A project report on

Shear Thinning Fluid Data Analysis using Classification ML Technique

Submitted by

PRAMIKHA K

Department of CSE IV sem / II year

RAMAPURAM

SRM

INSTITUTE OF SCIENCE & TECHNOLOGY

(Deemed to be University u/s 3 of UGC Act, 1956)

Department of Computer Science & Engineering

SRM IST Ramapuram

IIT GUWAHATI

DEPARTMENT OF CHEMICAL ENGINEERING

CERTIFICATE OF INTERNSHIP

This is to certify that **Ms. Pramikha K s**tudying in IV sem/ 2nd year in **COMPUTER SCIENCE & ENGINEERING** has satisfactorily completed project on **SHEAR THINNING FLUID DATA ANALYSIS USING CLASSIFICATION ML TECHNIQUE** under the guidance of professor **Dr. Senthilmurugan Subbiah** from 3 June 2023 to 15 July 2023.

SIGNATURE

ACKNOWLEDGEMENT

I would like to convey my heartfelt gratitude to Mr. Shanmugam V and Mr.Dinesh Gautam for their tremendous support and assistance in the completion of my project. I would also like to thank our **professor Senthilmurugan Subbiah,** for providing me with this wonderful opportunity to work on this project during the internship. The completion of the project would not have been possible without their help and insights.

PRAMIKHA K

BIBLIOGRAPHY

For successfully completing my project file. I have taken help from the following website links: -

www.google.com

Google scholar

ResearchGate

Scopus

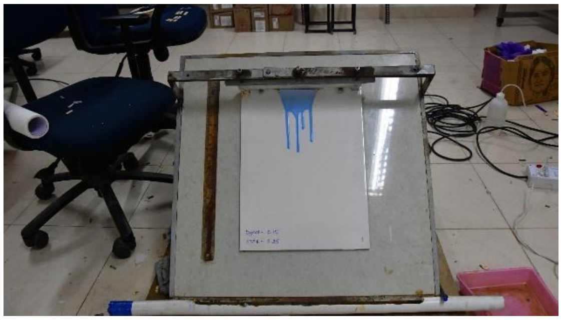
INDEX

ABSTRACT

The field of chemical engineering is data-rich. For the purpose of comprehending flow patterns, creating empirical models, engineering and optimizing chemical reactions, and monitoring and managing chemical processes and systems, practitioners gather and analyze data. It appears inevitable that artificial intelligence and machine learning would play a significant role in the field of chemical engineering given the high data dependence. Chemical engineering has undergone a revolution because of machine learning (ML) approaches that allow for data-driven modeling, optimization, and decision-making. We analyze the data collected for the shear thinning fluid displacement in an inclined plane and find the best fit ML model with highest accuracy. We particularly use the classification technique in our case. Classification machine learning algorithms have several applications in chemical engineering like Fault detection and diagnosis, quality control, process optimization, chemical compound classification and environmental monitoring.

INTRODUCTION

This experimental study involves getting shear thinning fluid to flow over a board. It is known to flow erratically, forming formations resembling fingers. Stable region or area without fingers is the term used to describe the absence of finger-like structures. Different fluids like Dynol, Non-ionic, Pushar, Hydrogel, Silica nanoparticles, CTAB with varying concentrations are added and their surface tension, contact angle and average viscosity are measured.



Dynol:0.15,CTAB:0.25

non ionic:0.8, hydro:0.00625

The entire experiment uses constant volume 2.5 ml of shear thinning fluid. The flow has been monitored for the predetermined duration of 3 minutes, and a picture is taken to allow for detailed examination. Once the experiment is complete, the tile surface is cleaned with IPA and tissue paper before being wiped down with a cotton towel. The tile surface is then dried using a commercial dryer, which helps to ensure the consistency of the fluid designed to thin shear. The same process is performed for each formulation, and a thorough study of the recorded images and videos is done to comprehend how the flow behaves when there are instabilities. 86 such data points were obtained in the experiment.

ANALYSIS OF RESULTS OBTAINED

For the study of shear thinning fluid data, inferential findings were obtained using Matlab. Different techniques were used, and in each instance the one with the best accuracy rate was recorded. Ten percent of the data was utilized for testing, while the remaining ninety percent was used for validation. 5-fold cross validation was used.

On the dataset, we experimented with applying several ML algorithms. However, the majority of them did not produce enough R Squared (Validation) findings. Finally, when the clustering approach was used, it provided accuracy better than 80% in a number of instances.

We will be approaching problem solving in 4 cases. The best results of each of the 4 cases will be recorded and observed. The overall best case will also be chosen based on accuracy.
Here, we have 3 derived inputs namely surface tension, contact angle and viscosity. The derived inputs are estimated after the composition is prepared. Whereas the other 6 inputs - Non ionic, Dynol, Pushar, Hydrogel, Silica Nanoparticles, CTAB are at experimental conditions.

During the classification process we define 2 labels. These labels(P,Q) are namely increase in effective area w.r.t to shear thinning fluid and another is decrease in finger length w.r.t. shear thinning fluid. If the effective area increases, the column gets value 0, if decreases then 1
If the finger length decreases then the column gets 1, if increases then 0
These 2 columns with 0 and 1 values will be used as output features for all 4 cases while performing in matlab.

We will also look at the results table of each of the cases in detail. We will also analyze the ROC curve and Confusion (Validation) matrix of the model yielding the highest accuracy in each of the cases.

ROC Curve

An ROC curve is a visual depiction of the effectiveness of a binary classifier. At various categorization criteria, it plots the true positive rate vs the false positive rate. The curve demonstrates how sensitivity and specificity can be traded off. The classifier's overall performance is measured by the area under the curve (AUC), with a greater AUC indicating stronger discriminating skills. The ROC curve aids in evaluating and contrasting the efficacy of various categorization methods.

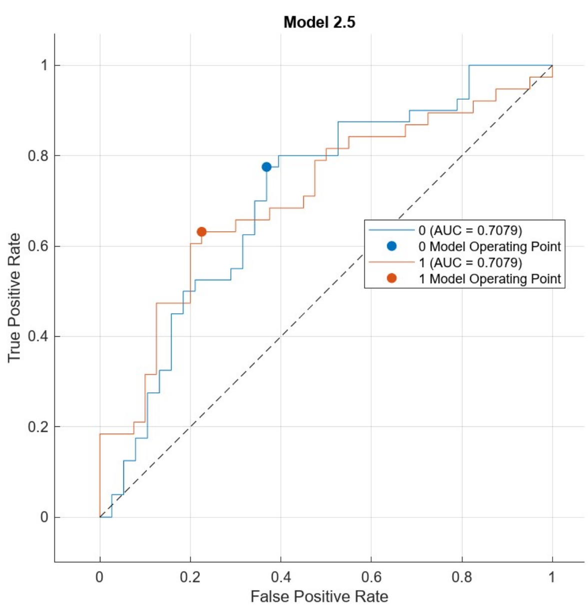
Confusion (Validation) Matrix

A table that lists a classification model's performance is known as a confusion matrix. The anticipated and actual classes for a piece of data are shown in detail. The matrix is split into four quadrants: true positives (instances that were properly predicted to be positive), true negatives (instances that were correctly predicted to be negative), false positives (instances that were wrongly predicted to be positive), and false negatives (instances that were incorrectly predicted to be negative). The matrix aids in assessing the model's F1 score, recall, accuracy, and precision. It offers perceptions into the kinds of faults the model is producing and can direct more research or model development.

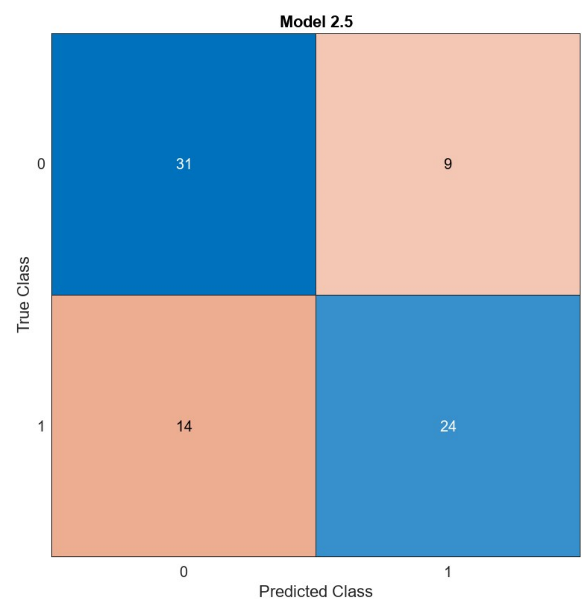
Now let's look at the 4 cases of our analysis

Case 1: 6 inputs vs 1 output increase in effective area w.r.t shear thinning fluid

Here, the 6 parameters - non ionic, dynol, pushar, Hydrogel, Silica Nanoparticles, CTAB are used as input features. The column P which is an increase in effective area w.r.t Shear thinning fluid is the output feature. We ran matlab and trained all the available models with 5 fold cross validation. The results table and summary tables are obtained, From the summary table we can infer which model gives the highest accuracy.



ROC Curve



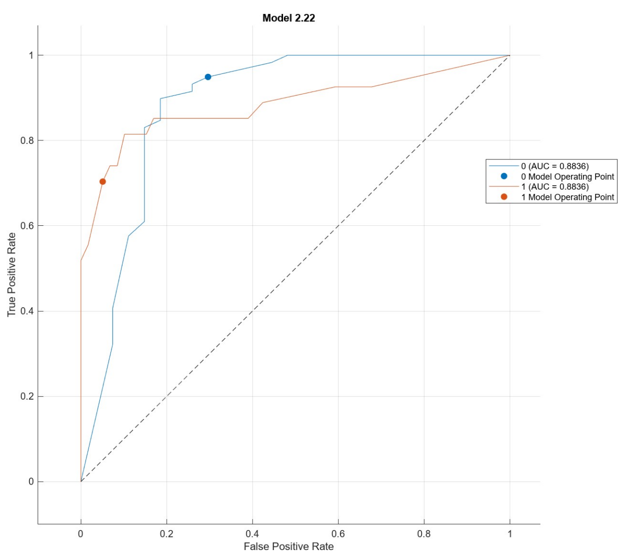
Confusion Matrix

The above 2 graphs represent the ROC curve and Confusion matrix of this case. When we look at the results table below we can conclude that ***Gaussian Naive Bayes*** is the most efficient model with an accuracy of **80.76%.**

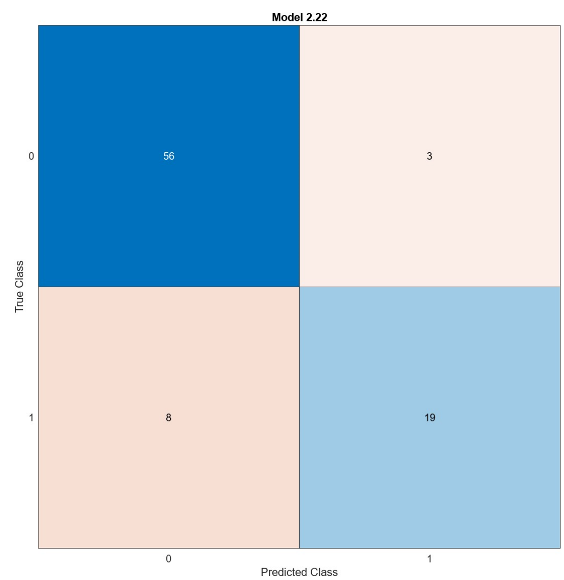
Model Type	Accuracy % (Validation)	Total Cost (Validation)
Fine Tree	70.51282051	23
Medium Tree	70.51282051	23
Coarse Tree	69.23076923	24
Linear Discriminant	71.79487179	22
Quadratic Discriminant	70.51282051	23
Logistic Regression	76.92307692	18
Gaussian Naive Bayes	80.76923077	15
Kernel Naive Bayes	73.07692308	21
Linear SVM	66.66666667	26
Quadratic SVM	70.51282051	23
Cubic SVM	69.23076923	24
Fine Gaussian SVM	64.1025641	28
Medium Gaussian SVM	71.79487179	22
Coarse Gaussian SVM	73.07692308	21
Fine KNN	67.94871795	25
Medium KNN	70.51282051	23
Coarse KNN	51.28205128	38
Cosine KNN	57.69230769	33
Cubic KNN	57.69230769	33
Weighted KNN	70.51282051	23
Boosted Trees	64.1025641	28
Bagged Trees	71.79487179	22
Subspace discriminant	74.35897436	20
Subspace KNN	66.66666667	26
RusBoosted Trees	67.94871795	25
Narrow Neural Network	73.07692308	21
Medium Neural Network	75.64102564	19
Wide Neural Network	73.07692308	21
Bilayered Neural Network	71.79487179	22
Trilayered Neural Network	74.35897436	20
SVM Kernel	66.66666667	26

Case 2: 6 inputs vs 1 output decrease in finger length w.r.t Shear thinning fluid

Here, the 6 parameters - non ionic, dynol, pushar, Hydrogel, Silica, CTAB are used as input features. The column P which is decrease in finger length w.r.t shear thinning fluid is the output feature. We run matlab and train all the available models with 5 fold cross validation. The results table and summary tables are obtained, From the summary table we can infer which model gives the highest accuracy.



ROC Curve



Confusion Matrix

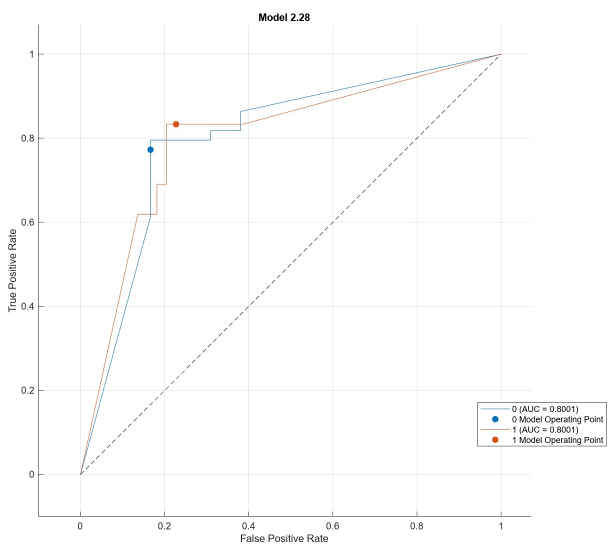
The above 2 graphs represent the ROC curve and Confusion matrix of this case. When we look at the results table below we can conclude that ***Ensemble Bagged Trees*** is the most efficient model with an accuracy of **87.209%.**

Results Table:

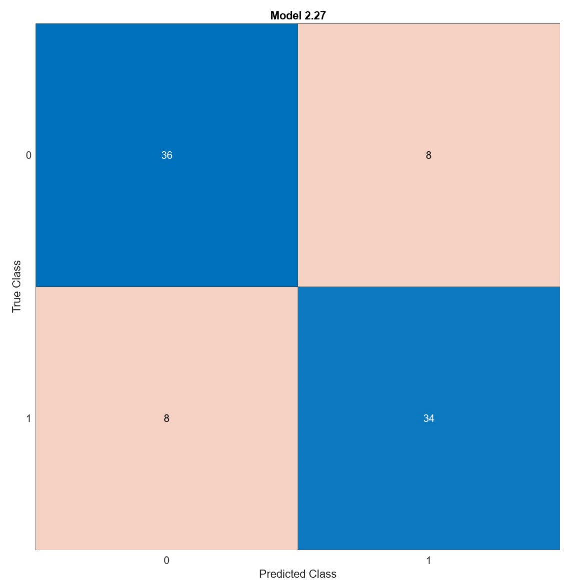
Model Type	Accuracy % (Validation)	Total Cost (Validation)
Fine Tree	82.55813953	15
Medium Tree	82.55813953	15
Coarse Tree	83.72093023	14
Linear Discriminant	82.55813953	15
Quadratic Discriminant	80.23255814	17
Logistic Regression	82.55813953	15
Gaussian Naive Bayes	81.39534884	16
Kernel Naive Bayes	70.93023256	25
Linear SVM	81.39534884	16
Quadratic SVM	79.06976744	18
Cubic SVM	81.39534884	16
Fine Gaussian SVM	68.60465116	27
Medium Gaussian SVM	82.55813953	15
Coarse Gaussian SVM	68.60465116	27
Fine KNN	68.60465116	27
Medium KNN	77.90697674	19
Coarse KNN	68.60465116	27
Cosine KNN	74.41860465	22
Cubic KNN	68.60465116	27
Weighted KNN	83.72093023	14
Boosted Trees	68.60465116	27
Bagged Trees	87.20930233	11
Subspace discriminant	80.23255814	17
Subspace KNN	70.93023256	25
RusBoosted Trees	86.04651163	12
Narrow Neural Network	76.74418605	20
Medium Neural Network	81.39534884	16
Wide Neural Network	79.06976744	18
Bilayered Neural Network	80.23255814	17
Trilayered Neural Network	83.72093023	14
SVM Kernel	76.74418605	20

Case 3: Three inputs vs 1 output increase in effective area w.r.t shear thinning fluid

Here, the 3 derived parameters - Surface Tension, Contact Angle and Viscosity are used as input features. The column P which is increase in effective area w.r.t shear thinning fluid is the output feature. We run matlab and train all the available models with 5 fold cross validation. The results table and summary tables are obtained, From the summary table we can infer which model gives the highest accuracy.



ROC Curve



Confusion Matrix

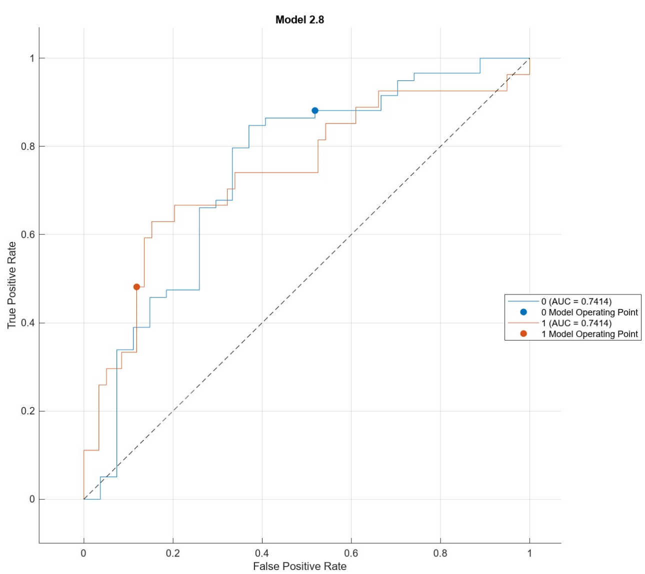
The above 2 graphs represent the ROC curve and Confusion matrix of this case. When we look at the results table below we can conclude that ***Medium Neural Network*** is the most efficient model with an accuracy of **81.395%**

Results Table:

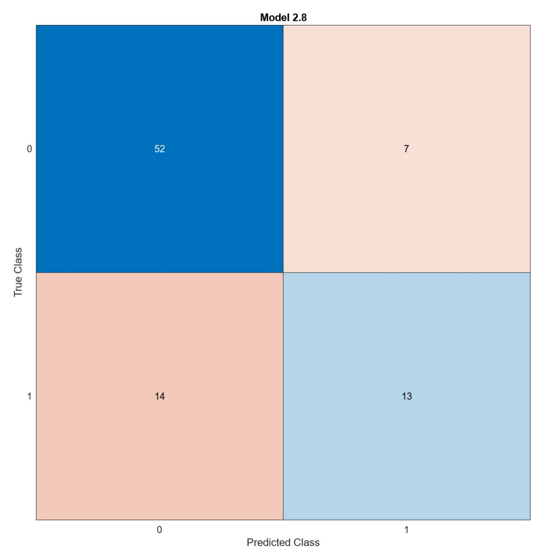
Model Type	Accuracy % (Validation)	Total Cost (Validation)
Fine Tree	74.41860465	22
Medium Tree	74.41860465	22
Coarse Tree	79.06976744	18
Linear Discriminant	74.41860465	22
Quadratic Discriminant	73.25581395	23
Logistic Regression	75.58139535	21
Gaussian Naive Bayes	69.76744186	26
Kernel Naive Bayes	72.09302326	24
Linear SVM	75.58139535	21
Quadratic SVM	73.25581395	23
Cubic SVM	70.93023256	25
Fine Gaussian SVM	67.44186047	28
Medium Gaussian SVM	76.74418605	20
Coarse Gaussian SVM	74.41860465	22
Fine KNN	72.09302326	24
Medium KNN	73.25581395	23
Coarse KNN	51.1627907	42
Cosine KNN	73.25581395	23
Cubic KNN	73.25581395	23
Weighted KNN	70.93023256	25
Boosted Trees	51.1627907	42
Bagged Trees	75.58139535	21
Subspace discriminant	73.25581395	23
Subspace KNN	79.06976744	18
RusBoosted Trees	52.3255814	41
Narrow Neural Network	73.25581395	23
Medium Neural Network	81.39534884	16
Wide Neural Network	75.58139535	21
Bilayered Neural Network	70.93023256	25
Trilayered Neural Network	70.93023256	25
SVM Kernel	70.93023256	25

Case 4: Three inputs vs 1 output w.r.t shear thinning fluid

Here, the 3 derived parameters - Surface Tension, Contact Angle and Viscosity are used as input features. The column P which is decrease in finger length w.r.t shear thinning fluid is the output feature. We run matlab and train all the available models with 5 fold cross validation. The results table and summary tables are obtained, From the summary table we can infer which model gives the highest accuracy.



Naive Bayes ROC Curve



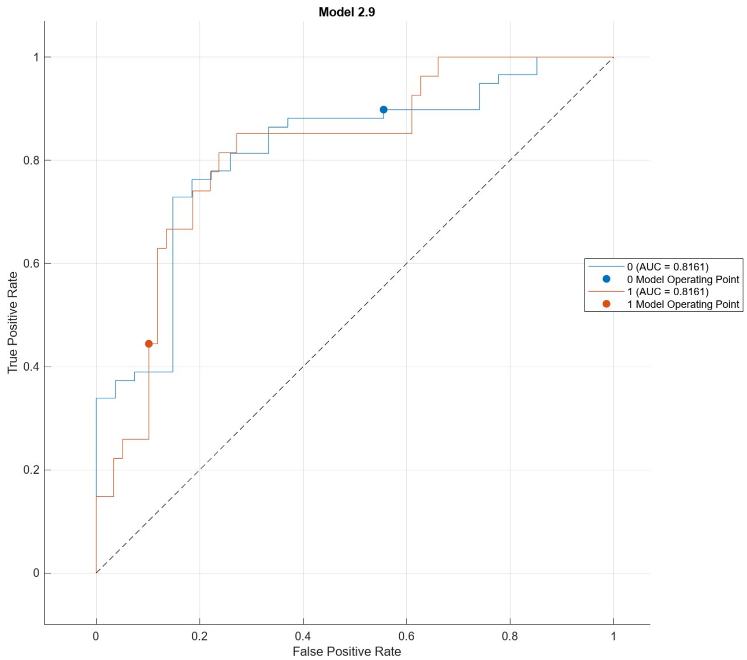
Naive Bayes Confusion matrix

In this particular case, we obtain 2 different models having the same accuracy, both ***kernel Naive Bayes*** and ***Linear SVM*** have the same accuracy of **75.581%.** Hence, we can deploy any of the 2 models .

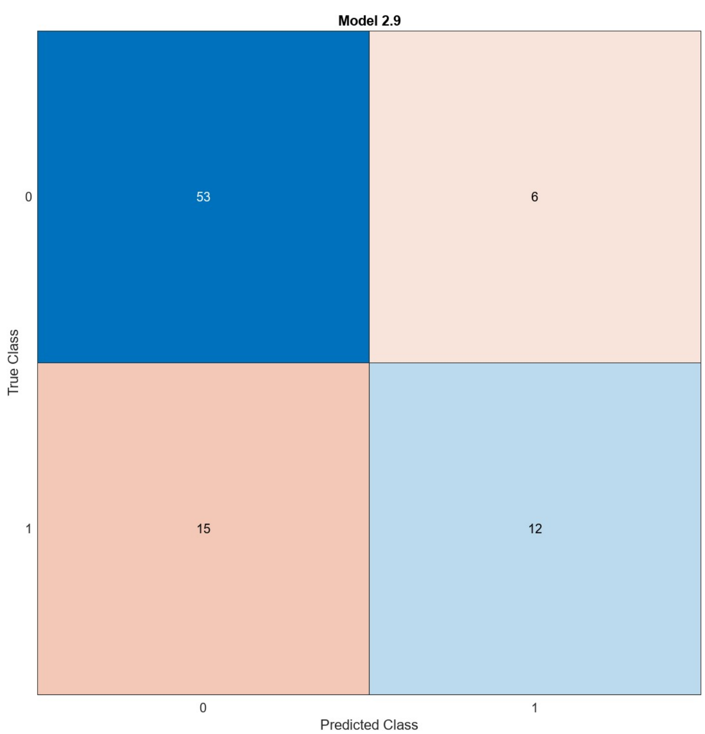
Results Table:

Model Type	Accuracy % (Validation)	Total Cost (Validation)
Fine Tree	67.44186047	28
Medium Tree	67.44186047	28
Coarse Tree	70.93023256	25
Linear Discriminant	74.41860465	22
Quadratic Discriminant	72.09302326	24
Logistic Regression	74.41860465	22
Gaussian Naive Bayes	73.25581395	23
Kernel Naive Bayes	75.58139535	21
Linear SVM	75.58139535	21
Quadratic SVM	68.60465116	27
Cubic SVM	74.41860465	22
Fine Gaussian SVM	65.11627907	30
Medium Gaussian SVM	72.09302326	24
Coarse Gaussian SVM	70.93023256	25
Fine KNN	65.11627907	30
Medium KNN	66.27906977	29
Coarse KNN	68.60465116	27
Cosine KNN	69.76744186	26
Cubic KNN	67.44186047	28
Weighted KNN	69.76744186	26
Boosted Trees	68.60465116	27
Bagged Trees	68.60465116	27
Subspace discriminant	73.25581395	23
Subspace KNN	67.44186047	28
RusBoosted Trees	68.60465116	27
Narrow Neural Network	66.27906977	29
Medium Neural Network	63.95348837	31
Wide Neural Network	63.95348837	31
Bilayered Neural Network	72.09302326	24
Trilayered Neural Network	73.25581395	23

Linear SVM ROC Curve and Confusion Matrix:



Linear SVM ROC Curve



Linear SVM Confusion Matrix

APPLICATIONS OF CLASSIFICATION IN CHEMICAL ENGINEERING:

Chemical engineering uses classification machine learning methods in a variety of ways. Here are a few instances:

1. **Fault Detection & Diagnosis:** Classification models may be used to identify and categorize problems in chemical processes. Fault Detection and Diagnosis. The model may learn patterns and correlations between process variables and known defects by being trained on historical data. In order to enable prompt intervention and the avoidance of expensive failures, the model can then categorize fresh observations and determine the type of problem happening in real-time.

2. **Quality Control**: Classification algorithms can help with quality control by categorizing goods or samples according to their quality characteristics. The quality of fresh samples or goods can be predicted using labeled data that specifies several quality levels. This aids in locating flawed goods or samples that fall short of quality requirements.

3. **Process Optimization:** By categorizing the operating circumstances or settings that produce desired results, classification models may be used to optimize chemical processes. The best settings for accomplishing certain goals, such as the highest yield, the least amount of energy used, or the lowest emissions, may be predicted by training the model on data with various process conditions and related outcomes.

4. **Classification of Chemical Compounds**: Chemical compounds can be categorized according to their characteristics or behaviors using classification algorithms. This can help with environmental analysis, material design, and medicine development. The model may predict the properties of novel compounds by being trained on known compounds and their qualities. This helps in the selection or creation of compounds with the required attributes.

5. **Environmental Monitoring**: To categorize pollution incidents or forecast environmental conditions, classification models can be used in environmental monitoring. The model can categorize real-time data into distinct pollution categories or forecast the likelihood of pollution incidents by being trained on data gathered from multiple sensors and past pollution events. This aids in proactive environmental management.

These are just a few instances of how chemical engineering might use classification machine learning. Classification can be applied to various chemical engineering data sets.

In conclusion, chemical engineering applications of machine learning have the potential to change a number of areas of the discipline. Machine learning may improve quality control procedures, accelerate chemical reactions, categorize chemical compounds, and enable proactive environmental monitoring by utilizing algorithms and models. The chemical engineering sector may benefit from these applications through higher productivity, reduced expenses, better product quality, and more environmentally friendly business practices. Chemical engineering stands to benefit significantly from the incorporation of machine learning as it develops and becomes more innovative.

REFERENCES

[1]Dobbelaere, M. R., Plehiers, P. P., Van de Vijver, R., Stevens, C. V., & Van Geem, K. M. (2021). Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. *Engineering*, *7*(9), 1201-1211.

[2]Schweidtmann, A. M., Esche, E., Fischer, A., Kloft, M., Repke, J. U., Sager, S., & Mitsos, A. (2021). Machine learning in chemical engineering: A perspective. *Chemie Ingenieur Technik*, *93*(12), 2029-2039.

[3] Venkatasubramanian, V. (2019). The promise of artificial intelligence in chemical engineering: Is it here, finally?. *AIChE Journal*, *65*(2), 466-478.

[4]Thebelt, A., Wiebe, J., Kronqvist, J., Tsay, C., & Misener, R. (2022). Maximizing information from chemical engineering data sets: Applications to machine learning. *Chemical Engineering Science*, *252*, 117469.

[5]Qin, S. J., & Chiang, L. H. (2019). Advances and opportunities in machine learning for process data analytics. *Computers & Chemical Engineering*, *126*, 465-473.