HUMIDITY LEVEL PREDICTOR USING GAS MULTISENSOR DATA

CHEMICAL INTELLIGENCE: ML DRIVEN CHALENGE

GOPAL GUPTA
PARITOSH SARANGI

TABLE OF CONTENTS

Problem Statement:

The project focuses on classifying RH levels into one of the five categories: "Dry", "Ideal", "Slightly Elevated", "Elevated", or "High".
The dataset contains hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. We will also use ground truth hourly averaged concentrations for CO, Non-Metallic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2) provided by a co-located reference certified analyzer.

Objective:

- Develop an accurate classification system for RH levels based on sensor data from a gas Multisensor device deployed in an Italian city.
- The classification system will be a valuable tool for industries such as air quality management, public health, and urban planning, supporting efforts to improve air quality in urban environments.
- The project aims to provide actionable insights to industry stakeholders to inform decision-making and proactive measures to reduce risks associated with poor air quality.

PROCESS PIPELINE

Step 1

Data Cleaning

Step 2

Feature Engineering (Calculating RH)

Step 3

Label Encoding and Normalisation of Data

Step 4

Classification using multiple models

Step 5

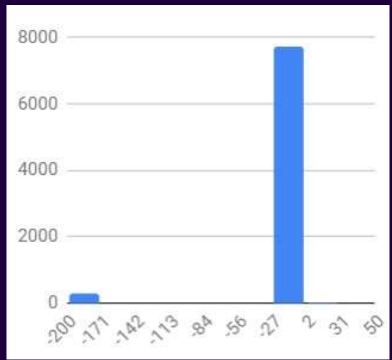
Comparison of predictions and Final Result

DATA CLEANING

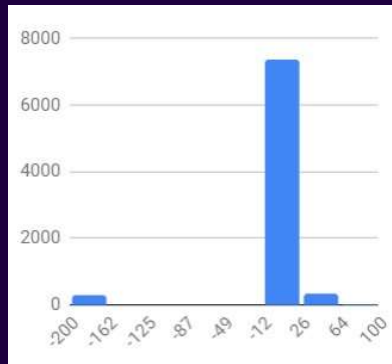
- We remove the Date and Time features from the dataset since those are unnecessary for the task.
- Removed NMHC(GT) since almost 80-90% of feature rows had NaN entries.
- Use MICE imputation to fill up the NaN values in the rest of the columns

Distribution of Feature Entries

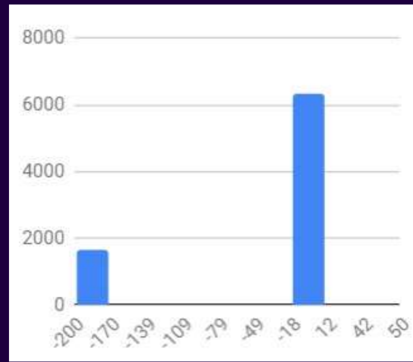
AH



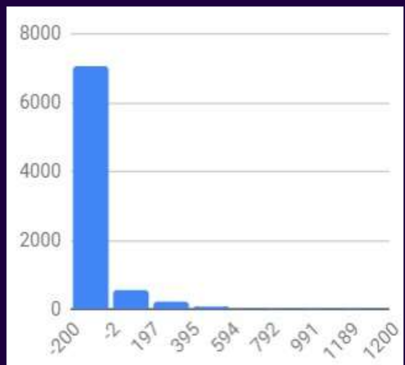
C6H6(GT)



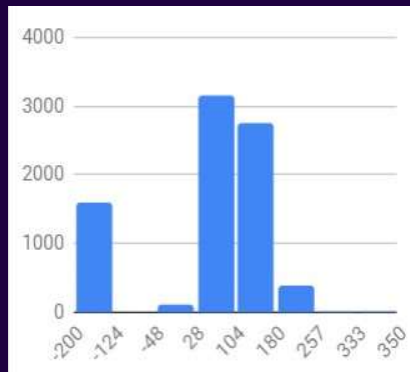
CO(GT)



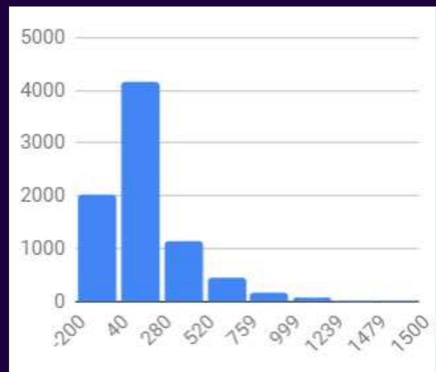
NHMC(GT)



NO2(GT)



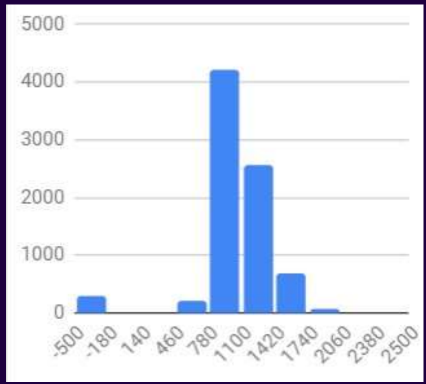
NOx(GT)



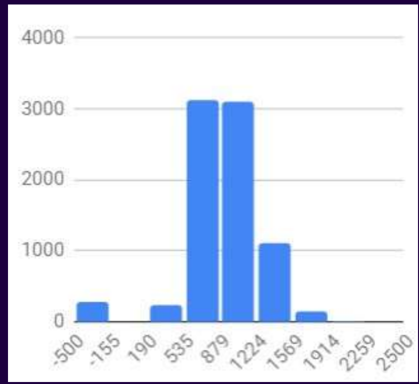
DATA CLEANING

Distribution of Feature Entries

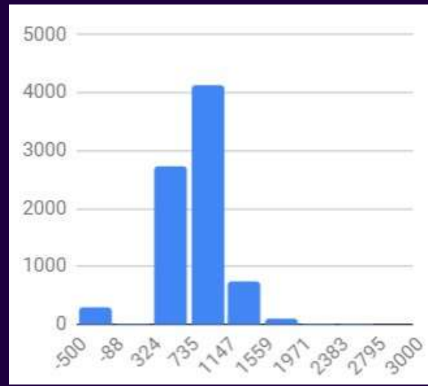
PT08.S1(CO)



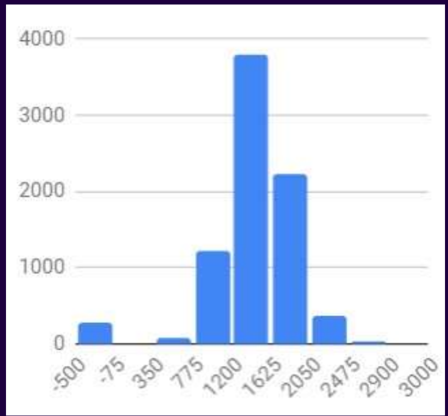
PT08.S2(NHMC)



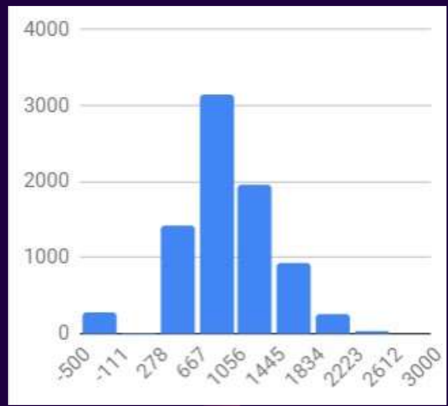
PT08.S3(NOx)



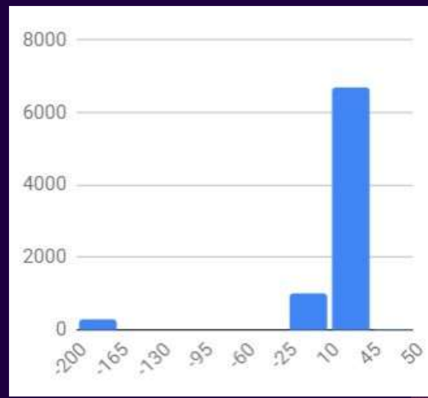
PT08.S4(NO2)



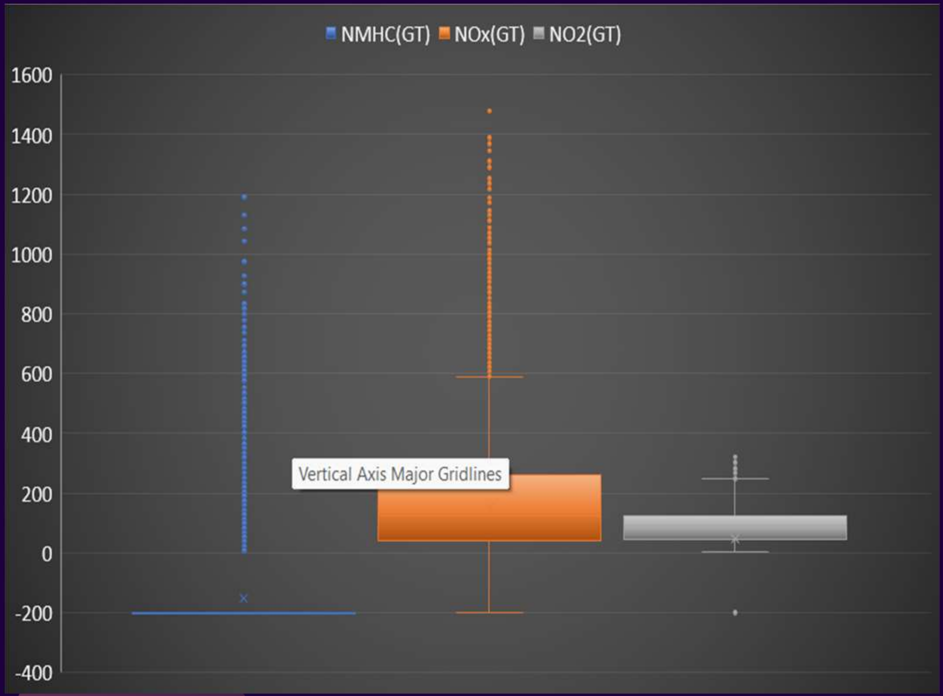
PT08.S5(O3)

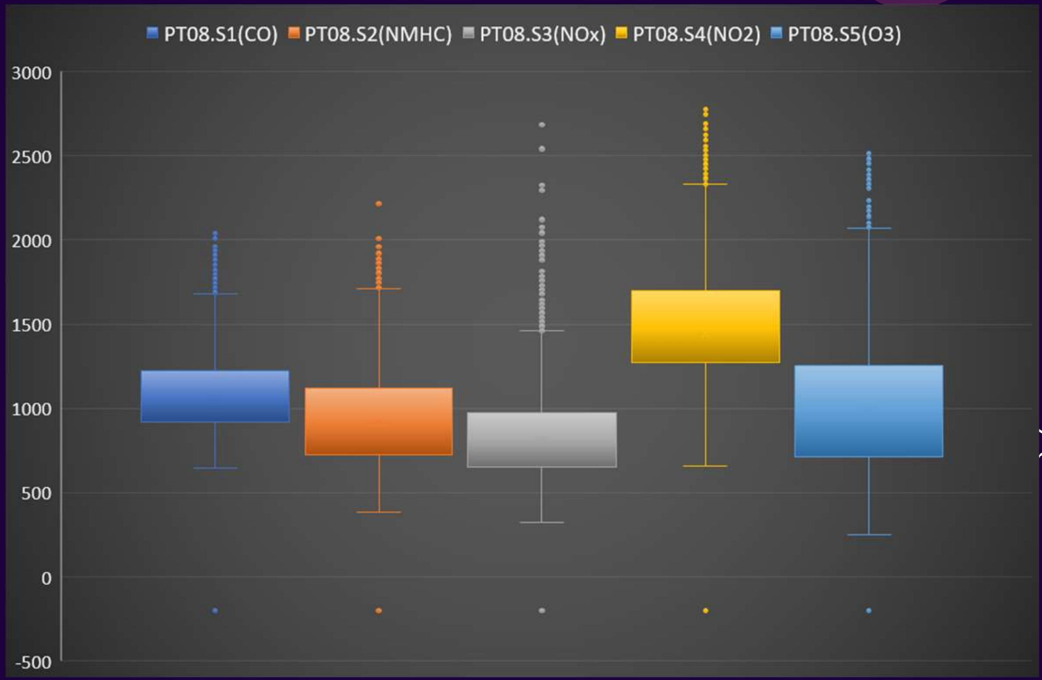


T



BOX AND WHISKER PLOT VISUALIZATION





FEATURE ENGINEERING - RELATIVE HUMIDITY

$$\text{Relative Humidity(RH)} = 100 \times \frac{P_a}{P_s}$$

Actual Vapour Pressure of water

Saturation Vapour Pressure of water

Relation between AH, T & RH:

$$\text{RH} = 100 \times \frac{AH \times R_W \times T(K)}{P_s}$$

$Where\ R_w = Specific\ gas\ constant\ for\ water\ vapour = 461.5\ J/(Kg.K)$

$P_s\ at\ given\ T(K)\ is\ calculated\ by\ equation\ proposed\ by\ Wagnus\ \&\ Pruss\ given\ as:$

$$P_s = P_c \times \exp\left\{\frac{T_c}{T}(a_1\vartheta + a_2\vartheta^{1.5} + a_3\,\vartheta^3 + a_4\,\vartheta^{3.5} + a_5\,\vartheta^4 + a_6\,\vartheta^{7.5})\right\}$$

$where\ P_c = Critical\ Pressure\ of\ Water(22.064\ MPa), T_c = Critical\ Temperature\ of\ Water(647.096\ K),$

$a_1, a_2, \ldots a_6 = Emperical\ Constant,\qquad \vartheta = 1 - \frac{T}{T_c}$

CLASSIFICATION MODELS USED

MODELS	F1 SCORES					ACCURACY
	Dry	Elevated	High	Ideal	Slightly Elevated	
Random Forest	0.97	0.98	0.98	0.98	0.98	0.98
XGBoost	0.97	0.98	0.98	0.98	0.97	0.98
Adaboost	0.97	0.99	0.99	0.98	0.98	0.98
Complement Naive Bayes	0.61	0.23	0.42	0.34	0.00	0.40
KNN	0.91	0.75	0.85	0.88	0.73	0.83
SVM (rbf kernel)	0.91	0.98	0.97	0.95	0.96	0.95
Gradient Boosting	0.94	0.98	0.99	0.97	0.96	0.97

COMBINING THE MODELS

We use different combinations of the models mentioned in the previous slide to create ensembled and stack generalized models for achieving better results.

These models did bring improvement in the F1 scores for the individual classification of the classes.

The Ensemble model (with VotingClassifier) and Stack Generalization model (with GradientBoostingClassifier) of XGBoost, Random Forest and Adaboost provided the best results as shown below :

MODELS	F1 SCORES					ACCURACY
	Dry	Elevated	High	Ideal	Slightly Elevated	
Ensemble	0.98	0.99	0.99	0.98	0.98	0.98
Stack Generalization	0.98	0.99	0.99	0.99	0.98	0.98

We decide to go with the Stack Generalization model due to its superior F1 score performance on individual classes.

CONCLUSION

After analyzing the performances of and outputs generated by all the models used by us. We finally decide to use the Stack Generalization model comprised of XGBoost, Random Forest, and AdaBoost Classifiers.

THANK YOU

REFRENCES

- van Buuren(TNO), S., & Groothuis-Oudshoorn(University of Twente), K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.
- Absolute Humidity Calculator." Omni Calculator, Omni Calculator Project, 2021, https://www.omnicalculator.com/physics/absolute-humidity#how-to-calculate-absolute-humidity-from-relative-humidity-and-temperature.
- Aqua-Calc Humidity Calculator." Aqua-Calc, Aqua-Calc.com, 2021, https://www.aqua-calc.com/calculate/humidity.