Room occupancy Estimation

Azzam alfurhud

Abstract

This project aims at prediction the number of people in the room giving some information gathered from sensors. The data is from UCI Machine Learning Repository as part of a research (Cited at the end of the document). As a multi-class classification problem, I used three models: Logistic Regression, Gaussian Naïve Bayes and XGboost. At the end of the project, I compared the models and chose the best one using appropriate metrics.

Design

The project has four classes 0, 1, 2, 3 that represent the number of people occupying the room. By building, a machine-learning model we hope to predict the number of people using features like the change of temperature, sound, CO2 level and other features. In addition, we may interpret what features affected by the number of people the most.

Data

The dataset contains 10,129 observations and 16 features. Some features are coming from different sensors so I believe it better to aggregate them to achieve a more accurate result.

Models

The data was split into training (80%) and testing (20%). The features were also standardized for comparison with non-standardized features.

The models used in this project are Logistic Regression, Gaussian Naïve Bayes and XGboost. Note that the standardized features were only

used in the logistic regression model due to the other models not being distance based (It will not affect the result). The results are as follows:

Metrics	Logistic Regression (Standardized features)	Gaussian Naïve Bayes	XGboost
F1 Score (Macro Average)	0.85	0.87	0.99
F1 Score (Weighted Average)	0.95	0.95	1.00
Accuracy	~ 0.9526	~ 0.9531	~ 0.9960

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting