AIR QUALITY PERSONAL RECOMMENDATION

Created on 16th August, 2023. Owned by AirQo.
Last Updated: 18th August, 2023

Overview

This model is designed to provide personal recommendations to users depending on their health status, age, activity and pm2_5 value. This document outlines the details of the training job for the machine learning model used in the task.

Intended Goal

The model is to be used in the AirQo mobile application that offers a variety of features such as real time pm2.5 readings, a dynamic map showing air quality of different locations, among others.

Data Description

We used two kinds of datasets : site data and user data.
The site data was hourly data from both AirQo and non-AirQo devices for the first months of the year 2023. This data was collected from devices from all locations, to reduce the likelihood of overfitting of the model. The data had various attributes such as tenant, timestamp,site_id,site_name,site_latitude,parish, site_longitude, pm2_5, device_number, device_id, pm2_5_raw_value, pm2_5_calibrated_value, and so many others. It was initially 250459 but then reduced to 52 in consideration to random pm2_5 value per site.
The following columns were selected site_name, site_latitude, ,site_longitude, and pm2_5 for further analysis.The pm2_5 category column was added on to them using the pm2.5 standard indices.
The user data was randomly generated for ages ranging from 18 to 60. It had various columns: Name, Date of Birth,Longitude, Latitude, Health_Conditions, Health_Status, and Activities. The following columns were selected: Longitude, Latitude, Age(calculated from date of birth), Health_Conditions, Health_Status, and Activities.
The two data frames containing the selected columns were merged to form one dataset. 70% of the data was used in the training process and 30% was used in testing the model.

Model Architecture

The model was built using different algorithm approaches: Decision Tree, Random Forest Classifier

For the **Decision Tree model**, it was developed with a random state of 42 and then trained with 70% of the data.
For the **Random Forest Classifier model**, it was trained with 100 estimators, random state of 42, 1 job, and maximum depth of 5.

Model Training

The data used for training the model was 70% of the whole data of the first months of 2023. Below are the steps that were taken in the data preprocessing;

- Merging the site and user data frames containing the selected columns.
- Dropping unwanted columns for model development.
- Obtaining the recommendation column
- Encoding categorical features
- Split the data into training and testing

The training time of all algorithm models were relatively the same.
The hyperparameters of the Random Forest Classifier were optimized using the Random Search CV method with forest_rf, param distributions of 5 iterations and 5 cv. The best parameters were 156 estimators and maximum depth of 15.

Model Evaluation

A simple user data relative to the training data was used and obtained the following results.
For the **Decision Tree model**, an accuracy of 0.88 was obtained with the corresponding classification report with the precision, recall, and f1-score ranging from 0.40 - 1.00.
For the **Random Forest Classifier,** an accuracy of 0.8311 was obtained before tuning and 0.89111 after tuning.

Potential Biases

Given that the data was not evenly distributed, there's a high possibility that the model started assuming recommendations.

Limitations

The model can not make future predictions given it depends on the pm2.5 value at that specific time.

References

“AirQo Low-Cost Air Quality Monitor Calibration Challenge.” *Zindi*,

https://zindi.africa/competitions/airqo-low-cost-air-quality-monitor-calibration-challenge.

“AQI Basics.” *AirNow.gov*, https://www.airnow.gov/aqi/aqi-basics/.

“PM2.5 particles in the air | Environment Protection Authority Victoria.” *EPA Victoria*,

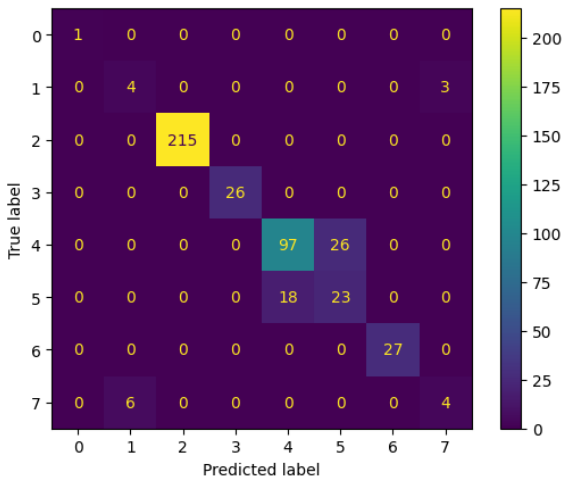
https://www.epa.vic.gov.au/for-community/environmental-information/air-quality/pm25-particles-in-the-air.

Qualitative Analysis

Considering Decision Tree model,
Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.40	0.57	0.47	7
2	1.00	1.00	1.00	215
3	1.00	1.00	1.00	26
4	0.84	0.79	0.82	123
5	0.47	0.56	0.51	41
6	1.00	1.00	1.00	27
7	0.57	0.40	0.47	10
accuracy			0.88	450
macro avg	0.79	0.79	0.78	450
weighted avg	0.89	0.88	0.88	450

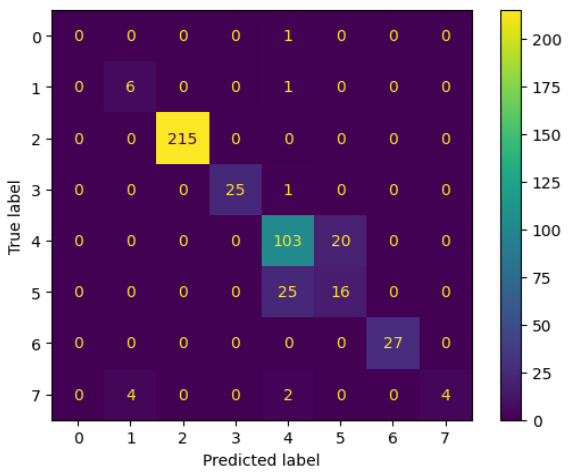
Confusion matrix



Considering Random Forest Classifier
Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.40	0.57	0.47	7
2	1.00	1.00	1.00	215
3	1.00	1.00	1.00	26
4	0.84	0.79	0.82	123
5	0.47	0.56	0.51	41
6	1.00	1.00	1.00	27
7	0.57	0.40	0.47	10
accuracy			0.88	450
macro avg	0.79	0.79	0.78	450
weighted avg	0.89	0.88	0.88	450

Confusion matrix



Quantitative Analysis