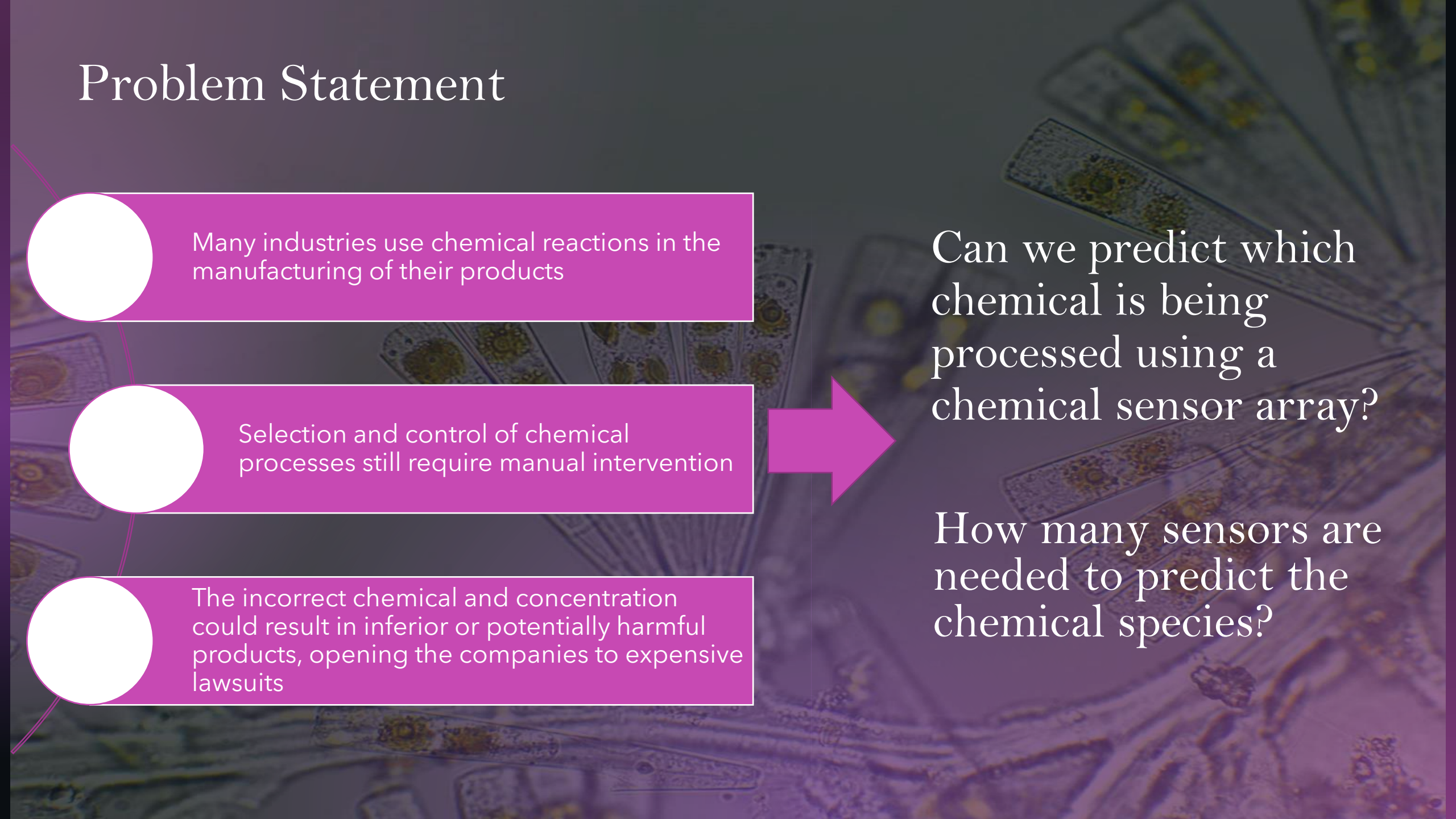


Predicting Chemical Species

SAMUEL SCOVRONSKI

DEC. 7TH, 2021

Problem Statement



Many industries use chemical reactions in the manufacturing of their products

Selection and control of chemical processes still require manual intervention

The incorrect chemical and concentration could result in inferior or potentially harmful products, opening the companies to expensive lawsuits



Can we predict which chemical is being processed using a chemical sensor array?

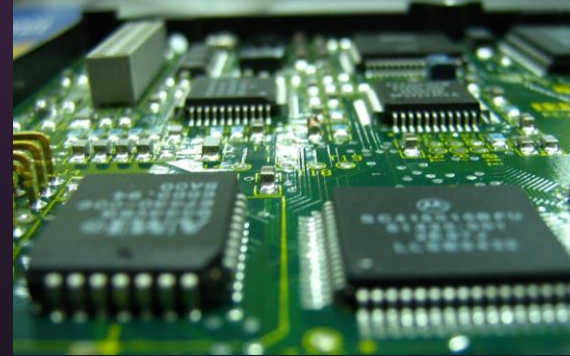
How many sensors are needed to predict the chemical species?

Why it is important?

Companies that use chemical processes



Their products



Data source

All data was provided by the BioCircuits Institute at UC – San Diego through UCI Machine Learning Repository

<http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset+at+Different+Concentrations>

Data provided was created using a sensor array comprising of 16 chemical sensors and their response to 6 chemical species at varying concentrations.

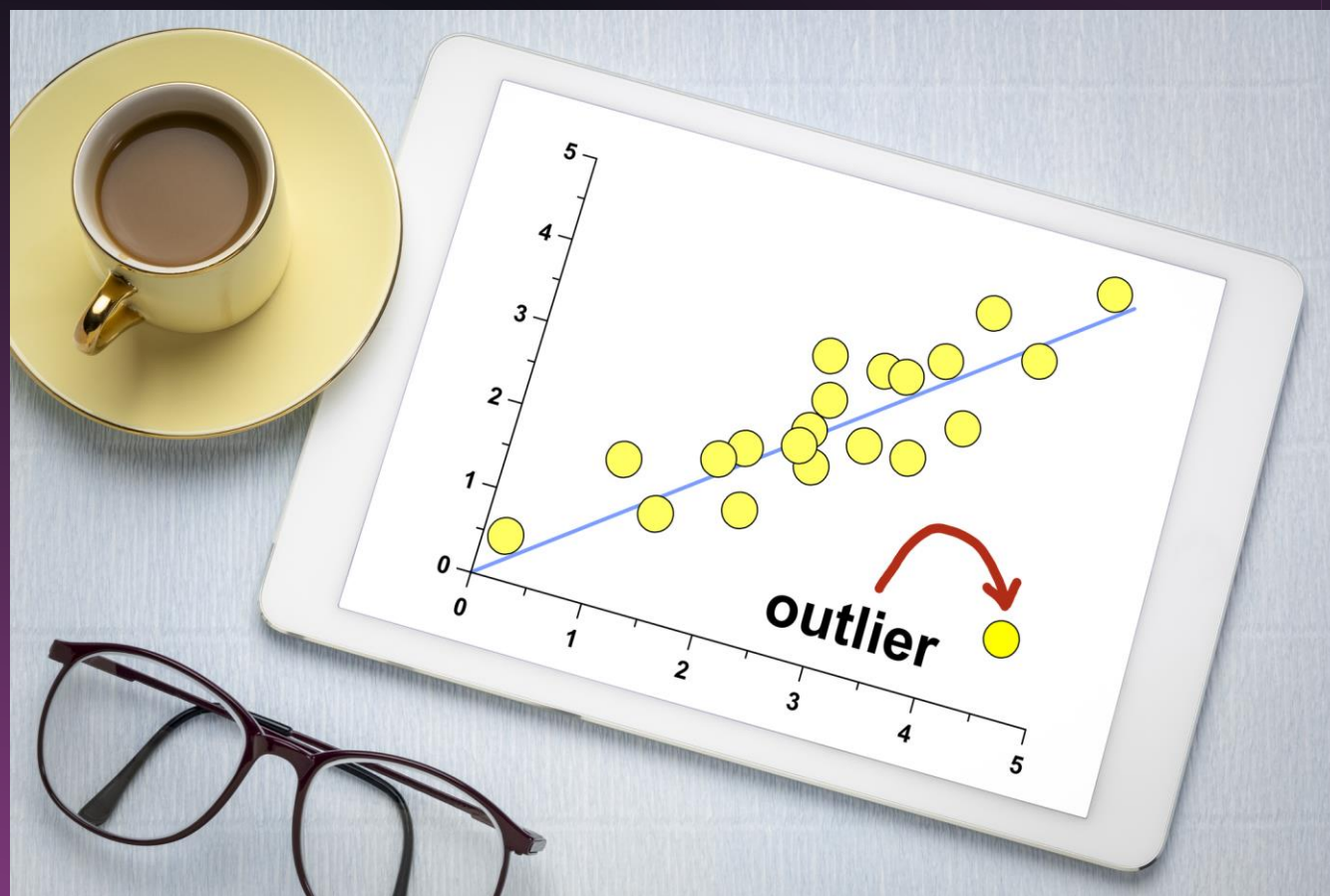
(Ammonia, acetaldehyde, Acetone, Ethylene, Ethanol, Toluene)

8 engineered features were taken from the raw electrical signal of each sensor resulting in a total of 128 predictor variables, and two target variables (chemical species & concentration)

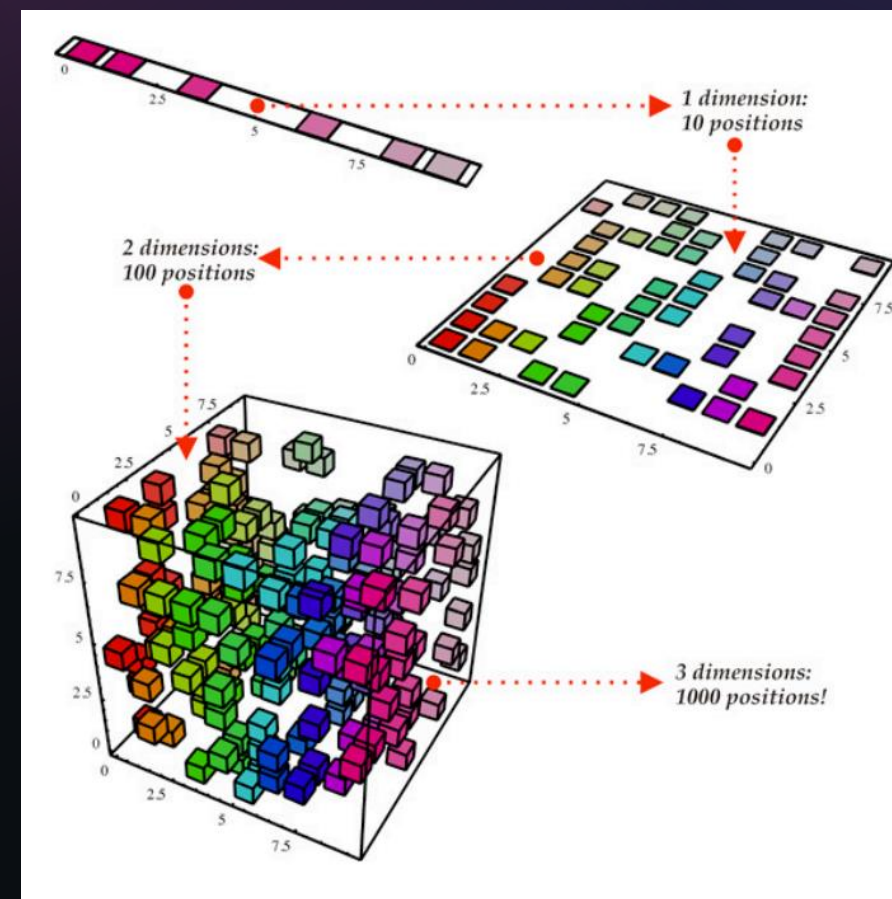


Data Exploration

Outliers

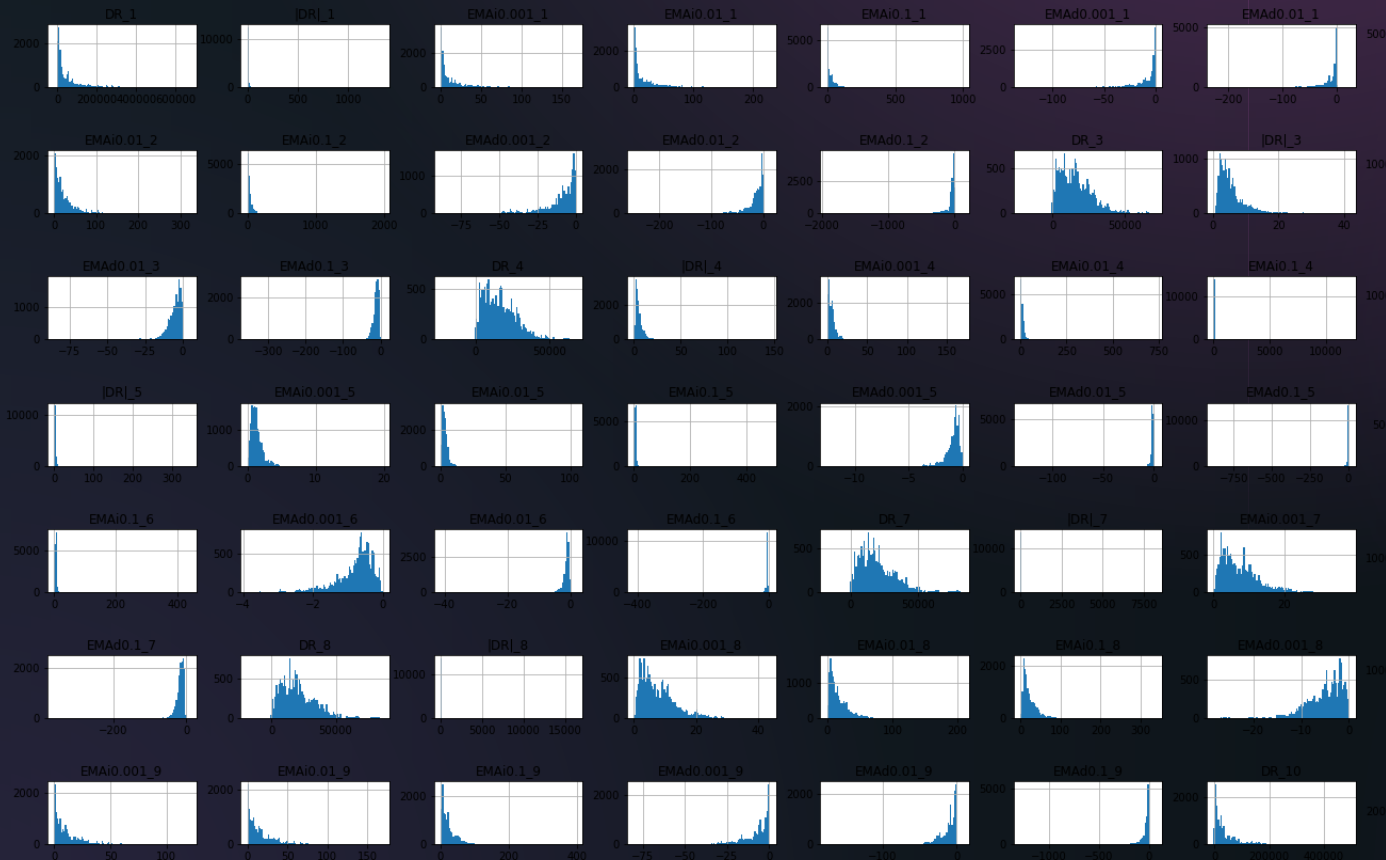


Dimensionality Reduction



Outliers

Selected predictor features distributions



Definition: values more than 1.5 times the IQR outside of the IQR (25 percentile to 75 percentile range)

35.3% of experimental runs had at least one feature that had an outlying value

Visually represented by the amount of white space in the distributions

Numerically represented with high Kurtosis values

Performed hypothesis tests to determine if a target variable was causing outliers

Outlier Hypothesis Tests

Note: Hypothesis testing was performed using 10,000 bootstrap samples

Null Hypothesis	P-value	Results
Outlier values are not associated with specific chemical species	N=6 -> p-value < 0.008	Acetone (chemical 5) is associated with outlying values
Outlier values are not associated with specific chemical concentrations	N=59 -> p-value < 0.00085	Concentrations > 350 ppm are associated with outlying values
Outlier values are not associated with specific test batches	N=10 -> p-value < 0.005	Experimental batches 1, 4 and 10 are associated with outlying values

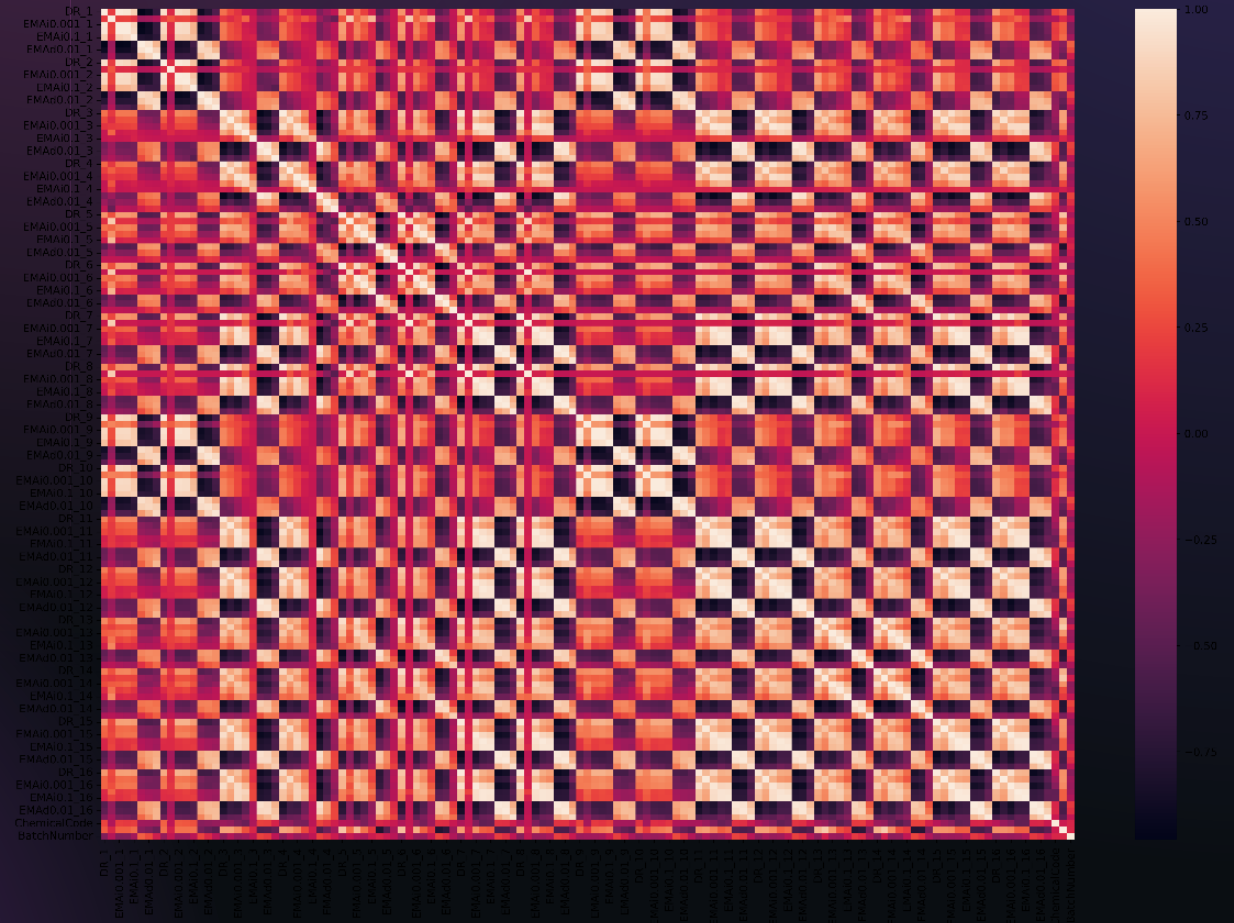
Dimensionality Reduction

Many of the predictor features were highly correlated with other features

High correlation is bad if we plan on making evaluating a linear regression model since it can lead to high variance in the weights

The feature set can be reduced by choosing one of the correlated features or doing a principal component analysis

Correlation Heatmap of Variables



Data Preparation

Data Splitting

- Data was split into three groups: train, test, and future
- Train and test groups were randomly chosen samples from the first five experimental batches (3,600 rows)
- The size of train and test was purposefully varied to determine the response on model accuracy
- Future group was used to evaluate on-going model performance

Feature Selection

- Various sensor combination feature sets are needed to determine the required number of sensors
- There are 65,519 potential sensor combinations
- A basic decision tree model was used to determine which sensors were the most impactful
- Combinations of these sensors were used in the evaluation

Baseline Model Performance

need a multi-class prediction model to determine the chemical species

Two simple models were used as a baseline to compare against complex models

Random Selection

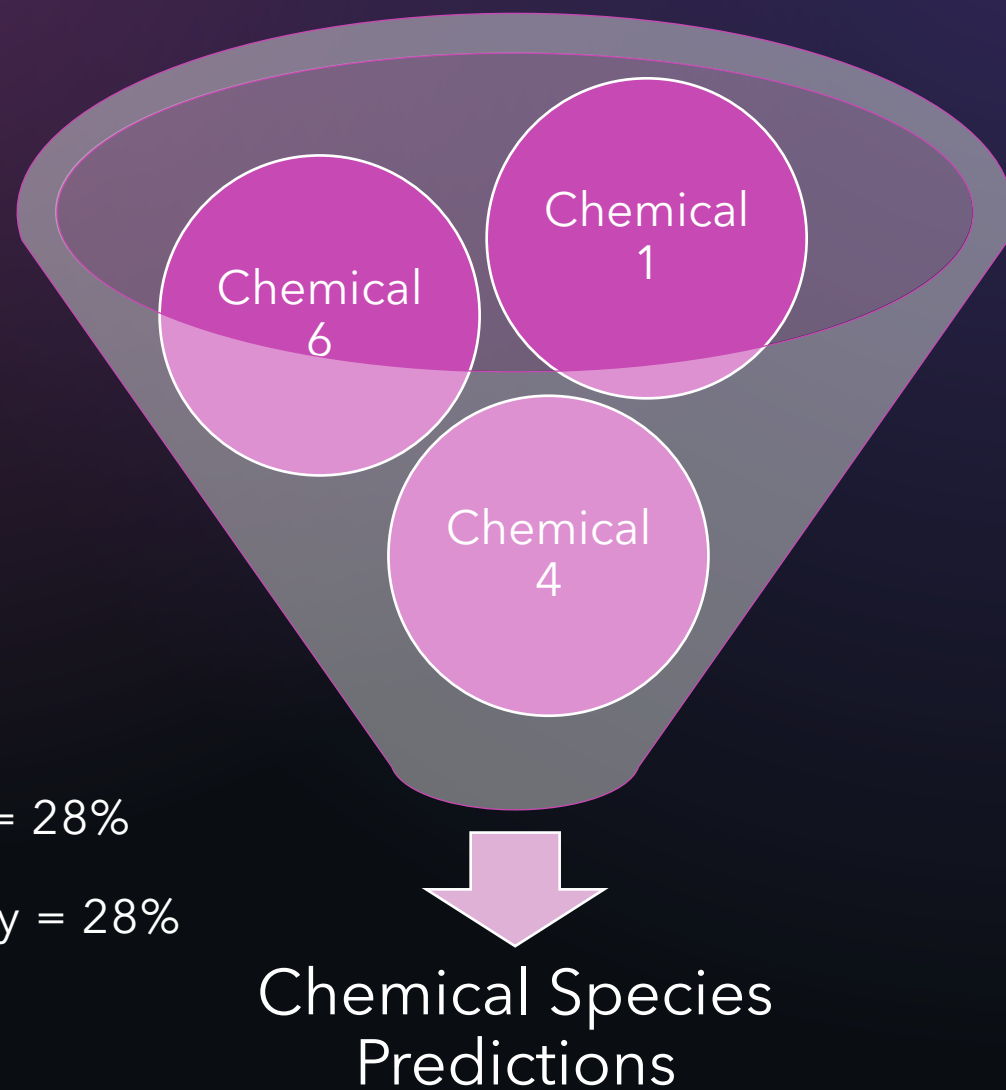
Theoretical Accuracy = 16.7%

Actual Model Accuracy = 15%

Most Frequent

Theoretical Accuracy = 28%

Actual Model Accuracy = 28%



Machine Learning Model Selection

- Five model algorithms were evaluated
- Linear SVC, k Nearest Neighbors, Random Forest, Gaussian Naïve Bayes, Decision Tree
- Evaluated using the same training and test data

Conclusion: Random Forest performs the best

Model Type	Fit Time	Score Time	Accuracy
Decision Tree	0.2933	0.0017	0.966
Random Forest	1.6627	0.0232	0.989
kNN	0.0848	0.1237	0.978
Linear SVC	0.6342	0.0055	0.983
Gaussian Naïve Bayes	0.0093	0.0032	0.803

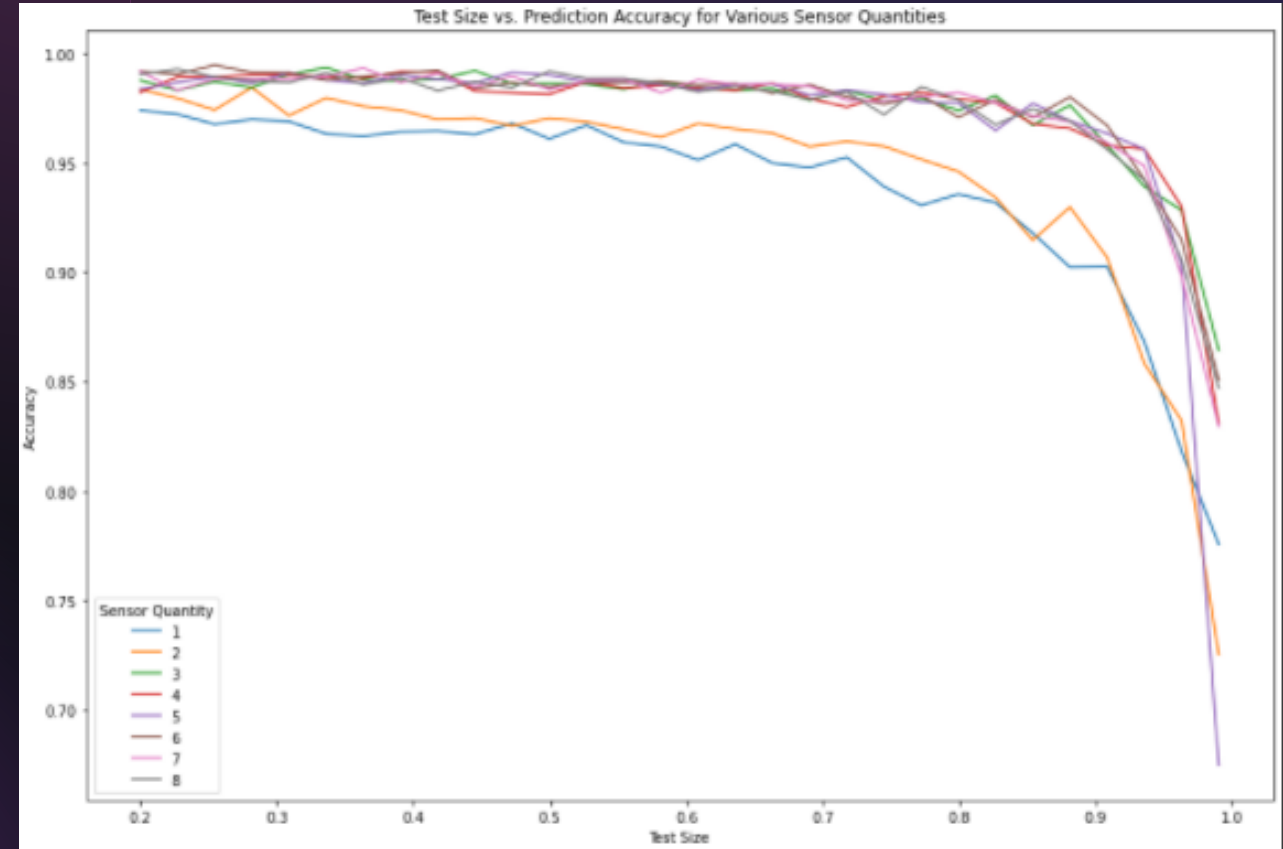
Model Tradeoffs

Tradeoff Setup

- 30 test sizes between 0.2 and 0.99
- 8 sensor combinations with each combination having an additional sensor
- Goal is to have >99% accuracy

Tradeoff Conclusions

- Need at least 3 sensors
- Sharp drop off in model accuracy when less than 800 rows are used in the training set



Poor Long-Term Performance



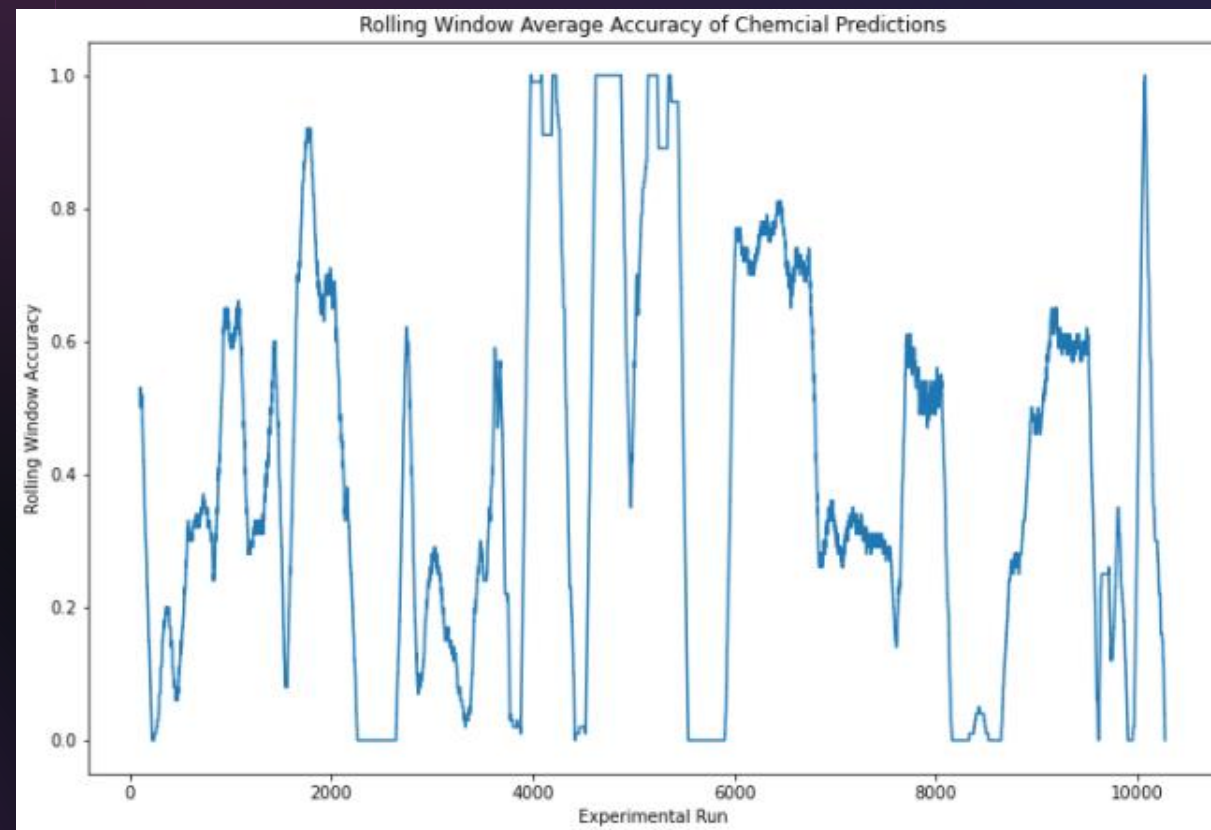
Based on previous results, a model using 5 sensors and a test size of 50% was created



After batch 5 the model accuracy dramatically decreases

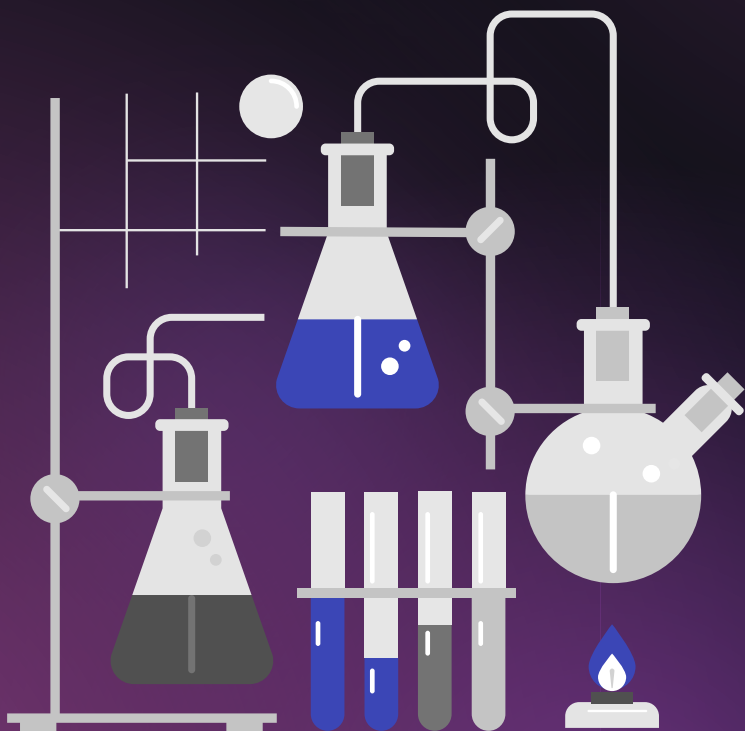


Unsure if this is due to over-fitting or a change in the test system

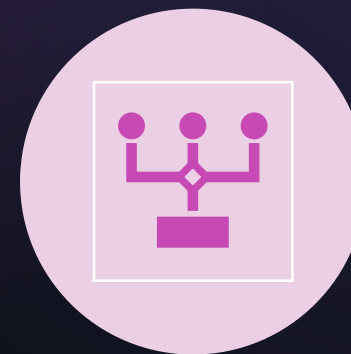


Future Work

PREDICT CHEMICAL
CONCENTRATION



IMPROVE LONG-TERM
PERFORMANCE



DID THE SYSTEM
CHANGE OR IS IT A
MODEL ISSUE?