\_\_max\_depth': 50, 'clf\_\_min\_samples\_leaf': 1, 'clf\_\_n\_estimators': 50}. The Decision Tree model also performed well with an accuracy score of 0.9305, and the best parameters were: {'clf\_\_criterion': 'entropy', 'clf\_\_max\_depth': 15, 'clf\_\_min\_samples\_leaf': 1}. The Logistic Regression model had the lowest accuracy score of 0.9275, and the best parameters were: {'clf\_\_C': 1, 'clf\_\_penalty': 'l2', 'clf\_\_solver': 'lbfgs'}.

In conclusion, the Random Forest model was the most accurate in predicting classroom usage, with an accuracy score of 0.9487. However, it's worth noting that the Decision Tree model also had good performance, with an accuracy score of 0.9305. The project highlighted the importance of collecting accurate data and preprocessing it effectively, as well as the value of exploring different models and tuning their hyperparameters

Graphical user interface

Description automatically generated

Chart, treemap chart

Description automatically generated