Analysis Of Gas Sensor Array Under Dynamic Gas Mixtures

Vaibhav Dnyandeo Ingale (016131167)
Lalitha Ramya Vemuri (016697317)
Hardwin Bui (011007156)

Introduction

- Chemical sensors are used in a wide variety of applications, including environmental monitoring, industrial process control, and medical diagnostics
- To ensure their accuracy and performance, it is important to be able to predict the concentration of the gas that they are sensing

Dataset Overview

- UCI Machine Learning repository.
  This data provides Ethylene concentration level readings of 16 different sensors in 2 different chemical environments.
- This data set consists of 4178504 instances and 19 attributes.

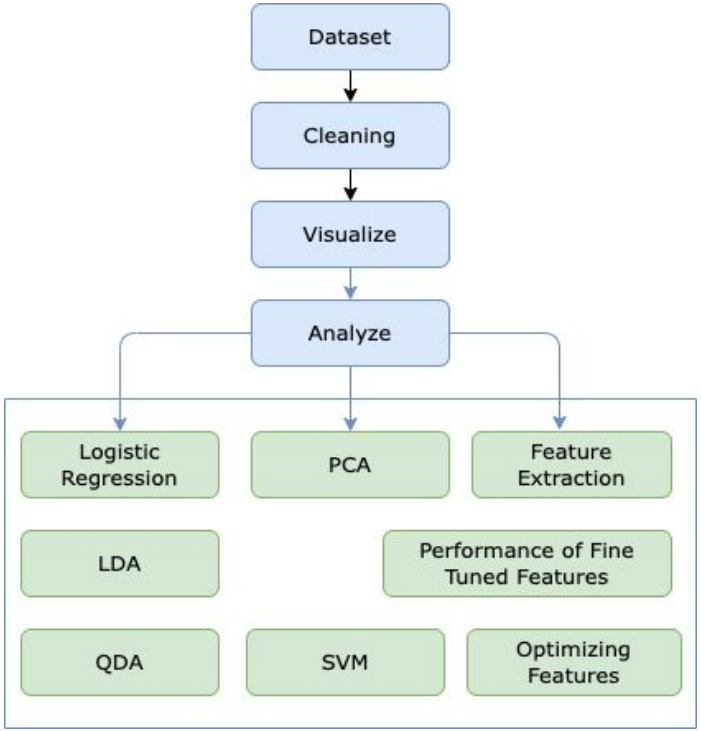
Project Goal

- Develop a data mining system that can be used to identify patterns in chemical sensor time series data
- Classify and predict the concentration of different gasses in the air with high accuracy

Algorithms to Consider

- Logistic Regression
  - An extension of linear regression that can be used to predict qualitative response for an observation
- LDA & QDA
  - Can provide accuracy score, classification report, and confusion matrix for the dataset
- K-Fold
  - Cross-validation can help evaluate the performance of the model
- PCA
  - Can be used to reduce the dimensionality of the data and capture the most important features

Process Flowchart



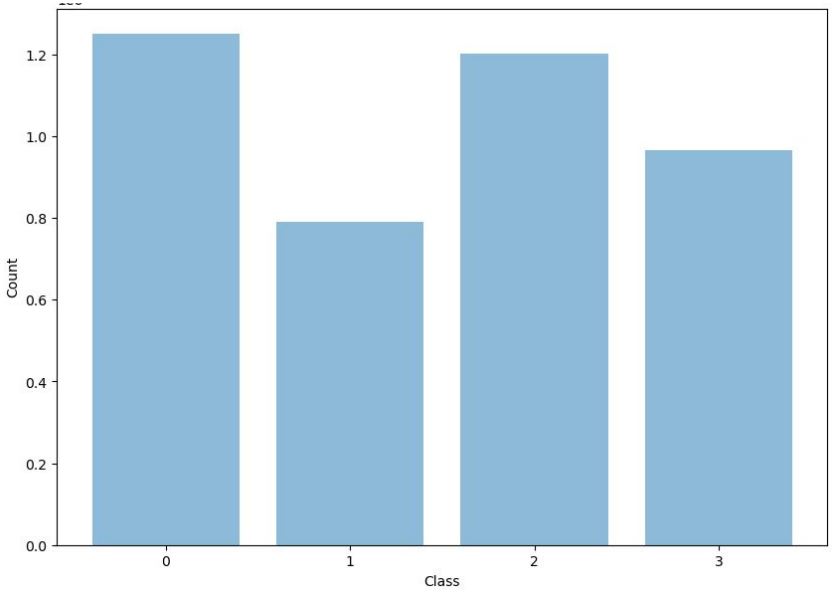
Importance of Sensors

- Industrial Safety
  - help prevent accidents caused by gas leaks and ensure the safety of workers
- Agriculture
  - monitor the levels of ethylene gas in greenhouses, which is naturally produced by ripening fruits and vegetables
- Environmental Monitoring
  - detect and monitor the levels of methane gas emissions in landfills and wastewater treatment facilities
- Food Quality Control
  - detect the presence of ethylene gas in food storage and transportation environments
- Medical Application
  - detect the presence of certain gasses in a patient's breath, which can be used to diagnose various medical conditions

Project Methodology

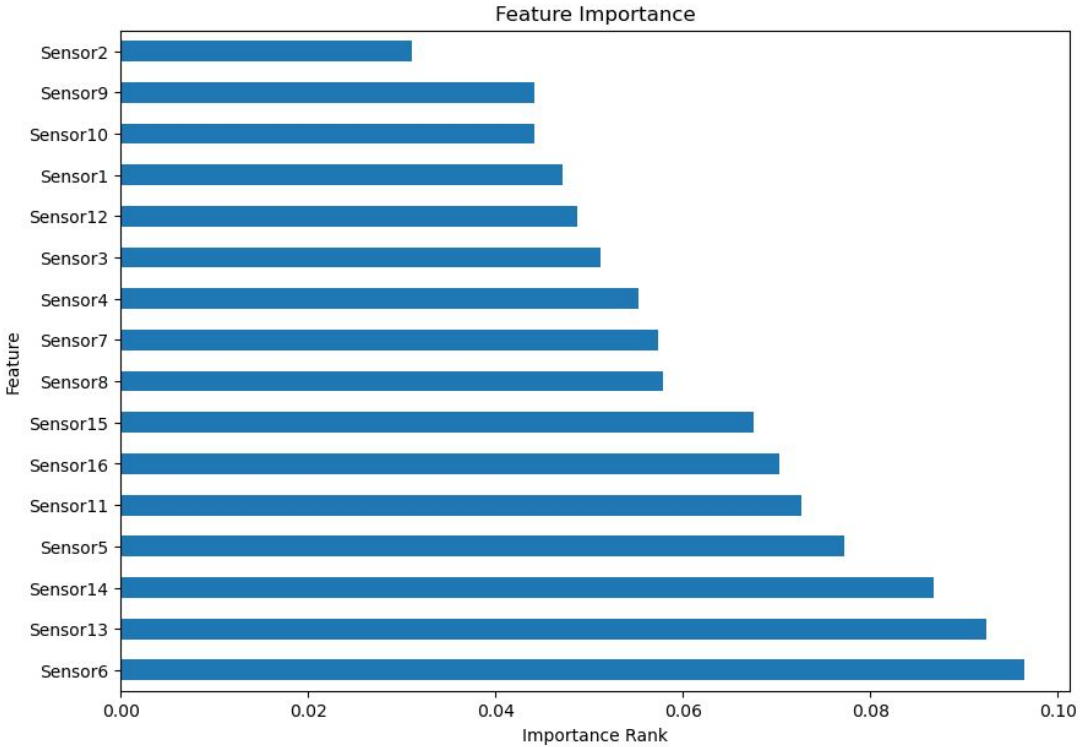
EDA/Preprocessing	Feature Selection	Model Selection	Model Training	Testing and Evaluation
Remove missing values and normalize sensor readings	Reduce dimensionality of data and select most important features that contribute to detection of Ethylene and Methane.	- Logistic Regression (LR) - Linear Discriminant Analysis (LDA) - Quadratic Discriminant Analysis (QDA) - KNN - SVM	For LDA, QDA, and LR, data is split into half for training and half for testing and accuracy is calculated using testing data.	For all models used, accuracy was obtained by comparing predicted labels to true labels.

Distribution of classes



Class	Ethylene	Methane
0	0	0
1	1	0
2	0	1
3	1	1

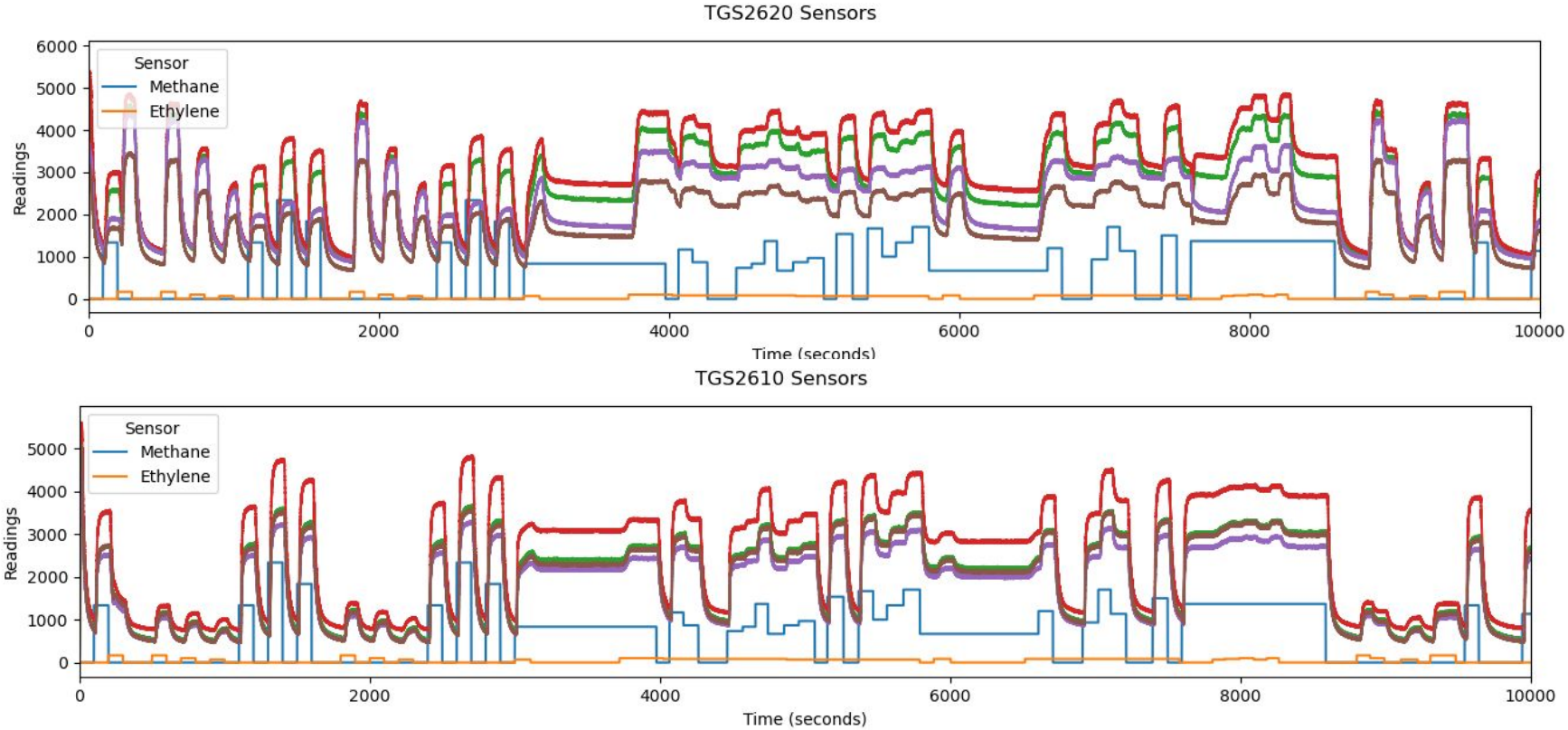
Feature Importance



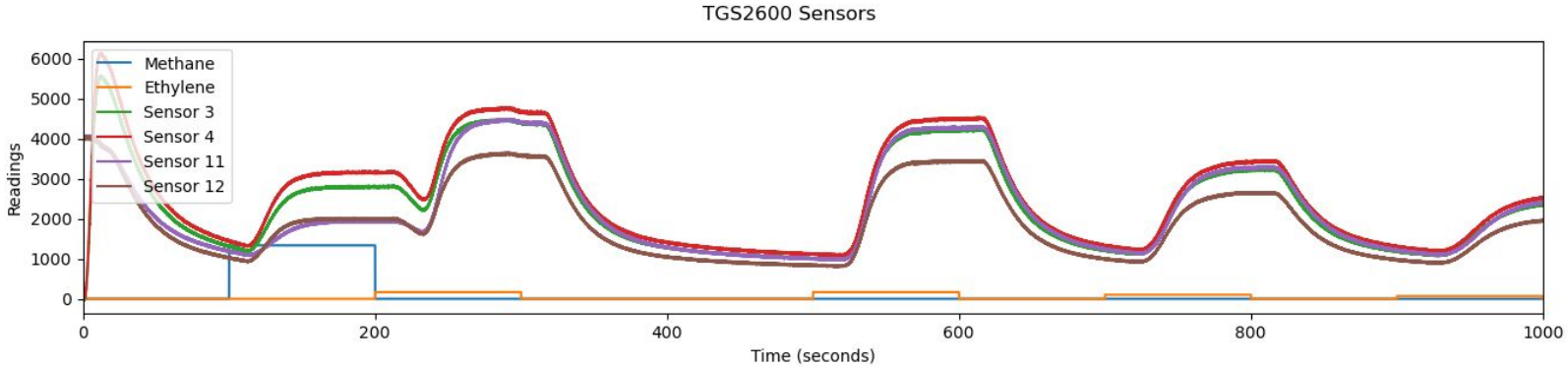
We used inbuilt class feature_importances of tree based classifiers - ExtraTreesClassifier

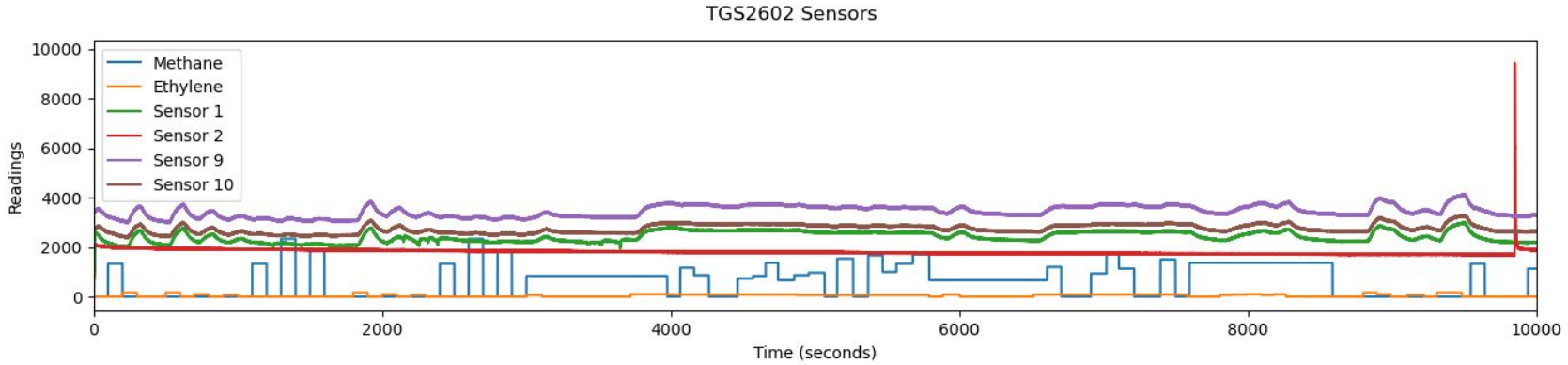
By knowing which features are the most important, we can focus our attention on those features and potentially discard less important ones, which can reduce the dimensionality of the dataset and improve the performance.

Sensor Readings vs Concentrations



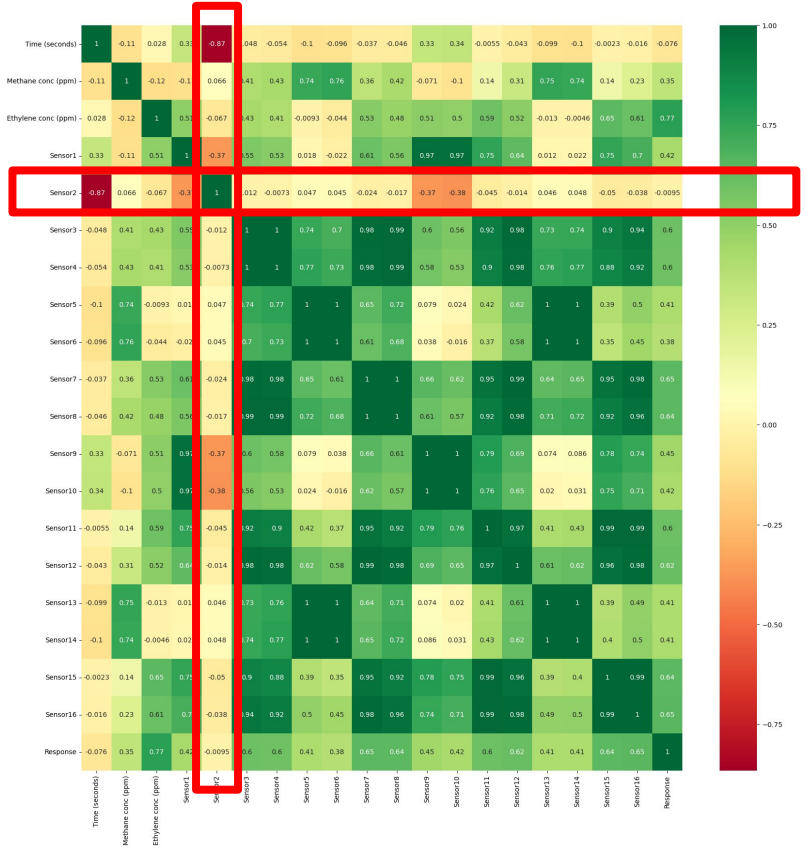
Sensor Readings vs Concentrations





Correlation Matrix Heat Map

Sensor 2 Readings



Classification Reports

LDA Methane model confusion matrix

	precision	recall	f1-score	support
0	0.95	0.94	0.95	1159945
1	0.93	0.94	0.93	919299
accuracy			0.94	2079244
macro avg	0.94	0.94	0.94	2079244
weighted avg	0.94	0.94	0.94	2079244

Logistic Regression Methane model confusion matrix

	precision	recall	f1-score	support
0	0.95	0.95	0.95	1159945
1	0.94	0.94	0.94	919299
accuracy			0.95	2079244
macro avg	0.94	0.94	0.94	2079244
weighted avg	0.95	0.95	0.95	2079244

QDA Methane model confusion matrix

	precision	recall	f1-score	support
0	0.95	0.94	0.95	1159945
1	0.93	0.94	0.93	919299
accuracy			0.94	2079244
macro avg	0.94	0.94	0.94	2079244
weighted avg	0.94	0.94	0.94	2079244

Analysis of results

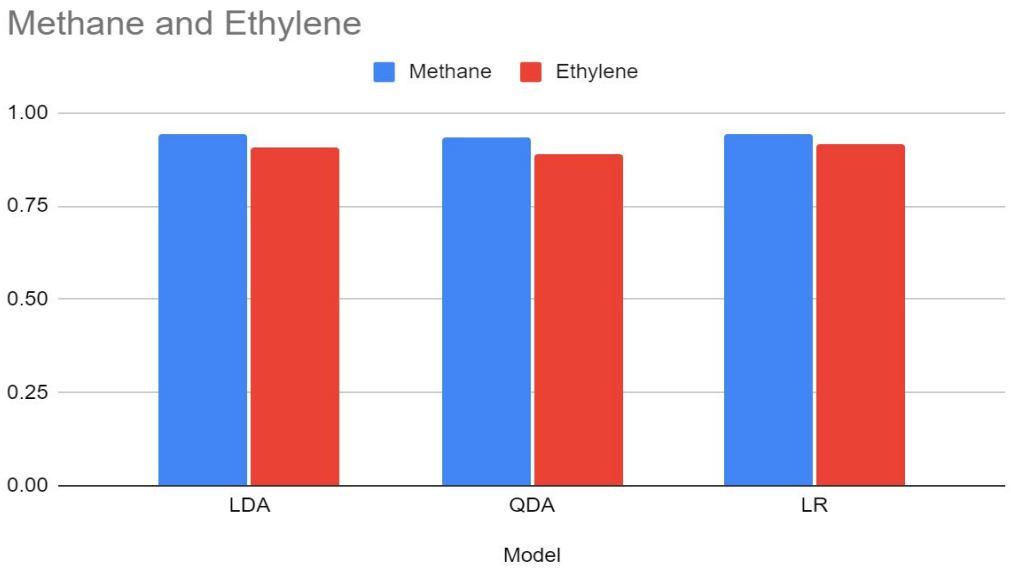
Accuracy After Preprocessing:

Model	Methane	Ethylene
LDA	0.9390203784822619	0.9044925771383171
QDA	0.9295770786802177	0.8857002422024843
LR	0.9436619515094852	0.9124920885010571

Performance of fine-tuned features

Model	Methane	Ethylene
LDA	0.9411863157955488	0.9060610491120811
QDA	0.9327039058427005	0.8911695789431159
LR	0.9451661276887177	0.9164210645792413

Analysis of results



Conclusion

- In conclusion, we analyzed gas mixtures using an array of sensors and compared the performance of three classification models: logistic regression, LDA, and QDA.
- After evaluating the results, we found that all three models performed well, but logistic regression was the most effective in this case. It provided the highest accuracy rate, indicating that it was able to correctly classify gas mixtures with a higher degree of accuracy than the other models.
- Overall, the use of an array of sensors and logistic regression can be a useful tool for accurately classifying gas mixtures, which can have a wide range of applications in various industries.