www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

COMPARATIVE ANALYSIS OF ANALYTICAL METHODS IN PRODUCTION ON BIOETHANOL: SUPERVISED MACHINE LEARNING AND DESIGN OF EXPERIMENTS

Vishal Murali[1]

[1]Institute/Organization, BV Raju Institute of Technology, Narsapur

DOI: https://www.doi.org/10.58257/IJPREMS37051

ABSTRACT

Bioethanol is a renewable, eco-friendly, and cost-effective alternative to fossil fuels derived from various biomass sources such as corn, sugarcane molasses, and other cellulosic materials. The usage and production of bioethanol has gained traction in recent years with good promise for future generations to come. In the Indian market, gasoline outlets are beginning to transition to a 20% blend of bioethanol called E-20 with gasoline from a 10% blend (E-10). This is done to maintain the calorific value of the blend along with reducing the emission of CO (carbon monoxide) into the environment. The study focuses on conducting a comparative analysis of analytical methods in the production of bioethanol, specifically aiming to compare the production of bioethanol from cellulosic material (Psidium Gujava) using Saccharomyces Cerevisiae (S. Cerevisiae). This is achieved by first collecting the leaves of Psidium Gujava and treating them initially to enhance the content of cellulose in the stock solution. After pre-treating the solution, the solution is sterilized and inoculated with S.Cerevisiae in the form of over-the-counter Bakers' Yeast. Post-fermentation, the yield is purified, and yield is measured using a UV-Vis Spectrophotometer. Post the production of bioethanol, the yield concentrations are collected, interpolated to the required amount and analyzed using a Design of Experiments approach.

The dataset collected in the form of CSV is put through various algorithms to predict the yield. The algorithms are developed, trained, and tested in Python using the sci-kit learn module in the case of Supervised learning models, and in the case of neural network regression, the algorithm will be developed in Python's TensorFlow module using Google Colab.

Regression algorithms like K-Nearest neighbors' regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression were some of the models that were developed, trained, and tested and yielded a promising result with a good prediction of the bioethanol yield. Amongst the models above, the random forest regressor showed more accurate results with a cleaner prediction. Analysis of different built-in kernel types in Support Vector Machine Regression was also performed where the radial-based function (RBF) kernel showed more promise in terms of accuracy but the prediction value, the maximum predicted value was at around 10mg/ml whereas the actual maximum concentration value was at around 16 mg/ml. This situation was not the case for the other kernels, but the accuracy dropped drastically. Apart from the accuracy values, mean squared error, average of errors, and standard deviation in errors were also taken into consideration for analysis purposes.

Keywords: Bioethanol, Fermentation, Design of Experiments, Supervised learning, Machine Learning

1. INTRODUCTION

Bioethanol is a renewable, eco-friendly, and cost-effective alternative to fossil fuels. It is derived from various biomass sources such as corn, sugarcane, and other lignocellulosic materials. The production of bioethanol has gained traction in recent years in developed and developing countries where bioethanol is blended with gasoline to produce cleaner by-products of combustion occurring in an IC Engine. The production process involves many stages such as pre-treatment, fermentation, and post-treatment methods. Optimization of the process is crucial to maximizing the bioethanol yield.

Supervised machine learning (SML) algorithms such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) have been widely used in various fields including Bioinformatics, genetics, and drug discovery to identify the patterns and relationships in labeled datasets. The result from the experimental method is collectively smaller in size, the data is interpolated to the needful using interpolation methods. The experimental process begins by collecting the leaves of Psidium Gujava commonly known as guava. The collected leaves are pre-treated and impurities are removed to increase cellulose accessibility to the yeast to act on. Fermentation is carried out using S.Cerevisiae to produce bioethanol from the stock cellulose solution. This paper aims to review the production and optimization studies of bioethanol using SML analysis, with a specific focus on Psidium Gujava leaves as a potential source of cellulose. The paper will begin by discussing the production process of bioethanol and the challenges faced in each stage. Then, it will review various SML algorithms that have been used to optimize and predict the yield of the product. The paper will conclude by summarizing the key findings of production and prediction of the yield and potential benefits acquired by using this source of cellulose for bioethanol production.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

2. BIOETHANOL SYNTHESIS

Bioethanol is synthesized by fermenting the stock solution, which is prepared by submerging the leaves of Psidium Gujava inside the buffer solutions made at different pH levels using a standard buffer chart (insert reference here) for 24 hours. The cellulose is absorbed into the solution. The leaves are filtered out and the solution is autoclaved at 100 °C and 2 psi for 10 minutes and transferred to a sterile area to be cooled. S, Cerevisiae in the form of solution is introduced to the cooled solution and is maintained in a sterile environment for 24 hours to allow fermentation of the solution. Post-fermentation time, the samples are analyzed under a UV-Vis spectrophotometer at 600nm. The results were interpolated to obtain an adequate amount of data to be implemented in supervised models.

Table 1: Bio-ethanol Concentration Values

pH	20 °C	30 °C	40 °C	50°C	Room Temperature (°C)
2	3.24	12.068	12.7547	1.529	7.264
3	0.294	9.96	3.196	0.843	7.640
4.5	6.627	12.90	5.647	5.205	5.5
5	11.28	2.46	11.235	3.0998	7.852
5.8	**16.08**	**14.372**	113.686	8.196	8.234
6	3.196	10.450	**14.617**	**16.137**	15.058
6.5	5.892	8.245	11.382	11.382	10.205
7.5	9.029	10.450	8.637	8	11.186
8	8.098	12.607	12.90	12.754	12.90

The concentration of yeast solution was varied at pH 6 and 60 °C and the following yield of bioethanol was obtained.

Table 2: Yeast Concentration Values

Yeast Concentration (ml)	Absorbance Values	Bio-Ethanol Concentration
2	0.239	9.91
4	0.370	16.33
6	0.460	**20.74**
8	0.410	18.29
10	0.385	17.06

The results were interpolated to obtain an adequate amount of data to be implemented in supervised models.

3. DATASET

The dataset was generated from the above data using MATLAB interpolation methods. The data were interpolated to contain approximately 21,000 rows for each of the columns i.e., Temperature, pH, and Concentration. The training-to-test ratio was chosen to be 80:20. The dataset was stored in the form of a comma-delimited file. An issue was faced while splitting the dataset where splitting resulted in the loss of rows with pH values greater than 6 not to be included in the training dataset which resulted in inadequate training of the model. To counter that the dataset's rows were shuffled randomly to evenly distribute and made sure every pH was evenly covered in both the training and test set. In the dataset wherein the temperature and concentration are the features in this situation, the features were separated from the predicted variable (need to find the alternative names for conc) using the "iloc" function which splits the columns in the dataset using the pandas module in Python.

4. DESIGN OF EXPERIMENTS

Design of Experiments (DOE) is a structured and efficient approach to experimentation that enables researchers to identify the relationships between multiple input variables (factors) and key output variables (responses). It helps them understand the factors that influence a process or system and optimize its performance. This is done by using Design Models developed using the experimental data. These models are used to predict the effect of changes in the input variables on the output variables and to identify the factors that have the most significant impact on the response.

4.1 TWO-LEVEL

Two-Level Design of Experiments (DOE) is a statistical technique that is used to optimize a process or product by studying the effect of two factors, each at two levels. This design model is beneficial when the relationship between the

factors and the response variable is expected to be linear. One of the key advantages of the Two-Level DOE is its simplicity, as it involves only two factors and two levels for each factor. This makes it a cost-effective and efficient approach for testing multiple factors simultaneously. The two-level design is written as a **2k** factorial design. It means that k factors are considered, each at 2 levels.

A R-square value of 0.8514 is obtained which indicates favorable accuracy for the model. The Adequate precision measures the noise-to-signal ratio, a ratio of 7.006 indicates an adequate signal and this model can be used to navigate design space.

The obtained Two Level Design Model Characteristic Equation is,

Bioethanol Yield = 16.66 * A+ 4.87 * B - 16.01* C + 4.44

A = pH, B = Temperature & C = Microbial Conc

4.2 CENTRAL COMPOSITE DESIGN

The Central Composite Design (CCD) is a Design of Experiments (DOE) technique used to model the relationship between multiple factors and a response variable. This design model is particularly useful when the response variable is not linear and when the researcher wants to model the curvature of the relationship between the factors and the response variable.

The Central Composite Design involves three types of runs:

1. Factorial design: A full or fractional factorial design is conducted with the chosen factors and their levels.
2. Star design: Additional runs are conducted at the midpoint of each factor range and the corners of the factorial design.
3. Center point design: Several runs are conducted at the center point of the design.

One of the key advantages of the Central Composite Design is its ability to model the curvature of the relationship between the factors and the response variable. This can help researchers identify the optimal settings for each factor with a high degree of precision, leading to improvements in the quality, efficiency, and effectiveness of the process or product being studied.

The number of models required is given by,

$N= 2^n + 2*n + n_c$

n = of factors are required

N = number of runs required

n_c = number of center points

In this case the number of models, $N= 2^3 + 2*3 + 6 = 20$ (8 factorial design, 6 axial design & 6 center points)

The axial value is given by 2^(k/2); where k is the number of factors. The axial value will be 1.682.

An R-square value of 0.8514 is obtained which indicates favorable accuracy for the model. The adequate precision measures the noise-to-signal ratio, a ratio of 7.666 indicates an adequate signal and this model can be used to navigate design space.

The obtained Central Composite Design Model Characteristic Equation is,

Bioethanol Yield = -737.44+33.47*A+2.19*B-1745*C-2.82*AB+38.53*AC+0.9960*BC-0.7987*A2+0.3393*B2-1015.85*C2

A = pH, B = Temperature & C = Microbial Conc

4.3 DEFINITIVE SCREENING

Definitive screening design is a statistical experimental design method used to identify important factors and interactions that affect the output of a system. This type of design involves a two-level design where each factor is either set at a high or low level. The design matrix is constructed in a way that allows for the efficient estimation of main effects and two-factor interactions, while also allowing for the identification of quadratic effects and higher-order interactions. The design matrix can also be augmented with center points to allow for the estimation of error variance and the testing of curvature in the response surface.

Definitive screening designs are different from traditional factorial designs in that they have fewer runs and are more efficient in terms of the number of experiments required to identify significant factors and interactions. This is achieved by using a set of non-estimable interaction effects that can be used to estimate the main effects of the input variables. The number of runs for Definitive screening design is given by,

N = 2k + 1 (for an even number of factors)

N = 2k + 3 (for an odd number of factors)

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

Where k = levels of factors are required

An R-square value of 0.9809 is obtained which indicates favorable accuracy for the model. The adequate precision measures the noise-to-signal ratio, a ratio of 15.6706 indicates an adequate signal and this model can be used to navigate design space.

The obtained Definitive Screening Design Model Characteristic Equation is,

Bioethanol Conc. = 10.37 + 2.54*A + 2.58 * B+ +0.0425 * C+ 0.1395 * D + 5.15 *A^2-0.8824*B^2+ 0.7301C^2-1.16 D^2

A = pH, B = Temperature & C = Microbial Conc

4.4 THREE LEVEL DESIGN MODEL

The Three Factor Design Model is a statistical design of experiments technique used to analyze the impact of three factors on a process or product. This model is useful for identifying the key factors that influence a response variable and optimizing the settings for these factors to improve the overall performance of the process or product. The Three Factor Design Model involves three factors, each with two or more levels. The response variable is measured for each combination of factor levels to determine the impact of each factor on the response variable. Analyzing the data allows for identifying the optimal settings for each factor that leads to the desired response variable. The three-level design is written as a **3k** factorial design. It means that k factors are considered, each at 3 levels. An R-square value of 0.9951 is obtained which indicates favorable accuracy for the model. The adequate precision measures the noise-to-signal ratio, a ratio of 65.5096 indicates an adequate signal and this model can be used to navigate design space. The obtained Three Level Design Model Characteristic Equation is,

Bioethanol Conc. = 10.37 + 2.57*A + 2.64 * B+ 0.945 AB- 5.34 *A^2-1.07*B^2-0.097C^2

A = pH, B = Temperature & C = Microbial Conc

4.5 BOX BEHNKEN METHOD

The Box-Behnken Design is a type of Design of Experiments (DOE) technique that is used to optimize a process or product by studying the interaction between three or more factors, each at three different levels. This design model is beneficial when the relationship between the factors and the response variable is not linear, and a quadratic or curved relationship is suspected. The Box-Behnken Design involves three levels of each factor, including low, middle, and high. It requires fewer experimental runs than a full factorial design, making it a more efficient approach for testing multiple factors simultaneously. One of the key advantages of the Box-Behnken Design is its ability to identify the optimal settings for the factors being studied with a minimal number of experimental runs. This can save time and resources, making it an attractive approach for industrial and scientific applications. An R-square value of 0.9979 is obtained which indicates favorable accuracy for the model. The adequate precision measures the noise-to-signal ratio, a ratio of 56.9675 indicates an adequate signal and this model can be used to navigate design space. The obtained Box Behnken Design Model Characteristic Equation is

Bioethanol Conc. = 10.55 + 2.61*A + 2.72 * B+ 0.945 AB- 5.39 *A2-1.12*B2-0.4513C2

A = pH, B = Temperature & C = Microbial Conc

The Box-Behnken design model is sometimes referred to as a response surface design because it can create a three-dimensional surface that shows the relationship between the input variables and the output response. This response surface can be used to visualize and analyze the impact of different input variables on the output response. For example, below is a response surface diagram for the Box Behnken model between two factors, pH and temperature, and a third actual factor, the concentration.

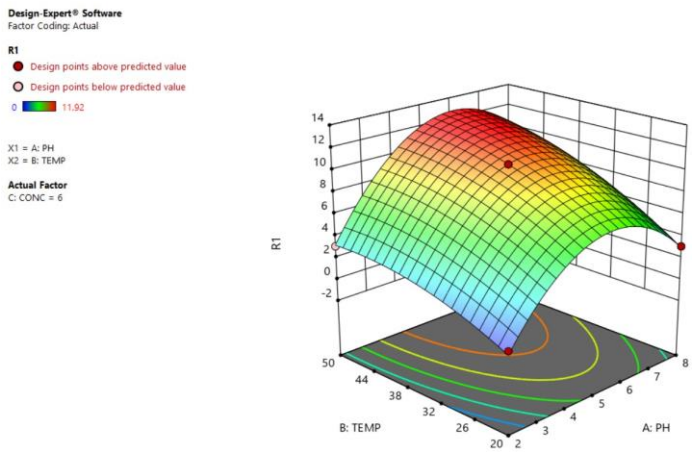

Image 2: Box Behnken Model Response Surface Diagram

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

5. ANALYSIS OF DOE MODELS

5.1 R^2, ADJUSTED R^2 & PREDICTED R^2

Adjusted R-squared is a modified version of R-squared that takes into account the number of predictor variables in the model. It adjusts the R-squared for the number of terms included in the model so that it penalizes the inclusion of unnecessary predictor variables that do not contribute significantly to the model's explanatory power. The adjusted R-squared value is always lower than the R-squared value and is often used as a more conservative measure of the model's explanatory power.

Predicted R-squared, on the other hand, is a measure of the model's predictive power. It is a variation of R-squared that is calculated using cross-validation, where the model is trained on a subset of the data and tested on the remaining data. Predicted R-squared estimates of how well the model will perform on new, unseen data are useful for assessing the model's generalization performance and can be used to compare the predictive power of different models.

Table: 3 R^2 Adjusted R^2 & Predicted R^2 Values

Model	R^2	Adjusted R^2	Predicted R^2
Regular Two-Level Design	0.8514	0.7400	NA
Central Composite Design	0.8386	0.6933	0.0073
Three Level Design	0.9951	0.9932	0.9889
Box-Behnken Design	0.9979	0.9952	0.9661
Definitive Screening Design	0.9809	0.9428	0.8927

5.2 ANOVA (ANALYSIS OF VARIANCE)

ANOVA (Analysis of Variance) is a statistical method used to compare the means of two or more groups and determine if there are any significant differences between them. It can help identify which factors have the biggest impact on a given outcome.

5.3 P-VALUE & F-VALUE

P-value is a statistical measure that helps to determine the probability of observing a test statistic. It is used to determine if the results of a study are statistically significant or not. The F-value, on the other hand, is a statistical test that measures the ratio of variances between two or more groups. It helps to determine if there are significant differences between the means of the groups being compared, and if so, which group(s) differ significantly from the others.

Table:4 Mean Square, F- Values, P-Values

Model	Mean Square	F-value	P-value
Regular Two-Level Design	56.60	7.64	0.0393
Central Composite Design	26.65	5.77	0.0057
Three Level Design	56.93	501.03	< 0.0001
Box-Behnken Design	27.85	366.73	< 0.0001
Definitive Screening Design	25.32	25.71	0.0035

5.4 CONCLUSION

From the above Design of Experiments analysis, it was concluded that both the Box Behnken Model & Three Level Model accurately fit the data. The Box Behnken model has a slightly higher R-square value (0.9979) than the Three Level model (0.9951). However, the adjusted R-square values are very similar for both models (0.9952 and 0.9932, respectively). The predicted R-square values are also very close, with the Box Behnken model having a slightly higher value (0.9661) than the Three-Level model (0.9889). The p-values for both models are very small (< 0.0001), indicating that both models are statistically significant. The F-values for both models are also very large (366.73 and 501.03, respectively), further supporting the conclusion that both models are statistically significant. Finally, the mean square values for both models are relatively small (27.85 and 56.93, respectively), indicating that both models have a low level of variability.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

ANOVA for Quadratic model

Response 1: R1

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	250.68	9	27.85	366.73	< 0.0001	significant
A-PH	54.55	1	54.55	718.21	< 0.0001	
B-TEMP	59.08	1	59.08	777.84	< 0.0001	
C-CONC	0.0000	1	0.0000	0.0000	1.0000	
AB	3.57	1	3.57	47.03	0.0002	
AC	0.0000	1	0.0000	0.0000	1.0000	
BC	0.0000	1	0.0000	0.0000	1.0000	
A^2	122.38	1	122.38	1611.30	< 0.0001	
B^2	5.27	1	5.27	69.38	< 0.0001	
C^2	0.8574	1	0.8574	11.29	0.0121	
Residual	0.5317	7	0.0760			
Lack of Fit	0.5317	3	0.1772			
Pure Error	0.0000	4	0.0000			
Cor Total	251.22	16				

Image 1: ANOVA for Box Behnken Model

ANOVA for Quadratic model

Response 1: Bioyield

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	512.34	9	56.93	501.03	< 0.0001	significant
A-ph	118.97	1	118.97	1047.04	< 0.0001	
B-temp	125.61	1	125.61	1105.53	< 0.0001	
C-conc	0.0000	1	0.0000	0.0000	1.0000	
AB	10.72	1	10.72	94.32	< 0.0001	
AC	0.0000	1	0.0000	0.0000	1.0000	
BC	0.0000	1	0.0000	0.0000	1.0000	
A^2	203.83	1	203.83	1793.98	< 0.0001	
B^2	8.12	1	8.12	71.46	< 0.0001	
C^2	0.0674	1	0.0674	0.5929	0.4495	
Residual	2.50	22	0.1136			
Lack of Fit	2.50	17	0.1470			
Pure Error	0.0000	5	0.0000			
Cor Total	514.84	31				

Image 2: ANOVA for Three-Level Model

From the ANOVA data, it was concluded that the pH and temperature factors significantly affected the yield of bioethanol.

6. SUPERVISED MACHINE LEARNING

The data obtained from the experiment as seen in the table above is interpolated to a subsequent amount of data enough to predict the concentration of bioethanol produced using supervised algorithms. Machine Learning models such as Decision Tree, Support Vector Machine, K-Nearest neighbors, and Random Forest were some of the models used for this data set.

6.1 DECISION TREE

A decision tree is a classification method that builds trees based on the dataset characteristics.. There are 2 types of nodes in the decision tree model. The internal nodes represent the features of the dataset and the leaf nodes refer decisions of the model. The internal and leaf nodes are connected via a branch which usually is the decision rules. One of the questions that arises is why choose a decision tree model. The simple answer to that is that a decision tree model makes choices and also mimics the thought process of a human being while making a decision. Moreover, the understanding of the decision tree model is quite easy. The data is represented in the form of an inverted tree.

6.2 SUPPORT VECTOR MACHINE

Support Vector Machine is a type of machine learning algorithm used in regression analysis. Its working can be simply explained by the best-fitting line or curve (hyperplane) that explains the relationship between input variables and output variables in a given dataset. In SVM or SVR, the data points are transformed into a higher dimensional space where a hyperplane is constructed to separate the data into two classes. The distance between the hyperplane and the points is called the “margin”. The margin is reduced which becomes the main goal while achieving good accuracy in predicting the output variables which is done by selecting the support vectors i.e., the points close to the hyperplane, and using the points to calculate the margin. In this study, SVR is developed, trained, tested, and analyzed using the scikit learn library using the produced data from experimental methods.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

6.3 K-NEAREST NEIGHBOURS

K-Nearest Neighbours Regressor (KNNR) is a simple straightforward algorithm that classifies the data based on the similarity between each other. KNNR calculates the average of the numerical target of the KNN. This method is widely used in statistical estimations, predictions, and pattern recognition. The base logic of this algorithm is identifying a fixed number of training examples that are close to the new data point and using it to predict its label. The number of samples can either be fixed or variable depending on the local density of points in radius-based neighbor learning. Euclidean distance is typically preferred for the KNNR algorithm in this study.

6.4 RANDOM FOREST

Random Forest also commonly known as random decision forest is an ensemble learning technique used in tasks such as classification and regression. It involves creating numerous decision trees during the training phase of the model building which are used to determine the prediction of the individual trees. In this study, sci-kit-learn's Random Forest Regressor (RFR) was used and the RFR was developed, trained, and evaluated on the experimental production data.

The following models above were executed in Python with the aid of the module Sci-Kit Learn which consists of each of the models described above pre-coded on the module. The results are visualized in the form of error plots, the difference in the prediction table, the overall average in the error, standard deviation, probability plots, and the 3D visualization of the prediction against the features.

7. ANALYSIS OF MACHINE LEARNING MODELS

The usage of the above models in the Sci-Kit learn library resulted in different Coefficients of determination (R^2). As we can see from the different prediction plots below (insert prediction plots after the inferences and results) the Random Forest Regressor (RFR) gave the highest possible R^2 followed by the Decision Tree Regressor (DTR), K-Nearest Neighbours Regressor (KNNR) and Support Vector Regressor (SVR) gave the least R^2 possible.

7.1 DECISION TREE

The decision tree regressor (DTR), was a straightforward procedure where the dataset was imported and split into training and testing with a split ratio of 0.20. The chosen parameters to be experimented with were 'splitter' which was set to random, 'max_depth' which was set to 4 to avoid a lengthy tree formation, the minimum number of samples to be split at the node (min_sample_split) was set to 4 which was same as min_samples_leaf which denoted the number of samples in a given leaf. The decision tree yielded an R^2 value of 0.92 and a mean squared error of 1.16 which denotes a good fit for the dataset. A pandas data frame was constructed which is stored as a CSV file that consists of the actual test concentration, predicted concentration, and error which is the difference between the actual and predicted concentration. The average of the errors was found to be -0.0123857. The negative value can be justified by the presence of more negative values in the difference column. The standard deviation for the errors in DTR was found to be 1.0770035. The average and standard deviation was computed using Python's numpy library.

7.2 K-NEAREST NEIGHBOURS

Regarding KNNR, the process of developing, training, and evaluating was similar to the process of DTR. The KNNR resulted in an R^2 value of 0.81 and a mean squared error of 1.91. A data frame was also constructed for the results with actual, predicted, and different columns which was saved as a CSV file. The average in errors for KNNR was computed to 0.6747749 with a standard deviation of 2.1099399. Parameters were chosen and experimented with. The number of neighbors was chosen a higher number to give a decent fit and prevent overfitting of the model.

7.3 SUPPORT VECTOR REGRESSION

In SVR, the R^2 resulting from developing, training, and evaluating the model was found to be the lowest R^2 value of 0.67 amongst the models next to KNNR. The mean squared error for the model was found to be 4.91 also the highest mean squared error amongst the models. The selection amongst the 3 inbuilt kernels: radian-based function (rbf), polynomial kernel (poly), and linear kernel (linear) showed a huge difference in the R^2 and mean squared error values. The results of the SVR with different built-in kernels are as follows:

Table:5 Results of SVR with various inbuilt kernels

Kernel Type	R^2	Mean Squared Error	Average of Errors	The standard deviation of errors
Radial Based Function	0.67	4.91	-0.2893805	1.3526230
Linear	0.55	6.76	0.6636520	2.6420797
Polynomial	0.51	7.42	0.4625533	2.5577162

From the table above we can infer that the use of different pre-built kernel types does affect the prediction quality. Also, by going through the predicted data in the CSV file, we can infer that the RBF kernel predicts with more accuracy

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

relative to other kernels. Still, the maximum predicted value using the RBF kernel was 10.96 mg/ml. The same goes for SVR with a linear kernel which had a maximum predicted value of 11.98 mg/ml. In the case of the polynomial kernel, we can observe that even though the R^2 and the mean squared error show a significantly lower accuracy of the model, the prediction range is significantly higher than the rest of the kernels of nearly 15.41 mg/ml.

7.4 RANDOM FOREST REGRESSION

In the case of RFR, the dataset when put through the model resulted in a perfect fit of R^2 value nearing 1 with 0.999971 as an exact value. The mean squared error was found to be nil. The average of errors was calculated to be 0.0009752 with a standard deviation in errors of 0.0206876.

Scatter plots and probability of error plots for each of the models were constructed using the matplotlib library in Python after the training and evaluation of the models. The units of concentration are in mg/ml.

Table:6 Overall analysis of results

MODEL	R^2	Mean Squared Error	Average of Errors	Standard Deviation in Errors
Decision Tree Regressor	0.92	1.16	-0.0123857	1.0770035
SVR (RBF)	0.67	4.91	-0.2893805	1.3526230
SVR (Linear)	0.55	6.76	0.6636520	2.6420797
SVR (Polynomial)	0.51	7.42	0.4625533	2.5577162
K-Nearest Neighbours	0.87	1.91	0.6747749	2.1099399
Random Forest Regressor	0.99	0	0.0009752	0.0206876

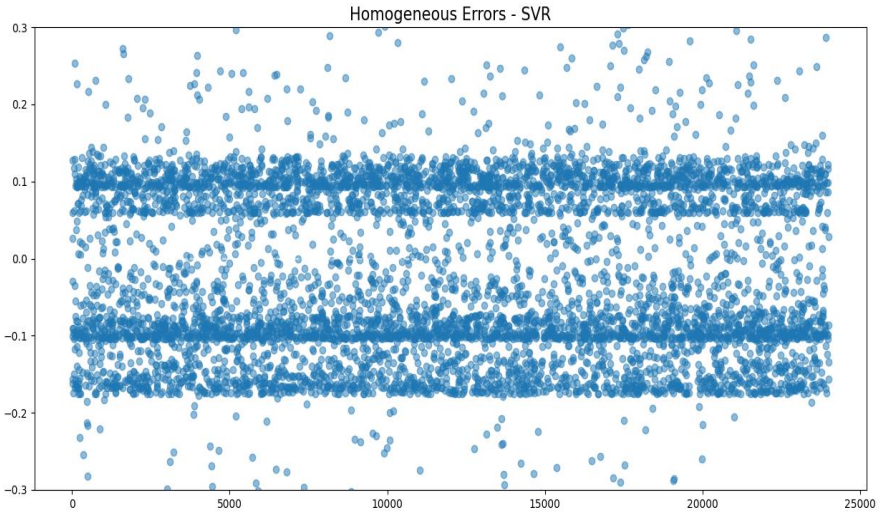

Figure: 3 Support Vector Machine Model Regression Error Plot

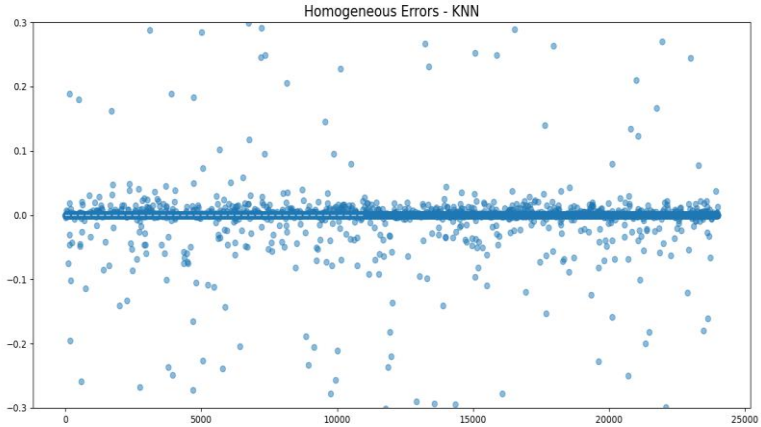

Figure:4 K-Nearest Neighbour Model Homogenous Error Plot

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1549-1559

e-ISSN : 2583-1062
Impact Factor : 7.001

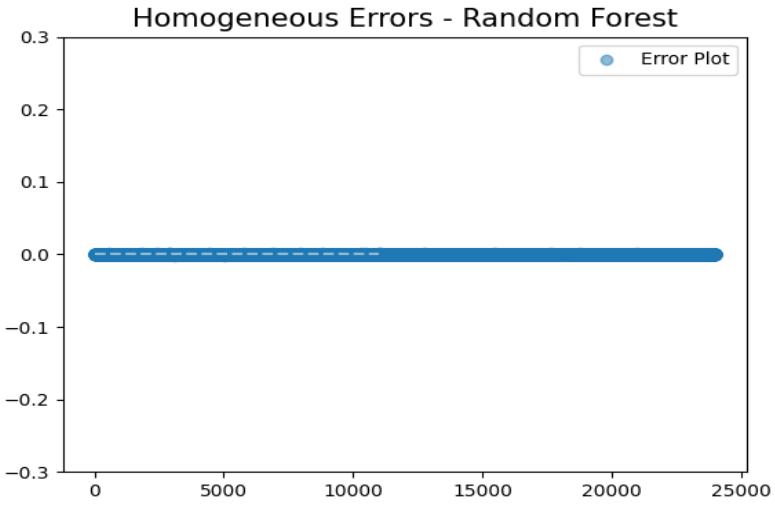

Figure:5 Random Forest Model Homogenous Error Plot

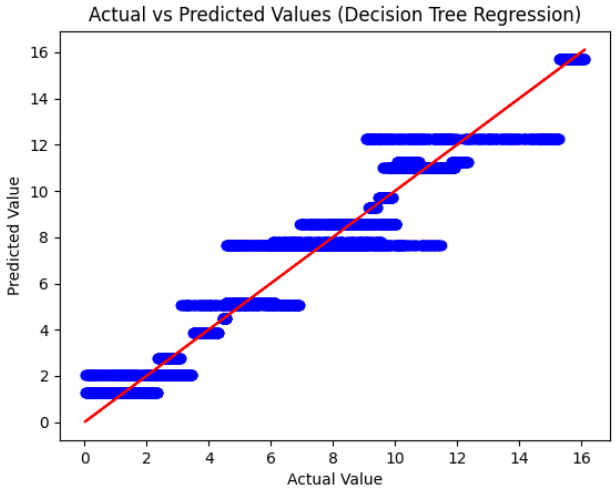

Figure: 6 Decision Tree Actual vs Predicted Values

DATA AVAILABILITY:

The data including the source code, data source, theory behind the implementation, data frames, and final graphs are available for usage in the following GitHub links:

ITEM	GitHub Link
Source Code	Source Code and Dataset CSV
CSV Files	CSV Results GitHub Link
Graphs	Graphs GitHub Link

8. CONCLUSION

The work carried out herein, with the detailed analysis of the production of bioethanol from Psidium guajava leaves using Saccharomyces cerevisiae, exposed one big possibility of combinations between experimental technique and computational methods regarding process optimization studies. The current study has also underlined the possible use of SML models and DOE for yield predictions and optimization processes.

Fermentation of pretreated Psidium guajava leaves was one of the promising sources of bioethanol from an experimental point of view. Optimization of fermentation conditions with respect to pH, temperature, and yeast concentration produced promising bioethanol results and substantiated the use of renewable biomass as an alternative to traditional feedstocks. Step-by-step pretreatment and controlled fermentation provide full accessibility of cellulose to the yeast to act upon.

The dataset generated during the experiments went through several SML models, such as Decision Tree Regressor, Support Vector Regressor, K-Nearest Neighbors Regressor, and Random Forest Regressor. Of these, the Random Forest

Regressor has always yielded the best performance, its R^2 value reaching almost 1, hence proving to be suitable for predictive analysis in this area. Another algorithm that worked really well in the analysis is the Decision Tree Regressor, while models such as the Support Vector Regression showed their limitation, especially at higher mean squared errors and lower ranges of the predictions. The results make it crystal clear that appropriate algorithms should be selected, keeping in mind the nature of the dataset and the requirements of the prediction.

On the other hand, for experimental design, the optimization of their bioethanol production process was investigated by DOE models: Box-Behnken, Central Composite, and Three-Level Designs. Both Box-Behnken and Three-Level Design models did very well in the prediction of bioethanol yield. The R^2 value was marginally higher for the Box-Behnken model, 0.9979, although the Three-Level Design model was also very strong in its predication. These models were further validated by statistical metrics of ANOVA and F-values; the noise-to-signal ratios implied their robustness in tracing the design space.

Comparisons of the DOE models showed that the pH and temperature of the medium are the most relevant parameters to bioethanol yield. This is in good agreement with other previously reported studies on the optimization of fermentation conditions and further confirms the importance of control of process parameters in bioprocess engineering. Furthermore, integration of predictive tools into experimental design will give more insight into variable-variable interactions, which can allow an informed decision on large-scale production.

The present study adds to the fast-increasing growth in the bioethanol optimization domain by presenting a well-integrated approach wherein experimental rigors are poised against certain computational sophistication. The promising results from DOE and SML analyses hint at probably a better way of using Psidium guajava leaves as feedstock in bioethanol production-a feedstock that is sustainable and economically viable. Besides, the scalability of the methods and models presented in this work allows for bright prospects for bioethanol advanced technologies in both academic and industrial sectors.

Other directions for the future might involve methodology extension to other feedstocks, optimization of fermentation conditions, and refinement of machine learning algorithms for better generalization with reduced computational complexity. This could be one way toward developing more efficient and greener systems of bioethanol production for the normalization of renewable energy solutions across the world.

9. REFERENCES

[1] DEMİRBAŞ, A. (2005). Bioethanol from Cellulosic Materials: A Renewable Motor Fuel from Biomass. Energy Sources, 27(4), pp.327–337. doi: https://doi.org/10.1080/00908310390266643

[2] Uma Maheswari, C., Obi Reddy, K., Muzenda, E., Guduri, B.R. and Varada Rajulu, A. (2012). Extraction and characterization of cellulose microfibrils from agricultural residue – Cocos nucifera L. Biomass and Bioenergy, 46, pp.555–563. doi:https://doi.org/10.1016/j.biombioe.2012.06.039

[3] Izmirlioglu, G. and Demirci, A. (2012). Ethanol Production from Waste Potato Mash by Using Saccharomyces Cerevisiae. Applied Sciences, 2(4), pp.738–753. doi:https://doi.org/10.3390/app2040738

[4] T. Tasnim, A. Farasat, The Bioproduction of Ethanol through Isolation of Some Local Bacteria, Medbiotech J. 2018; 2(3): 132-135, DOI:10.22034/mbt.2018.80815

[5] Singh, K., Kumar, R., Chaudhary, V., Vaishali, Sunil, Arya, A.M. and Sharma, S. (2019). Sugarcane bagasse: Foreseeable biomass of bioâ€• products and biofuel: An overview. Journal of Pharmacognosy and Phytochemistry, [online] 8(2), pp.2356–2360. Available at: https://www.phytojournal.com/archives/2019.v8.i2.8022/sugarcane-bagasse-foreseeable-biomass-of-bioaeurproducts-and-biofuel-an-overview [Accessed 13 Nov. 2023]

[6] Savadekar, N.R. and Mhaske, S.T. (2012). Synthesis of nano cellulose fibers and effect on thermoplastics starch-based films. Carbohydrate Polymers, 89(1), pp.146–151. doi:https://doi.org/10.1016/j.carbpol.2012.02.063

[7] Liu, Z., Si, B., Li, J., He, J., Zhang, C., Lu, Y., Zhang, Y. and Xing, X.-H. (2018). Bioprocess engineering for biohythane production from low-grade waste biomass: technical challenges towards scale up. Current Opinion in Biotechnology, 50, pp.25–31. doi:https://doi.org/10.1016/j.copbio.2017.08.014

[8] Lin, Y. and Tanaka, S. (2005). Ethanol fermentation from biomass resources: current state and prospects. Applied Microbiology and Biotechnology, [online] 69(6), pp.627–642. doi:https://doi.org/10.1007/s00253-005-0229-x

[9] www.lifescienceglobal.com. (n.d.). Abstract: From Beverages to Biofuels: The Journeys of Ethanol-Producing Microorganisms - Lifescience Global. [online] Available at: https://www.lifescienceglobal.com/journals/international-journal-of-biotechnology-for-wellness-industries/volume-3-number-3/86-abstract/ijbwi/1164-abstract-from-beverages-to-biofuels-the-journeys-of-ethanol-producing-microorganisms [Accessed 13 Nov. 2023]

[10] Penjumras, P., Rahman, R.B.A., Talib, R.A. and Abdan, K. (2014). Extraction and Characterization of Cellulose from Durian Rind. Agriculture and Agricultural Science Procedia, 2, pp.237–243. doi:https://doi.org/10.1016/j.aaspro.2014.11.034

[11] Goyal, G., Tsai, S.-L., Madan, B., DaSilva, N.A. and Chen, W. (2011). Simultaneous cell growth and ethanol production from cellulose by an engineered yeast consortium displaying a functional mini-cellulosome. Microbial Cell Factories, 10(1), p.89. doi:https://doi.org/10.1186/1475-2859-10-89

[12] Reddy, J.P. and Rhim, J.-W. (2018). Extraction and Characterization of Cellulose Microfibers from Agricultural Wastes of Onion and Garlic. Journal of Natural Fibers, 15(4), pp.465–473. doi:https://doi.org/10.1080/15440478.2014.945227

[13] Reddy, K.O., Uma Maheswari, C., Muzenda, E., Shukla, M. and Rajulu, A.V. (2015). Extraction and Characterization of Cellulose from Pretreated Ficus (Peepal Tree) Leaf Fibers. Journal of Natural Fibers, 13(1), pp.54–64. doi:https://doi.org/10.1080/15440478.2014.984055

[14] ResearchGate. (n.d.). (PDF) Colorimetric Method for the Estimation of Ethanol in Alcoholic-Drinks. [online] Available at: https://www.researchgate.net/publication/228058306_Colorimetric_Method_for_the_Estimation_of_Ethanol_in_Alcoholic-Drinks

[15] Flamholz, A., Noor, E., Bar-Even, A., Liebermeister, W. and Milo, R. (2013). Glycolytic strategy as a tradeoff between energy yield and protein cost. Proceedings of the National Academy of Sciences, 110(24), pp.10039–10044. doi:https://doi.org/10.1073/pnas.1215283110

[16] Sibaly, S. and Jeetah, P. (2017). Production of paper from pineapple leaves. Journal of Environmental Chemical Engineering, 5(6), pp.5978–5986. doi:https://doi.org/10.1016/j.jece.2017.11.026

[17] www.who.int. (n.d.). Laboratory biosafety manual, 3rd edition. [online] Available at: https://www.who.int/publications/i/item/9241546506

[18] Hackeling, G. (n.d.). Mastering Machine Learning with scikit-learn Apply effective learning algorithms to real-world problems using scikit-learn. [online] Available at: https://pdfs.semanticscholar.org/2ff7/555c46302fd599c609522342d8e44a52e164.pdf [Accessed 13 Nov. 2023]

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, [online] 12(85), pp.2825–2830. Available at: https://www.jmlr.org/papers/v12/pedregosa11a.html