CPC251- Machine Learning & Computational Intelligence

PUSAT PENGAJIAN SAINS KOMPUTER
SEMESTER II
2020/2021

PROJECT REPORT

GROUP4_QSAR

Submission Date: 9 July 2021

System demo link: https://youtu.be/7-wak5x_qr0

NO.	NAME	MATRIC NUMBER
1.	OMSYARAN A/L CHANDRAN	148869
2.	THANISH A/L NATARAJAN	149156
3.	PRAVEEN NAIR A/L SANGARA NARAYANAN	149311
4.	MOHAMMED TAYFOUR ABDALLA MOHAMMED	140562

Table of Contents

1.0 Background study

The ability to test and determine the biodegradability of chemicals without turning to expensive tests is both ecologically and economically desirable. Pushing forward in that direction opens many avenues to test and resolve issues without affecting the environment or the economy in a negative way.

In this project, we are tasked with working in that direction. Thus, the aim of the project can be said as follows where we are asked to build 2 models. A machine learning model and a fuzzy logic model both used with the purpose of predicting the target variable of our chosen dataset, the QSAR biodegradation dataset.

In the recent years, many countries round the globe have noticed the general increase of the overall problems and issues faced in our natural environment. They have realized that there is a need to reduce the amount of non-biodegradable materials used to intensify measures for the environment and encourage the recycling of materials.

Many of those countries have assigned agencies and regulations that are responsible for the use of chemical substances and evaluation of their potential impacts on both human health and the environment.

The European Chemicals Agency of the EU (European Union) is notable here due to their efforts in advocating and promoting the use of alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals that are generally at a risk of being harmed with such experiments. Such alternative methods include the biodegradability predictions of chemicals from quantitative structure-activity relationship (QSAR) models.

The QSAR biodegradation dataset was built in the Milano Chemometrics and QSAR Research Group (University of Milan Bicocca).

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Cheminformatics (ECO) project.

The data has been used to develop QSAR (Quantitative Structure Activity Relationships) models for the study of the relationships between chemical structure and biodegradation of molecules. Biodegradation experimental values of 1055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan (NITE). Classification models were developed in order to discriminate ready (356) and not ready (699) biodegradable molecules by the means of three different modelling methods:

1. k Nearest Neighbors
2. Partial Least Squares - Discriminant Analysis
3. Support Vector Machines

Our task is to create new models to be used for prediction and report the accuracy, recall, precision, and F1-score of the machine learning models, and to compare it with the fuzzy logic model and discuss both models that are built in terms of their performance.

2.0 Feature Selection process

As previously stated, we are tasked with building 2 models. A machine learning model, and a fuzzy logic model.

The first model we will discuss is the machine learning model. A machine learning model is a model that has been trained to recognize patterns in the given dataset. A model is trained over a given set of data, provided it has an algorithm that it can use to reason over and learn from said data. The data features that you use to train the machine learning model have a huge impact on the overall performance that can be achieved.

The process of selecting a subset of the relevant features for building the predictive models is known as feature selection.

Given a set of features,

$$\boldsymbol{x} = \{x1, x2, \ldots, xd\}$$

where d is the number of features (dimensions), we want to find k of d dimensions that provide the most information and discard $(p - k)$ dimensions.

There are many advantages which justify why feature selection could be a powerful and viable option to use, these include:

- It reduces overfitting of data by reducing the number of redundant data.
- It improves the overall accuracy of our model.
- The training time is also greatly reduced because there are a fewer number of data points that need to be used after the process.

The goal is to find the subset that contains the least number of dimensions that contributes the most to the overall accuracy. The approaches to feature selection can be divided into 2 main methods.

- Filter Method: It filters out the less relevant features using a statistical measure by assigning scores to each feature.

- Wrapper Method: It uses a predictive model as a black box to evaluate the features and assign scores based on predictive model's accuracy. There are two types of the wrapper method: Forward Sequential Feature Selection (Forward SFS) and Backward Sequential Backward Selection (Backward SFS), both of which have been used in this project.

In forward sequential feature selection, we start with an empty set of features, and then features are added, one by one. The model is trained with one feature at a time and its performance is evaluated accordingly on the training set. The feature that is chosen to be added into the set is the one which has a minimum error.

Backward sequential feature selection is the opposite of that. We initially start with all the features, and then we proceed to remove features one by one. A model is trained with all features save for the removed feature. Given that it is implied that the removed feature is the least relevant, the feature to be removed is the one which yields the minimum error.

3.0 Model Construction and selection

Now we begin to create our models to compare the two with each other in terms of their performance.

First and foremost, we will discuss the machine learning model. Feature selection is performed to determine the relevant features from the QSAR biodegradation dataset, specifically the wrapper method. As previously stated, it has 2 types: Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), both of which have been used in this project. For the machine learning model, we have implemented the following models which are Support Vector Machine (SVC), K-nearest neighbors (KNN), Decision Tree (DT), Perceptron, Logistic Regression and Neural Network.

We will perform the feature selection process for each of the machine learning models and build the model based on the obtained features. Basically, each machine learning model will have its own set of selected features through Forward SFS and Backward SFS. With the selected features, the models will be built, and the performance results will be obtained and recorded. Hence, for each type of machine learning model that is being created, a best model will be picked. When the best model for each type of machine learning model is picked, an overall comparison will be done to determine which machine learning model is the best to be used for this QSAR dataset.

The second model we were tasked to build was the fuzzy logic model. Fuzzy logic, like the name implies, is a form of logic. The truth values of variables in this approach are real values between 0 and 1. The process to construct a fuzzy logic model generally consists of three major steps: fuzzification, inference and defuzzification, all of which were implemented for this project. Further explanation on how we implemented the fuzzy system will be explained in the report as well.

4.0 Result and Discussion

A brief explanation on we are doing is as follows. We will perform Forward SFS and Backward SFS in the sequence of 5,10,15,20,25,30, 35 and 40. Record all the results obtained from it. The results will consist of training and testing set. After that, a best model for that particular machine learning model will be picked. This process will be repeated for all the other type of machine learning model. After picking the best model for each type of machine learning model, we will determine the best overall machine learning model for this QSAR dataset.

Since the problem we are trying to solve is a binary classification problem. Class 'NRB' is represented by 0 while 'RB' is represented by 1. Firstly, we run the feature selection process via wrapper method. Forward and backward sequential feature selection is done. Since the output, as in the features that was selected through this method will differ, we created tables which will be used to store the outputs. This outputs will be used later for the analysis of the best model within the set of machine learning model created.

We set the number of features to be selected in increments of 5. So, the sequence would be 5,10,15,20,25,30,35 and 40. With this sequence, we did feature selection for 'Forward SFS' and 'Backward SFS'. The results are recorded and shown in the tables below. The following tables shows the feature that was selected along with the best parameter that is being selected. The second and third table will consist of images that contains the model's performance result that is created using the best parameter. The second table is for the training set while the third table is for the testing set. This same set of tables will be created for the 'Backward SFS' as well. Hence, for each machine learning model, we will have 2 sets of tables where one is for 'Forward SFS' and the other is for 'Backward SFS'. And within each SFS, will have 3 tables.

4.1 Support Vector Machine (SVM)
4.1.1 SVM Model Forward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'nHM', 'nCp', 'SpMax_A', 'nN', 'nArCOOR'	{'C': 2, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.260551408086333 4
10	'nHM', 'nCp', 'F03[C-N]', 'LOC', 'SpMax_A', 'SpMax_B(m)', 'Psi_i_A','nN', 'SM6_B(m)','nArCOOR'	{{'C': 2, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.333284056471284 3
15	'nHM', 'F04[C-N]', 'nCp', 'F03[C-N]', 'LOC', 'nN-N', 'nCIR', 'SpMax_A','Psi_i_1d', 'nCrt','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)','nArCOOR'	{'C': 3, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.375711434695838 95
20	'nHM', 'F04[C-N]', 'nCp', 'F03[C-N]', 'LOC', 'F03[C-O]', 'nN-N','nCRX3', 'nCIR', 'B03[C-Cl]', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]','nCrt', 'F02[C-N]', 'SpMax_B(m)','Psi_i_A','nN','SM6_B(m)','nArCOOR'	{'C': 1.5, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.375711434695838 95
25	'nHM', 'F01[N-N]', 'F04[C-N]', 'nCp', 'F03[C-N]', 'LOC', 'F03[C-O]', 'nN-N', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]','N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'TI2_L', 'nCrt','F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	{'C': 2, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.369650380663759 66
30	'J_Dz(e)','nHM','F01[N-N]','F04[C-N]','NssssC', 'C%', 'nCp','F03[C-N]', 'LOC', 'F03[C-O]', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)','nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d','B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'F02[C-N]','SpMax_B(m)','Psi_i_A','nN','SM6_B(m)','nArCOOR'	{'C': 2, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.454505137112868 94
35	'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'C%', 'nCp','F03[C-N]', 'HyWi_B(m)', 'LOC', 'F03[C-O]', 'Me', 'Mi', 'nN-N','nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]','N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt','F02[C-N]', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)','nArCOOR', 'nX'	{'C': 3, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.448444083080789 65

40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-','C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)','nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d','B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'F02[C-N]', 'nHDon', 'SpMax_B(m)','Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'C': 2, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	0.460566191144948 1

TRAINING SET RESULT FOR SVM MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.15582655826558264 Mean squared error : 0.15582655826558264 r2 score : 0.3029787863108879 The max error value : 1 Accuracy score: 0.8441734417344173 [[452 37] [78 171]] precision recall f1-score support 0 0.85 0.92 0.89 489 1 0.82 0.69 0.75 249 accuracy 0.84 738 macro avg 0.84 0.81 0.82 738 weighted avg 0.84 0.84 0.84 738
10	Mean absolute error : 0.12330623306233063 Mean squared error : 0.12330623306233063 r2 score : 0.44844408308078954 The max error value : 1 Accuracy score: 0.8766937669376694 [[464 25] [66 183]] precision recall f1-score support 0 0.88 0.95 0.91 489 1 0.88 0.73 0.80 249 accuracy 0.88 738 macro avg 0.88 0.84 0.86 738 weighted avg 0.88 0.88 0.87 738
15	Mean absolute error : 0.12059620596205962 Mean squared error : 0.12059620596205962 r2 score : 0.460566191144948 The max error value : 1 Accuracy score: 0.8794037940379403 [[463 26] [63 186]] precision recall f1-score support 0 0.88 0.95 0.91 489 1 0.88 0.75 0.81 249 accuracy 0.88 738 macro avg 0.88 0.85 0.86 738 weighted avg 0.88 0.88 0.88 738

20	Mean absolute error : 0.12059620596205962 Mean squared error : 0.12059620596205962 r2 score : 0.460566191144948 The max error value : 1 Accuracy score: 0.8794037940379403 [[459 30] [59 190]] precision recall f1-score support 0 0.89 0.94 0.91 489 1 0.86 0.76 0.81 249 accuracy 0.88 738 macro avg 0.87 0.85 0.86 738 weighted avg 0.88 0.88 0.88 738
25	Mean absolute error : 0.11246612466124661 Mean squared error : 0.11246612466124661 r2 score : 0.49693251533742344 The max error value : 1 Accuracy score: 0.8875338753387534 [[458 31] [52 197]] precision recall f1-score support 0 0.90 0.94 0.92 489 1 0.86 0.79 0.83 249 accuracy 0.89 738 macro avg 0.88 0.86 0.87 738 weighted avg 0.89 0.89 0.89 738
30	Mean absolute error : 0.1016260162601626 Mean squared error : 0.1016260162601626 r2 score : 0.5454209475940572 The max error value : 1 Accuracy score: 0.8983739837398373 [[461 28] [47 202]] precision recall f1-score support 0 0.91 0.94 0.92 489 1 0.88 0.81 0.84 249 accuracy 0.90 738 macro avg 0.89 0.88 0.88 738 weighted avg 0.90 0.90 0.90 738

35	Mean absolute error : 0.08401084010840108 Mean squared error : 0.08401084010840108 r2 score : 0.6242146500110874 The max error value : 1 Accuracy score: 0.9159891598915989 [[468 21] [41 208]] precision recall f1-score support 0 0.92 0.96 0.94 489 1 0.91 0.84 0.87 249 accuracy 0.92 738 macro avg 0.91 0.90 0.90 738 weighted avg 0.92 0.92 0.92 738
40	Mean absolute error : 0.09620596205962059 Mean squared error : 0.09620596205962059 r2 score : 0.5696651637223742 The max error value : 1 Accuracy score: 0.9037940379403794 [[461 28] [43 206]] precision recall f1-score support 0 0.91 0.94 0.93 489 1 0.88 0.83 0.85 249 accuracy 0.90 738 macro avg 0.90 0.89 0.89 738 weighted avg 0.90 0.90 0.90 738

TEST SET RESULT FOR SVM MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.1640378548895899 Mean squared error : 0.1640378548895899 r2 score : 0.2663996439697375 The max error value : 1 0.8359621451104101 [[187 23] [29 78]] precision recall f1-score support 0 0.87 0.89 0.88 210 1 0.77 0.73 0.75 107 accuracy 0.84 317 macro avg 0.82 0.81 0.81 317 weighted avg 0.83 0.84 0.83 317

10	Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 0.8580441640378549 [[194 16] [29 78]] precision recall f1-score support 0 0.87 0.92 0.90 210 1 0.83 0.73 0.78 107 accuracy 0.86 317 macro avg 0.85 0.83 0.84 317 weighted avg 0.86 0.86 0.86 317
15	Mean absolute error : 0.138801261829653 Mean squared error : 0.138801261829653 r2 score : 0.37926123720516247 The max error value : 1 0.861198738170347 [[196 14] [30 77]] precision recall f1-score support 0 0.87 0.93 0.90 210 1 0.85 0.72 0.78 107 accuracy 0.86 317 macro avg 0.86 0.83 0.84 317 weighted avg 0.86 0.86 0.86 317
20	Mean absolute error : 0.15457413249211358 Mean squared error : 0.15457413249211358 r2 score : 0.3087227414330218 The max error value : 1 0.8454258675078864 [[191 19] [30 77]] precision recall f1-score support 0 0.86 0.91 0.89 210 1 0.80 0.72 0.76 107 accuracy 0.85 317 macro avg 0.83 0.81 0.82 317 weighted avg 0.84 0.85 0.84 317
25	Mean absolute error : 0.15457413249211358 Mean squared error : 0.15457413249211358 r2 score : 0.3087227414330218 The max error value : 1 0.8454258675078864 [[190 20] [29 78]] precision recall f1-score support 0 0.87 0.90 0.89 210 1 0.80 0.73 0.76 107 accuracy 0.85 317 macro avg 0.83 0.82 0.82 317 weighted avg 0.84 0.85 0.84 317

30	Mean absolute error : 0.12618296529968454 Mean squared error : 0.12618296529968454 r2 score : 0.4356920338228749 The max error value : 1 0.8738170347003155 [[194 16] [24 83]] precision recall f1-score support 0 0.89 0.92 0.91 210 1 0.84 0.78 0.81 107 accuracy 0.87 317 macro avg 0.86 0.85 0.86 317 weighted avg 0.87 0.87 0.87 317
35	Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 0.886435331230284 [[196 14] [22 85]] precision recall f1-score support 0 0.90 0.93 0.92 210 1 0.86 0.79 0.83 107 accuracy 0.89 317 macro avg 0.88 0.86 0.87 317 weighted avg 0.89 0.89 0.89 317
40	Mean absolute error : 0.12302839116719243 Mean squared error : 0.12302839116719243 r2 score : 0.44979973297730313 The max error value : 1 0.8769716088328076 [[193 17] [22 85]] precision recall f1-score support 0 0.90 0.92 0.91 210 1 0.83 0.79 0.81 107 accuracy 0.88 317 macro avg 0.87 0.86 0.86 317 weighted avg 0.88 0.88 0.88 317

Now we have completed the ‘Forward SFS’ for the SVM model. The models are created, and it is evaluated as well. The results are stored. By looking at the table above, the model with 15, 30, 35 and 40 features selected has a good accuracy value which ranges from 0.86 to 0.88. It looks promising even though it has misclassified some of the data. Moving on to ‘Backward SFS’, the table for its outputs are as follows.

4.1.2 SVM Model Backward SFS

Number of features selected	Features that were selected	Best parameters
5	'SpMax_L', 'nHM', 'SdO', 'nCrt', 'nN'	{'C': 2.5, 'degree': 2, 'gamma': ' scale', 'kernel': 'rbf'}
10	'SpMax_L', 'nHM', 'nCp', 'HyWi_B(m)', 'LOC', 'nArNO2', 'SdO', 'nCr t','Psi_i_A', 'nN'	{'C': 1.5, 'degree': 2, 'gamma': ' scale', 'kernel': 'rbf'}
15	'SpMax_L', 'nHM', 'C%', 'nCp', 'SdssC', 'HyWi_B(m)', 'LOC', 'nArNO 2','B03[C-Cl]', 'Psi_i_1d', 'SdO', 'nCrt', 'Psi_i_A', 'nN', 'nArCOOR'	{'C': 2, 'degree': 2, 'gamma': 'sc ale', 'kernel': 'linear'}
20	'SpMax_L', 'nHM', 'F01[N-N]', 'F04[C-N]', 'C%', 'nCp', 'nO', 'SdssC',' HyWi_B(m)', 'LOC', 'nN-N', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'Psi_i_1d', 'SdO', 'nCrt', 'Psi_i_A', 'nN', 'nArCOOR'	{'C': 1.5, 'degree': 2, 'gamma': ' scale', 'kernel': 'rbf'}
25	'SpMax_L', 'nHM', 'F01[N-N]', 'F04[C-N]', 'C%', 'nCp', 'nO', 'F03[C-N]','SdssC', 'HyWi_B(m)', 'LOC', 'F03[C-O]', 'Mi', 'nN-N', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'Psi_i_1d', 'SdO', 'nCrt', 'nHDon', 'Psi_i_A','nN', 'nArCOOR', 'nX'	{'C': 2, 'degree': 2, 'gamma': 'sc ale', 'kernel': 'rbf'}
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'F03[C-O] ','Mi', 'nN-N', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A ', 'Psi_i_1d', 'SdO', 'nCrt', 'nHDon', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'C': 1.5, 'degree': 2, 'gamma': ' scale', 'kernel': 'rbf'}
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCIR', 'B01[C-Br]','B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'SdO', 'TI2_L', 'nCrt', 'nHDon', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'C': 2.5, 'degree': 2, 'gamma': ' scale', 'kernel': 'rbf'}
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)','nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d','B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'C': 3.5, 'degree': 2, 'gamma': ' scale', 'kernel': 'rbf'}

TRAINING SET RESULT FOR SVM MODEL BACKWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.13414634146341464 Mean squared error : 0.13414634146341464 r2 score : 0.39995565082415563 The max error value : 1 Accuracy score: 0.8658536585365854 [[458 31] [68 181]] precision recall f1-score support 0 0.87 0.94 0.90 489 1 0.85 0.73 0.79 249 accuracy 0.87 738 macro avg 0.86 0.83 0.84 738 weighted avg 0.87 0.87 0.86 738
10	Mean absolute error : 0.1043360433604336 Mean squared error : 0.1043360433604336 r2 score : 0.5332988395298989 The max error value : 1 Accuracy score: 0.8956639566395664 [[460 29] [48 201]] precision recall f1-score support 0 0.91 0.94 0.92 489 1 0.87 0.81 0.84 249 accuracy 0.90 738 macro avg 0.89 0.87 0.88 738 weighted avg 0.89 0.90 0.89 738
15	Mean absolute error : 0.11653116531165311 Mean squared error : 0.11653116531165311 r2 score : 0.4787493532411857 The max error value : 1 Accuracy score: 0.8834688346883469 [[453 36] [50 199]] precision recall f1-score support 0 0.90 0.93 0.91 489 1 0.85 0.80 0.82 249 accuracy 0.88 738 macro avg 0.87 0.86 0.87 738 weighted avg 0.88 0.88 0.88 738

20	Mean absolute error : 0.0921409214092141 Mean squared error : 0.0921409214092141 r2 score : 0.587848325818612 The max error value : 1 Accuracy score: 0.907859078590786 [[463 26] [42 207]] precision recall f1-score support 0 0.92 0.95 0.93 489 1 0.89 0.83 0.86 249 accuracy 0.91 738 macro avg 0.90 0.89 0.90 738 weighted avg 0.91 0.91 0.91 738
25	Mean absolute error : 0.08943089430894309 Mean squared error : 0.08943089430894309 r2 score : 0.5999704338827705 The max error value : 1 Accuracy score: 0.9105691056910569 [[463 26] [40 209]] precision recall f1-score support 0 0.92 0.95 0.93 489 1 0.89 0.84 0.86 249 accuracy 0.91 738 macro avg 0.90 0.89 0.90 738 weighted avg 0.91 0.91 0.91 738
30	Mean absolute error : 0.08807588075880758 Mean squared error : 0.08807588075880758 r2 score : 0.6060314879148496 The max error value : 1 Accuracy score: 0.9119241192411924 [[465 24] [41 208]] precision recall f1-score support 0 0.92 0.95 0.93 489 1 0.90 0.84 0.86 249 accuracy 0.91 738 macro avg 0.91 0.89 0.90 738 weighted avg 0.91 0.91 0.91 738

35	Mean absolute error : 0.08943089430894309 Mean squared error : 0.08943089430894309 r2 score : 0.5999704338827705 The max error value : 1 Accuracy score: 0.9105691056910569 [[464 25] [41 208]] precision recall f1-score support 0 0.92 0.95 0.93 489 1 0.89 0.84 0.86 249 accuracy 0.91 738 macro avg 0.91 0.89 0.90 738 weighted avg 0.91 0.91 0.91 738
40	Mean absolute error : 0.08265582655826559 Mean squared error : 0.08265582655826559 r2 score : 0.6302757040431666 The max error value : 1 Accuracy score: 0.9173441734417345 [[466 23] [38 211]] precision recall f1-score support 0 0.92 0.95 0.94 489 1 0.90 0.85 0.87 249 accuracy 0.92 738 macro avg 0.91 0.90 0.91 738 weighted avg 0.92 0.92 0.92 738

TEST SET RESULT FOR SVM MODEL BACKWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.14511041009463724 Mean squared error : 0.14511041009463724 r2 score : 0.35104583889630625 The max error value : 1 0.8548895899053628 [[189 21] [25 82]] precision recall f1-score support 0 0.88 0.90 0.89 210 1 0.80 0.77 0.78 107 accuracy 0.85 317 macro avg 0.84 0.83 0.84 317 weighted avg 0.85 0.85 0.85 317
10	Mean absolute error : 0.13249211356466878 Mean squared error : 0.13249211356466878 r2 score : 0.4074766355140187 The max error value : 1 0.8675078864353313 [[190 20] [22 85]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.81 0.79 0.80 107 accuracy 0.87 317 macro avg 0.85 0.85 0.85 317 weighted avg 0.87 0.87 0.87 317
15	Mean absolute error : 0.13564668769716087 Mean squared error : 0.13564668769716087 r2 score : 0.3933689363595906 The max error value : 1 0.8643533123028391 [[185 25] [18 89]] precision recall f1-score support 0 0.91 0.88 0.90 210 1 0.78 0.83 0.81 107 accuracy 0.86 317 macro avg 0.85 0.86 0.85 317 weighted avg 0.87 0.86 0.87 317

20	Mean absolute error : 0.13249211356466878 Mean squared error : 0.13249211356466878 r2 score : 0.4074766355140187 The max error value : 1 0.8675078864353313 [[189 21] [21 86]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.80 0.80 0.80 107 accuracy 0.87 317 macro avg 0.85 0.85 0.85 317 weighted avg 0.87 0.87 0.87 317
25	Mean absolute error : 0.12302839116719243 Mean squared error : 0.12302839116719243 r2 score : 0.44979973297730313 The max error value : 1 0.8769716088328076 [[193 17] [22 85]] precision recall f1-score support 0 0.90 0.92 0.91 210 1 0.83 0.79 0.81 107 accuracy 0.88 317 macro avg 0.87 0.86 0.86 317 weighted avg 0.88 0.88 0.88 317
30	Mean absolute error : 0.13249211356466878 Mean squared error : 0.13249211356466878 r2 score : 0.4074766355140187 The max error value : 1 0.8675078864353313 [[191 19] [23 84]] precision recall f1-score support 0 0.89 0.91 0.90 210 1 0.82 0.79 0.80 107 accuracy 0.87 317 macro avg 0.85 0.85 0.85 317 weighted avg 0.87 0.87 0.87 317
35	Mean absolute error : 0.12302839116719243 Mean squared error : 0.12302839116719243 r2 score : 0.44979973297730313 The max error value : 1 0.8769716088328076 [[193 17] [22 85]] precision recall f1-score support 0 0.90 0.92 0.91 210 1 0.83 0.79 0.81 107 accuracy 0.88 317 macro avg 0.87 0.86 0.86 317 weighted avg 0.88 0.88 0.88 317

40	Mean absolute error : 0.1167192429022082 Mean squared error : 0.1167192429022082 r2 score : 0.47801513128615936 The max error value : 1 0.8832807570977917 [[194 16] [21 86]] precision recall f1-score support 0 0.90 0.92 0.91 210 1 0.84 0.80 0.82 107 accuracy 0.88 317 macro avg 0.87 0.86 0.87 317 weighted avg 0.88 0.88 0.88 317

4.1.3 Picking the best SVM (SVC) Model.

Now, the 'Backward SFS' is done as well. If we take a closer look at the outputs of the 'Backward SFS' we can say that the model with 30 and 35 does look good where its accuracy is 0.86 and 0.87, respectively. However, its f1-score is almost similar. Hence, the best model for SVM is the model with 20 features selected via 'Backward SFS'. It is because that model has an accuracy score of 0.86 and a f1-score of 0.87 where it means there are some misclassified data which is around 40 to be exact. This means that the models are not overfitted. Hence, this model can be used to predict data correctly. So, this model with the following parameter is picked : {'C': 1.5, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}. The result for that best model is shown below.

```
Mean absolute error : 0.13249211356466878
Mean squared error : 0.13249211356466878
r2 score : 0.4074766355140187
The max error value : 1

0.8675078864353313
[[189  21]
 [ 21  86]]
              precision    recall  f1-score   support

           0       0.90      0.90      0.90       210
           1       0.80      0.80      0.80       107

    accuracy                           0.87       317
   macro avg       0.85      0.85      0.85       317
weighted avg       0.87      0.87      0.87       317
```

4.2 K-Nearest Neighbor (KNN) model

4.2.1 KNN Forward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'F04[C-N]', 'F03[C-O]', 'SdO', 'nCrt', 'SpMax_B(m)'	{'metric': 'euclidean', 'n_neighbors': 5, 'weights': 'uniform'}	0.8428184281842818
10	'F01[N-N]', 'F04[C-N]', 'NssssC', 'F03[C-O]', 'nN-N', 'nArNO2', 'SdO', 'nCrt', 'SpMax_B(m)', 'nArCOOR'	{'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'uniform'}	0.8631436314363143
15	'F01[N-N]', 'F04[C-N]', 'NssssC', 'F03[C-O]', 'nN-N', 'nArNO2', 'nCRX3','B01[C-Br]', 'B03[C-Cl]', 'N-073', 'B04[C-Br]', 'SdO', 'nCrt', 'SpMax_B(m)', 'nArCOOR'	{'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'uniform'}	0.8550135501355013
20	'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nO', 'F03[C-N]', 'F03[C-O]', 'nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'B04[C-Br]', 'SdO', 'nCrt', 'SpMax_B(m)', 'nArCOOR', 'nX'	{'metric': 'euclidean', 'n_neighbors': 11, 'weights': 'distance'}	0.8509485094850948
25	'SpMax_L', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nO', 'F03[C-N]', 'SdssC', 'F03[C-O]', 'nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'nCrt', 'SpMax_B(m)', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'metric': 'manhattan', 'n_neighbors': 5, 'weights': 'distance'}	0.8766937669376693
30	'SpMax_L', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nO', 'F03[C-N]', 'SdssC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]','SdO', 'TI2_L', 'nCrt', 'C-026', 'SpMax_B(m)', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'metric': 'euclidean', 'n_neighbors': 11, 'weights': 'distance'}	0.8644986449864499
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'distance'}	0.8644986449864499
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'uniform'}	0.8577235772357724

TRAINING SET RESULT FOR KNN MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Training set Mean absolute error : 0.11653116531165311 Mean squared error : 0.11653116531165311 r2 score : 0.4787493532411857 The max error value : 1 accuracy score: 0.8834688346883469 [[448 41] [45 204]] precision recall f1-score support 0 0.91 0.92 0.91 489 1 0.83 0.82 0.83 249 accuracy 0.88 738 macro avg 0.87 0.87 0.87 738 weighted avg 0.88 0.88 0.88 738
10	Training set Mean absolute error : 0.08536585365853659 Mean squared error : 0.08536585365853659 r2 score : 0.6181535959790081 The max error value : 1 accuracy score: 0.9146341463414634 [[465 24] [39 210]] precision recall f1-score support 0 0.92 0.95 0.94 489 1 0.90 0.84 0.87 249 accuracy 0.91 738 macro avg 0.91 0.90 0.90 738 weighted avg 0.91 0.91 0.91 738

15	Training set Mean absolute error : 0.08943089430894309 Mean squared error : 0.08943089430894309 r2 score : 0.5999704338827705 The max error value : 1 accuracy score: 0.9105691056910569 [[463 26] [40 209]] precision recall f1-score support 0 0.92 0.95 0.93 489 1 0.89 0.84 0.86 249 accuracy 0.91 738 macro avg 0.90 0.89 0.90 738 weighted avg 0.91 0.91 0.91 738	
20	Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738	

25		Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738
30		Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738

35		Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738
40		Training set Mean absolute error : 0.11517615176151762 Mean squared error : 0.11517615176151762 r2 score : 0.48481040727326496 The max error value : 1 accuracy score: 0.8848238482384824 [[444 45] [40 209]] precision recall f1-score support 0 0.92 0.91 0.91 489 1 0.82 0.84 0.83 249 accuracy 0.88 738 macro avg 0.87 0.87 0.87 738 weighted avg 0.89 0.88 0.89 738

TEST SET RESULT FOR KNN MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.1640378548895899 Mean squared error : 0.1640378548895899 r2 score : 0.2663996439697375 The max error value : 1 accuracy score: 0.8359621451104101 confusion matrix: [[185 25] [27 80]] precision recall f1-score support 0 0.87 0.88 0.88 210 1 0.76 0.75 0.75 107 accuracy 0.84 317 macro avg 0.82 0.81 0.82 317 weighted avg 0.84 0.84 0.84 317
10	Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.3651535380507343 6 The max error value : 1 accuracy score: 0.8580441640378549 confusion matrix: [[189 21] [24 83]] precision recall f1-score support 0 0.89 0.90 0.89 210 1 0.80 0.78 0.79 107 accuracy 0.86 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317
15	Mean absolute error : 0.13564668769716087 Mean squared error : 0.13564668769716087 r2 score : 0.3933689363595906 The max error value : 1 accuracy score: 0.8643533123028391 confusion matrix: [[192 18] [25 82]] precision recall f1-score support 0 0.88 0.91 0.90 210 1 0.82 0.77 0.79 107 accuracy 0.86 317 macro avg 0.85 0.84 0.85 317 weighted avg 0.86 0.86 0.86 317
20	Mean absolute error : 0.15141955835962145 Mean squared error : 0.15141955835962145 r2 score : 0.3228304405874499 The max error value : 1 accuracy score: 0.8485804416403786 confusion matrix: [[185 25] [23 84]] precision recall f1-score support 0 0.89 0.88 0.89 210 1 0.77 0.79 0.78 107 accuracy 0.85 317 macro avg 0.83 0.83 0.83 317 weighted avg 0.85 0.85 0.85 317

25		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion matrix: [[187 23] [22 85]] precision recall f1-score support 0 0.89 0.89 0.89 210 1 0.79 0.79 0.79 107 accuracy 0.86 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317
30		Mean absolute error : 0.15141955835962145 Mean squared error : 0.15141955835962145 r2 score : 0.3228304405874499 The max error value : 1 accuracy score: 0.8485804416403786 confusion matrix: [[183 27] [21 86]] precision recall f1-score support 0 0.90 0.87 0.88 210 1 0.76 0.80 0.78 107 accuracy 0.85 317 macro avg 0.83 0.84 0.83 317 weighted avg 0.85 0.85 0.85 317
35		Mean absolute error : 0.14826498422712933 Mean squared error : 0.14826498422712933 r2 score : 0.33693813974187803 The max error value : 1 accuracy score: 0.8517350157728707 confusion matrix: [[184 26] [21 86]] precision recall f1-score support 0 0.90 0.88 0.89 210 1 0.77 0.80 0.79 107 accuracy 0.85 317 macro avg 0.83 0.84 0.84 317 weighted avg 0.85 0.85 0.85 317
40		Mean absolute error : 0.14511041009463724 Mean squared error : 0.14511041009463724 r2 score : 0.35104583889630625 The max error value : 1 accuracy score: 0.8548895899053628 confusion matrix: [[183 27] [19 88]] precision recall f1-score support 0 0.91 0.87 0.89 210 1 0.77 0.82 0.79 107 accuracy 0.85 317 macro avg 0.84 0.85 0.84 317 weighted avg 0.86 0.85 0.86 317

The KNN Forward SFS has now been completed for all number of features. The accuracy number for all of them float around 2 values. 0.85 and 0.86. Currently, the most promising model is the one with the number of features being 15. The accuracy score for it is 0.85, however its f1-score is 0.90, which is the highest. We will now proceed to complete the KNN Backward SFS

4.2.2 KNN Backward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'SpMax_L', 'nHM', 'Me', 'SpMax_A', 'nN'	{'metric': 'euclidean', 'n_neighbors': 19, 'weights': 'distance'}	0.857723577235772 4
10	'SpMax_L', 'nHM', 'nCp', 'F03[C-O]', 'Me', 'SpPosA_B(p)', 'SpMax_A', 'C-026', 'nN', 'nArCOOR'	{'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'distance'}	0.860433604336043 3
15	'SpMax_L', 'nHM', 'nCp', 'SdssC', 'SM6_L', 'F03[C-O]', 'Me','SpPosA_B(p)', 'SpMax_A', 'C-026', 'SpMax_B(m)', 'Psi_i_A', 'nN','SM6_B(m)','nArCOOR'	{'metric': 'manhattan', 'n_neighbors': 5, 'weights': 'uniform'}	0.856368563685637
20	'SpMax_L', 'J_Dz(e)', 'nHM', 'nCp', 'SdssC', 'SM6_L', 'F03[C-O]', 'Me', 'nArNO2', 'SpPosA_B(p)', 'SpMax_A', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	{'metric': 'euclidean', 'n_neighbors': 11, 'weights': 'uniform'}	0.860433604336043 4
25	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'NssssC', 'nCp', 'nO','F03[C-N]', 'SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nArNO2', 'SpPosA_B(p)', 'SpMax_A', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	{'metric': 'euclidean', 'n_neighbors': 19, 'weights': 'uniform'}	0.857723577235772 4
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'C%','nCp', 'nO', 'F03[C-N]', 'SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'B03[C-Cl]', 'SpMax_A', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)','nArCOOR', 'nX'	{'metric': 'manhattan', 'n_neighbors': 19, 'weights': 'distance'}	0.860433604336043 3
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'metric': 'euclidean', 'n_neighbors': 11, 'weights': 'distance'}	0.856368563685637
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-	{'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'uniform'}	0.857723577235772 4

	Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'		

TRAINING SET RESULT FOR KNN MODEL BACKWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Training set Mean absolute error : 0.008130081300813009 Mean squared error : 0.008130081300813009 r2 score : 0.9636336758075246 The max error value : 1 accuracy score: 0.991869918699187 [[489 0] [6 243]] precision recall f1-score support 0 0.99 1.00 0.99 489 1 1.00 0.98 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738
10	Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738

15		Training set Mean absolute error : 0.0989159891598916 Mean squared error : 0.0989159891598916 r2 score : 0.5575430556582157 The max error value : 1 accuracy score: 0.9010840108401084 [[454 35] [38 211]] precision recall f1-score support 0 0.92 0.93 0.93 489 1 0.86 0.85 0.85 249 accuracy 0.90 738 macro avg 0.89 0.89 0.89 738 weighted avg 0.90 0.90 0.90 738	
20		Training set Mean absolute error : 0.12466124661246612 Mean squared error : 0.12466124661246612 r2 score : 0.4423830290487103 The max error value : 1 accuracy score: 0.8753387533875339 [[446 43] [49 200]] precision recall f1-score support 0 0.90 0.91 0.91 489 1 0.82 0.80 0.81 249 accuracy 0.88 738 macro avg 0.86 0.86 0.86 738 weighted avg 0.87 0.88 0.87 738	
25		Training set Mean absolute error : 0.12737127371273713 Mean squared error : 0.12737127371273713 r2 score : 0.4302609209845518 The max error value : 1 accuracy score: 0.8726287262872628 [[435 54] [40 209]] precision recall f1-score support 0 0.92 0.89 0.90 489 1 0.79 0.84 0.82 249 accuracy 0.87 738 macro avg 0.86 0.86 0.86 738 weighted avg 0.87 0.87 0.87 738	

30		Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738	
35		Training set Mean absolute error : 0.0 Mean squared error : 0.0 r2 score : 1.0 The max error value : 0 accuracy score: 1.0 [[489 0] [0 249]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738	
40		Training set Mean absolute error : 0.11517615176151762 Mean squared error : 0.11517615176151762 r2 score : 0.48481040727326496 The max error value : 1 accuracy score: 0.8848238482384824 [[444 45] [40 209]] precision recall f1-score support 0 0.92 0.91 0.91 489 1 0.82 0.84 0.83 249 accuracy 0.88 738 macro avg 0.87 0.87 0.87 738 weighted avg 0.89 0.88 0.89 738	

TEST SET RESULT FOR KNN MODEL BACKWARD SFS

Number of features selected	The model output when the best parameter is fitted
5	Mean absolute error : 0.167192429022082 Mean squared error : 0.167192429022082 r2 score : 0.25229194481530937 The max error value : 1 accuracy score: 0.832807570977918 confusion matrix: [[186 24] [29 78]] precision recall f1-score support 0 0.87 0.89 0.88 210 1 0.76 0.73 0.75 107 accuracy 0.83 317 macro avg 0.81 0.81 0.81 317 weighted avg 0.83 0.83 0.83 317
10	Mean absolute error : 0.14826498422712933 Mean squared error : 0.14826498422712933 r2 score : 0.33693813974187803 The max error value : 1 accuracy score: 0.8517350157728707 confusion matrix: [[186 24] [23 84]] precision recall f1-score support 0 0.89 0.89 0.89 210 1 0.78 0.79 0.78 107 accuracy 0.85 317 macro avg 0.83 0.84 0.83 317 weighted avg 0.85 0.85 0.85 317
15	Mean absolute error : 0.138801261829653 Mean squared error : 0.138801261829653 r2 score : 0.37926123720516247 The max error value : 1 accuracy score: 0.861198738170347 confusion matrix: [[189 21] [23 84]] precision recall f1-score support 0 0.89 0.90 0.90 210 1 0.80 0.79 0.79 107 accuracy 0.86 317 macro avg 0.85 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317

20		Mean absolute error : 0.15141955835962145 Mean squared error : 0.15141955835962145 r2 score : 0.3228304405874499 The max error value : 1 accuracy score: 0.8485804416403786 confusion matrix: [[188 22] [26 81]] precision recall f1-score support 0 0.88 0.90 0.89 210 1 0.79 0.76 0.77 107 accuracy 0.85 317 macro avg 0.83 0.83 0.83 317 weighted avg 0.85 0.85 0.85 317	
25		Mean absolute error : 0.15457413249211358 Mean squared error : 0.15457413249211358 r2 score : 0.3087227414330218 The max error value : 1 accuracy score: 0.8454258675078864 confusion matrix: [[181 29] [20 87]] precision recall f1-score support 0 0.90 0.86 0.88 210 1 0.75 0.81 0.78 107 accuracy 0.85 317 macro avg 0.83 0.84 0.83 317 weighted avg 0.85 0.85 0.85 317	
30		Mean absolute error : 0.14826498422712933 Mean squared error : 0.14826498422712933 r2 score : 0.33693813974187803 The max error value : 1 accuracy score: 0.8517350157728707 confusion matrix: [[185 25] [22 85]] precision recall f1-score support 0 0.89 0.88 0.89 210 1 0.77 0.79 0.78 107 accuracy 0.85 317 macro avg 0.83 0.84 0.84 317 weighted avg 0.85 0.85 0.85 317	
35		Mean absolute error : 0.13564668769716087 Mean squared error : 0.13564668769716087 r2 score : 0.3933689363595906 The max error value : 1 accuracy score: 0.8643533123028391 confusion matrix: [[185 25] [18 89]] precision recall f1-score support 0 0.91 0.88 0.90 210 1 0.78 0.83 0.81 107 accuracy 0.86 317 macro avg 0.85 0.86 0.85 317 weighted avg 0.87 0.86 0.87 317	

40		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.3651535380507343 6 The max error value : 1 accuracy score: 0.8580441640378549 confusion matrix: [[185 25] [20 87]] precision recall f1-score support 0 0.90 0.88 0.89 210 1 0.78 0.81 0.79 107 accuracy 0.86 317 macro avg 0.84 0.85 0.84 317 weighted avg 0.86 0.86 0.86 317	

Both the KNN forward SFS and backward SFS have been completed. Initially, we found the KNN forward SFS with 15 features, to be promising. With the KNN backward SFS now completed, the model with 10 number of features is also very promising. Its accuracy rating is 0.85, with an f1-score of 0.89. The best feature selected for KNN has been determined to be one with an f1-score of 0.90.

4.2.3 Picking the best KNN model.

Optimal model with features

1) Forward
   Number of features = 15

2) Backward
   Number of features = 10

Best Feature selected for KNN.

The KNN model with 10 features where the features are obtained through Backward SFS.

```
Mean absolute error : 0.13249211356466878
Mean squared error : 0.13249211356466878
r2 score : 0.4074766355140187
The max error value : 1

accuracy score:  0.8675078864353313
confusion matrix: [[190  20]
 [ 22  85]]
              precision    recall  f1-score   support

           0       0.90      0.90      0.90       210
           1       0.81      0.79      0.80       107

    accuracy                           0.87       317
   macro avg       0.85      0.85      0.85       317
weighted avg       0.87      0.87      0.87       317
```

4.3 Decision Tree (DT) Model

4.3.1 Decision Tree Forward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'nO', 'F03[C-O]', 'SpMax_A', 'nN', 'SM6_B(m)'	{'criterion': 'entropy', 'max_depth': 4}	0.8089430894308943
10	'C%', 'nO', 'LOC', 'F03[C-O]', 'nArNO2', 'B03[C-Cl]', 'SpMax_A', 'nN', 'SM6_B(m)', 'nArCOOR'	{'criterion': 'entropy', 'max_depth': 12}	0.8089430894308943
15	'SpMax_L', 'J_Dz(e)', 'C%', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'F03[C-O]', 'nArNO2', 'B03[C-Cl]', 'SpMax_A', 'Psi_i_A', 'nN','SM6_B(m)', 'nArCOOR'	{'criterion': 'entropy', 'max_depth': 6}	0.8075880758807589
20	'SpMax_L', 'J_Dz(e)', 'nHM', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nArNO2', 'B03[C-Cl]', 'SpMax_A', 'F02[C-N]', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	{'criterion': 'entropy', 'max_depth': 6}	0.8089430894308943
25	'SpMax_L', 'J_Dz(e)', 'nHM', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nN-N', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	{'criterion': 'gini', 'max_depth': 4}	0.8075880758807589
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nO','F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me','nN-N', 'nArNO2', 'nCRX3', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'criterion': 'entropy', 'max_depth': 6}	0.8089430894308943
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'SdO', 'TI2_L', 'C-026', 'F02[C-N]','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'criterion': 'entropy', 'max_depth': 6}	0.8116531165311653
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'criterion': 'entropy', 'max_depth': 6}	0.8143631436314364

TRAINING SET RESULT FOR DT MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Training set Mean absolute error : 0.006775067750677507 Mean squared error : 0.006775067750677507 r2 score : 0.9696947298396038 The max error value : 1 accuracy score: 0.9932249322493225 [[487 2] [3 246]] precision recall f1-score support 0 0.99 1.00 0.99 489 1 0.99 0.99 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738
10	Training set Mean absolute error : 0.15040650406504066 Mean squared error : 0.15040650406504066 r2 score : 0.3272230024392049 The max error value : 1 accuracy score: 0.8495934959349594 [[446 43] [68 181]] precision recall f1-score support 0 0.87 0.91 0.89 489 1 0.81 0.73 0.77 249 accuracy 0.85 738 macro avg 0.84 0.82 0.83 738 weighted avg 0.85 0.85 0.85 738
15	Training set Mean absolute error : 0.009485094850948509 Mean squared error : 0.009485094850948509 r2 score : 0.9575726217754453 The max error value : 1 accuracy score: 0.9905149051490515 [[488 1] [6 243]] precision recall f1-score support 0 0.99 1.00 0.99 489 1 1.00 0.98 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738

20		Training set Mean absolute error : 0.13279132791327913 Mean squared error : 0.13279132791327913 r2 score : 0.40601670485623487 The max error value : 1 accuracy score: 0.8672086720867209 [[451 38] [60 189]] precision recall f1-score support 0 0.88 0.92 0.90 489 1 0.83 0.76 0.79 249 accuracy 0.87 738 macro avg 0.86 0.84 0.85 738 weighted avg 0.87 0.87 0.87 738	
25		Training set Mean absolute error : 0.006775067750677507 Mean squared error : 0.006775067750677507 r2 score : 0.9696947298396038 The max error value : 1 accuracy score: 0.9932249322493225 [[487 2] [3 246]] precision recall f1-score support 0 0.99 1.00 0.99 489 1 0.99 0.99 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738	
30		Training set Mean absolute error : 0.0921409214092141 Mean squared error : 0.0921409214092141 r2 score : 0.587848325818612 The max error value : 1 accuracy score: 0.907859078590786 [[450 39] [29 220]] precision recall f1-score support 0 0.94 0.92 0.93 489 1 0.85 0.88 0.87 249 accuracy 0.91 738 macro avg 0.89 0.90 0.90 738 weighted avg 0.91 0.91 0.91 738	

35		Training set Mean absolute error : 0.0013550135501355014 Mean squared error : 0.0013550135501355014 r2 score : 0.9939389459679208 The max error value : 1 accuracy score: 0.9986449864498645 [[489 0] [1 248]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 1.00 1.00 1.00 249 accuracy 1.00 738 macro avg 1.00 1.00 1.00 738 weighted avg 1.00 1.00 1.00 738	
40		Training set Mean absolute error : 0.07452574525745258 Mean squared error : 0.07452574525745258 r2 score : 0.666642028235642 The max error value : 1 accuracy score: 0.9254742547425474 [[469 20] [35 214]] precision recall f1-score support 0 0.93 0.96 0.94 489 1 0.91 0.86 0.89 249 accuracy 0.93 738 macro avg 0.92 0.91 0.92 738 weighted avg 0.93 0.93 0.92 738	

TEST SET RESULT FOR DECISION TREE MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted	
5	Mean absolute error : 0.2082018927444795 Mean squared error : 0.2082018927444795 r2 score : 0.06889185580774371 The max error value : 1 accuracy score: 0.7917981072555205 confusion_matrix [[184 26] [40 67]] precision recall f1-score support 0 0.82 0.88 0.85 210 1 0.72 0.63 0.67 107 accuracy 0.79 317 macro avg 0.77 0.75 0.76 317 weighted avg 0.79 0.79 0.79 317	

10		Mean absolute error : 0.2113564668769716 Mean squared error : 0.2113564668769716 r2 score : 0.0547841566533156 The max error value : 1 accuracy score: 0.7886435331230284 confusion_matrix [[174 36] [31 76]] precision recall f1-score support 0 0.85 0.83 0.84 210 1 0.68 0.71 0.69 107 accuracy 0.79 317 macro avg 0.76 0.77 0.77 317 weighted avg 0.79 0.79 0.79 317
15		Mean absolute error : 0.17034700315457413 Mean squared error : 0.17034700315457413 r2 score : 0.23818424566088114 The max error value : 1 accuracy score: 0.8296529968454258 confusion_matrix [[193 17] [37 70]] precision recall f1-score support 0 0.84 0.92 0.88 210 1 0.80 0.65 0.72 107 accuracy 0.83 317 macro avg 0.82 0.79 0.80 317 weighted avg 0.83 0.83 0.82 317
20		Mean absolute error : 0.1829652996845426 Mean squared error : 0.1829652996845426 r2 score : 0.1817534490431687 The max error value : 1 accuracy score: 0.8170347003154574 confusion_matrix [[191 19] [39 68]] precision recall f1-score support 0 0.83 0.91 0.87 210 1 0.78 0.64 0.70 107 accuracy 0.82 317 macro avg 0.81 0.77 0.78 317 weighted avg 0.81 0.82 0.81 317

25		Mean absolute error : 0.15772870662460567 Mean squared error : 0.15772870662460567 r2 score : 0.2946150422785937 The max error value : 1 accuracy score: 0.8422712933753943 confusion_matrix [[192 18] [32 75]] precision recall f1-score support 0 0.86 0.91 0.88 210 1 0.81 0.70 0.75 107 accuracy 0.84 317 macro avg 0.83 0.81 0.82 317 weighted avg 0.84 0.84 0.84 317
30		Mean absolute error : 0.14511041009463724 Mean squared error : 0.14511041009463724 r2 score : 0.35104583889630625 The max error value : 1 accuracy score: 0.8548895899053628 confusion_matrix [[186 24] [22 85]] precision recall f1-score support 0 0.89 0.89 0.89 210 1 0.78 0.79 0.79 107 accuracy 0.85 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.86 0.85 0.86 317
35		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion_matrix [[192 18] [27 80]] precision recall f1-score support 0 0.88 0.91 0.90 210 1 0.82 0.75 0.78 107 accuracy 0.86 317 macro avg 0.85 0.83 0.84 317 weighted avg 0.86 0.86 0.86 317

40		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion_matrix [[191 19] [26 81]] precision recall f1-score support 0 0.88 0.91 0.89 210 1 0.81 0.76 0.78 107 accuracy 0.86 317 macro avg 0.85 0.83 0.84 317 weighted avg 0.86 0.86 0.86 317

The decision tree forward SFS model is now completed. With what we have observed, the accuracy rating of each model hovers around 0.79 and 0.86. The most promising one is the one with 35 number of features selected. It has an accuracy rating of 0.86 and an f1-score of 0.90. We will proceed with the decision tree backward SFS model.

4.3.2 Decision Tree Backward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'SpMax_L', 'HyWi_B(m)', 'SM6_L', 'Psi_i_A', 'nN'	{'criterion': 'gini', 'max_depth': 6}	0.7655826558265583
10	'SpMax_L', 'HyWi_B(m)', 'LOC', 'SM6_L', 'Mi', 'nArNO2', 'B03[C-Cl]','Psi_i_A', 'nN', 'nArCOOR'	{'criterion': 'gini', 'max_depth': 10}	0.7859078590785907
15	'SpMax L', 'J Dz(e)', 'nO', 'F03[C-N]', 'HyWi B(m)', 'LOC', 'SM6 _L','F03[C-O]', 'Mi', 'nArNO2', 'B03[C-Cl]', 'Psi_i_A', 'nN', 'nArC OOR','nX'	{'criterion': 'entropy', 'max_depth': 8}	0.8252032520325203
20	'SpMax_L', 'J_Dz(e)', 'nCb-', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)','LOC', 'SM6 L', 'F03[C-O]', 'Me', 'Mi', 'nArNO2', 'B03[C-Cl]','F0 2[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'criterion': 'gini', 'max_depth': 6}	0.8089430894308943
25	'SpMax_L', 'J_Dz(e)', 'NssssC', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'Sdss C','HyWi B(m)', 'LOC', 'SM6 L', 'F03[C-O]', 'Me', 'Mi', 'nArNO2', 'SpPosA B(p)', 'B03[C-Cl]', 'N-073', 'SpMax A', 'F02[C-N]','SpMa x_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX']	{'criterion': 'entropy', 'max_depth': 6}	.8062330623306232
30	'SpMax_L', 'J_Dz(e)', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO','F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me','Mi', 'nN-N', 'nArNO2', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]','N-073', 'SpMax_A', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'criterion': 'entropy', 'max_depth': 6}	0.8089430894308943

35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'SdO', 'TI2_L', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'criterion': 'entropy', 'max_depth': 6}	0.8075880758807589
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-','C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'criterion': 'entropy', 'max_depth': 6}	0.8102981029810298

TRAINING SET RESULT FOR DECISION TREE MODEL BACKWARD SFS

Number of selected	The model output when the best parameter is fitted
5	Training set Mean absolute error : 0.16937669376693767 Mean squared error : 0.16937669376693767 r2 score : 0.24236824599009554 The max error value : 1 accuracy score: 0.8306233062330624 [[450 39] [86 163]] precision recall f1-score support 0 0.84 0.92 0.88 489 1 0.81 0.65 0.72 249 accuracy 0.83 738 macro avg 0.82 0.79 0.80 738 weighted avg 0.83 0.83 0.83 738
10	Training set Mean absolute error : 0.008130081300813009 Mean squared error : 0.008130081300813009 r2 score : 0.9636336758075246 The max error value : 1 accuracy score: 0.991869918699187 [[488 1] [5 244]] precision recall f1-score support 0 0.99 1.00 0.99 489 1 1.00 0.98 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738

15		Training set Mean absolute error : 0.037940379403794036 Mean squared error : 0.037940379403794036 r2 score : 0.8302904871017814 The max error value : 1 accuracy score: 0.962059620596206 [[482 7] [21 228]] precision recall f1-score support 0 0.96 0.99 0.97 489 1 0.97 0.92 0.94 249 accuracy 0.96 738 macro avg 0.96 0.95 0.96 738 weighted avg 0.96 0.96 0.96 738	
20		Training set Mean absolute error : 0.008130081300813009 Mean squared error : 0.008130081300813009 r2 score : 0.9636336758075246 The max error value : 1 accuracy score: 0.991869918699187 [[484 5] [1 248]] precision recall f1-score support 0 1.00 0.99 0.99 489 1 0.98 1.00 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738	
25		Training set Mean absolute error : 0.005420054200542005 Mean squared error : 0.005420054200542005 r2 score : 0.975755783871683 The max error value : 1 accuracy score: 0.994579945799458 [[487 2] [2 247]] precision recall f1-score support 0 1.00 1.00 1.00 489 1 0.99 0.99 0.99 249 accuracy 0.99 738 macro avg 0.99 0.99 0.99 738 weighted avg 0.99 0.99 0.99 738	

30		Training set Mean absolute error : 0.04065040650406504 Mean squared error : 0.04065040650406504 r2 score : 0.818168379037623 The max error value : 1 accuracy score: 0.959349593495935 [[474 15] [15 234]] precision recall f1-score support 0 0.97 0.97 0.97 489 1 0.94 0.94 0.94 249 accuracy 0.96 738 macro avg 0.95 0.95 0.95 738 weighted avg 0.96 0.96 0.96 738	
35		Training set Mean absolute error : 0.07452574525745258 Mean squared error : 0.07452574525745258 r2 score : 0.666642028235642 The max error value : 1 accuracy score: 0.9254742547425474 [[469 20] [35 214]] precision recall f1-score support 0 0.93 0.96 0.94 489 1 0.91 0.86 0.89 249 accuracy 0.93 738 macro avg 0.92 0.91 0.92 738 weighted avg 0.93 0.93 0.92 738	
40		Training set Mean absolute error : 0.07452574525745258 Mean squared error : 0.07452574525745258 r2 score : 0.666642028235642 The max error value : 1 accuracy score: 0.9254742547425474 [[469 20] [35 214]] precision recall f1-score support 0 0.93 0.96 0.94 489 1 0.91 0.86 0.89 249 accuracy 0.93 738 macro avg 0.92 0.91 0.92 738 weighted avg 0.93 0.93 0.92 738	

TEST SET RESULT FOR DECISION TREE MODEL BACKWARD SFS

Number of selected	The model output when the best parameter is fitted
5	Mean absolute error : 0.23659305993690852 Mean squared error : 0.23659305993690852 r2 score : -0.05807743658210951 The max error value : 1 accuracy score: 0.7634069400630915 confusion_matrix [[176 34] [41 66]] precision recall f1-score support 0 0.81 0.84 0.82 210 1 0.66 0.62 0.64 107 accuracy 0.76 317 macro avg 0.74 0.73 0.73 317 weighted avg 0.76 0.76 0.76 317
10	Mean absolute error : 0.1861198738170347 Mean squared error : 0.1861198738170347 r2 score : 0.1676457498887406 The max error value : 1 accuracy score: 0.8138801261829653 confusion_matrix [[181 29] [30 77]] precision recall f1-score support 0 0.86 0.86 0.86 210 1 0.73 0.72 0.72 107 accuracy 0.81 317 macro avg 0.79 0.79 0.79 317 weighted avg 0.81 0.81 0.81 317
15	Mean absolute error : 0.1608832807570978 Mean squared error : 0.1608832807570978 r2 score : 0.2805073431241656 The max error value : 1 accuracy score: 0.8391167192429022 confusion_matrix [[187 23] [28 79]] precision recall f1-score support 0 0.87 0.89 0.88 210 1 0.77 0.74 0.76 107 accuracy 0.84 317 macro avg 0.82 0.81 0.82 317 weighted avg 0.84 0.84 0.84 317

20		Mean absolute error : 0.17665615141955837 Mean squared error : 0.17665615141955837 r2 score : 0.20996884735202492 The max error value : 1 accuracy score: 0.8233438485804416 confusion_matrix [[181 29] [27 80]] precision recall f1-score support 0 0.87 0.86 0.87 210 1 0.73 0.75 0.74 107 accuracy 0.82 317 macro avg 0.80 0.80 0.80 317 weighted avg 0.82 0.82 0.82 317
25		Mean absolute error : 0.15772870662460567 Mean squared error : 0.15772870662460567 r2 score : 0.2946150422785937 The max error value : 1 accuracy score: 0.8422712933753943 confusion_matrix [[185 25] [25 82]] precision recall f1-score support 0 0.88 0.88 0.88 210 1 0.77 0.77 0.77 107 accuracy 0.84 317 macro avg 0.82 0.82 0.82 317 weighted avg 0.84 0.84 0.84 317
30		Mean absolute error : 0.17665615141955837 Mean squared error : 0.17665615141955837 r2 score : 0.20996884735202492 The max error value : 1 accuracy score: 0.8233438485804416 confusion_matrix [[182 28] [28 79]] precision recall f1-score support 0 0.87 0.87 0.87 210 1 0.74 0.74 0.74 107 accuracy 0.82 317 macro avg 0.80 0.80 0.80 317 weighted avg 0.82 0.82 0.82 317

35		Mean absolute error : 0.138801261829653 Mean squared error : 0.138801261829653 r2 score : 0.37926123720516247 The max error value : 1 accuracy score: 0.861198738170347 confusion_matrix [[192 18] [26 81]] precision recall f1-score support 0 0.88 0.91 0.90 210 1 0.82 0.76 0.79 107 accuracy 0.86 317 macro avg 0.85 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317	
40		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion_matrix [[192 18] [27 80]] precision recall f1-score support 0 0.88 0.91 0.90 210 1 0.82 0.75 0.78 107 accuracy 0.86 317 macro avg 0.85 0.83 0.84 317 weighted avg 0.86 0.86 0.86 317	

Both the Decision Tree forward SFS and the Decision Tree backward SFS models are completed, and the results are compared. The accuracy ratings of the DT backward SFS models are in the range of 0.76 to 0.86, and much like DT forward SFS model, the most promising model is the one with 35 number of features selected. The accuracy rating is 0.86 and the f1-score is 0.90. The best Decision Tree model is determined to be one with 35 number of features selected in a Decision Tree backward SFS model. It has an accuracy rating of 0.86 and f1 score of 0.90.

4.3.3 Picking the best Decision Tree model.

Optimal model with features

1) Forward
   Number of features = 35
2) Backward
   Number of features = 35

Best Feature selected for DT.

The DT model with 35 features where the features are obtained through Backward SFS.

```
Mean absolute error : 0.138801261829653
Mean squared error : 0.138801261829653
r2 score : 0.37926123720516247
The max error value : 1

accuracy score:  0.861198738170347
confusion_matrix [[192  18]
 [ 26  81]]
              precision    recall  f1-score   support

           0       0.88      0.91      0.90       210
           1       0.82      0.76      0.79       107

    accuracy                           0.86       317
   macro avg       0.85      0.84      0.84       317
weighted avg       0.86      0.86      0.86       317
```

4.4 Perceptron model

4.4.1 Perceptron Forward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'nO', 'F03[C-O]', 'SpMax_A', 'nN', 'SM6_B(m)'	{'alpha': 0.01, 'loss': 'modified_huber', 'penalty': 'none'}	0.799457994 5799458
10	'C%', 'nO', 'LOC', 'F03[C-O]', 'nArNO2', 'B03[C-Cl]', 'SpMax_A', 'nN', 'SM6_B(m)', 'nArCOOR'	'alpha': 0.1, 'loss': 'modified_huber', 'penalty': 'none'	0.844173441 7344174
15	'SpMax_L', 'J_Dz(e)', 'C%', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'F03[C-O]', 'nArNO2', 'B03[C-Cl]', 'SpMax_A', 'Psi_i_A', 'nN','SM6_B(m)', 'nArCOOR'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.853658536 5853658
20	'SpMax_L', 'J_Dz(e)', 'nHM', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nArNO2', 'B03[C-Cl]', 'SpMax_A', 'F02[C-N]', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'],	'alpha': 0.001, 'loss': 'hinge', 'penalty': 'l2'	0.849593495 9349594
25	'SpMax_L', 'J_Dz(e)', 'nHM', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nN-N', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.849593495 9349594
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nN-N', 'nArNO2', 'nCRX3', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A','C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l1'	0.855013550 1355015
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]','Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'SdO', 'TI2_L', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.863143631 4363143
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.861788617 8861789

TEST SET RESULT FOR PERCEPTRON MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.1829652996845426 Mean squared error : 0.1829652996845426 r2 score : 0.1817534490431687 The max error value : 1 accuracy score: 0.8170347003154574 confusion matrix: [[186 24] [34 73]] precision recall f1-score support 0 0.85 0.89 0.87 210 1 0.75 0.68 0.72 107 accuracy 0.82 317 macro avg 0.80 0.78 0.79 317 weighted avg 0.81 0.82 0.81 317
10	Mean absolute error : 0.13564668769716087 Mean squared error : 0.13564668769716087 r2 score : 0.3933689363595906 The max error value : 1 accuracy score: 0.8643533123028391 confusion matrix: [[189 21] [22 85]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.80 0.79 0.80 107 accuracy 0.86 317 macro avg 0.85 0.85 0.85 317 weighted avg 0.86 0.86 0.86 317
15	Mean absolute error : 0.12618296529968454 Mean squared error : 0.12618296529968454 r2 score : 0.4356920338228749 The max error value : 1 accuracy score: 0.8738170347003155 confusion matrix: [[190 20] [20 87]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.81 0.81 0.81 107 accuracy 0.87 317 macro avg 0.86 0.86 0.86 317 weighted avg 0.87 0.87 0.87 317

20		Mean absolute error : 0.1167192429022082 Mean squared error : 0.1167192429022082 r2 score : 0.47801513128615936 The max error value : 1 accuracy score: 0.8832807570977917 confusion matrix: [[193 17] [20 87]] precision recall f1-score support 0 0.91 0.92 0.91 210 1 0.84 0.81 0.82 107 accuracy 0.88 317 macro avg 0.87 0.87 0.87 317 weighted avg 0.88 0.88 0.88 317	
25		Mean absolute error : 0.13564668769716087 Mean squared error : 0.13564668769716087 r2 score : 0.3933689363595906 The max error value : 1 accuracy score: 0.8643533123028391 confusion matrix: [[188 22] [21 86]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.80 0.80 0.80 107 accuracy 0.86 317 macro avg 0.85 0.85 0.85 317 weighted avg 0.86 0.86 0.86 317	
30		Mean absolute error : 0.12933753943217666 Mean squared error : 0.12933753943217666 r2 score : 0.4215843346684468 The max error value : 1 accuracy score: 0.8706624605678234 confusion matrix: [[191 19] [22 85]] precision recall f1-score support 0 0.90 0.91 0.90 210 1 0.82 0.79 0.81 107 accuracy 0.87 317 macro avg 0.86 0.85 0.85 317 weighted avg 0.87 0.87 0.87 317	

35		Mean absolute error : 0.12933753943217666 Mean squared error : 0.12933753943217666 r2 score : 0.4215843346684468 The max error value : 1 accuracy score: 0.8706624605678234 confusion matrix: [[190 20] [21 86]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.81 0.80 0.81 107 accuracy 0.87 317 macro avg 0.86 0.85 0.86 317 weighted avg 0.87 0.87 0.87 317
40		Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 accuracy score: 0.886435331230284 confusion matrix: [[193 17] [19 88]] precision recall f1-score support 0 0.91 0.92 0.91 210 1 0.84 0.82 0.83 107 accuracy 0.89 317 macro avg 0.87 0.87 0.87 317 weighted avg 0.89 0.89 0.89 317

The Perceptron Forward SFS model is completed and all the outputs depending on the number of features selected is printed out as well. The accuracies depending on the number of features is selected. All of them generally float with an accuracy that is higher than 0.85, with the model with 40 selected features landing on a score of 0.89, the highest of all of them. We now proceed with the Perceptron Backward SFS model.

4.4.2 Perceptron Backward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'SpMax_L', 'HyWi_B(m)', 'SM6_L', 'Psi_i_A', 'nN'	{'alpha': 0.001, 'loss': 'log', 'penalty': 'l2'}	0.82926829 26829268
10	'SpMax_L', 'HyWi_B(m)', 'LOC', 'SM6_L', 'Mi', 'nArNO2', 'B03[C-Cl]', 'Psi_i_A', 'nN', 'nArCOOR'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'none'	0.86449864 49864499
15	'SpMax_L', 'J_Dz(e)', 'nO', 'F03[C-N]', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Mi', 'nArNO2', 'B03[C-Cl]', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	'alpha': 0.001, 'loss': 'hinge', 'penalty': 'l1'	0.86043360 43360433
20	'SpMax_L', 'J_Dz(e)', 'nCb-', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)','LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nArNO2', 'B03[C-Cl]','F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.85907859 07859079
25	'SpMax_L', 'J_Dz(e)', 'NssssC', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'SdssC','HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi','nArNO2','SpPosA_B(p)', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'F02[C-N]','SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.85501355 01355013
30	'SpMax_L', 'J_Dz(e)', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO','F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'C-026', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.86043360 43360433
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp','nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'SdO', 'TI2_L', 'C-026', 'F02[C-N]','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2'	0.86314363 14363143
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	'alpha': 0.01, 'loss': 'modified_huber', 'penalty': 'l2'	0.86856368 56368563

TEST SET RESULT FOR PERCEPTRON MODEL BACKWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.19873817034700317 Mean squared error : 0.19873817034700317 r2 score : 0.11121495327102804 The max error value : 1 accuracy score: 0.8012618296529969 confusion matrix: [[183 27] [36 71]] precision recall f1-score support 0 0.84 0.87 0.85 210 1 0.72 0.66 0.69 107 accuracy 0.80 317 macro avg 0.78 0.77 0.77 317 weighted avg 0.80 0.80 0.80 317
10	Mean absolute error : 0.12933753943217666 Mean squared error : 0.12933753943217666 r2 score : 0.4215843346684468 The max error value : 1 accuracy score: 0.8706624605678234 confusion matrix: [[191 19] [22 85]] precision recall f1-score support 0 0.90 0.91 0.90 210 1 0.82 0.79 0.81 107 accuracy 0.87 317 macro avg 0.86 0.85 0.85 317 weighted avg 0.87 0.87 0.87 317
15	Mean absolute error : 0.13249211356466878 Mean squared error : 0.13249211356466878 r2 score : 0.4074766355140187 The max error value : 1 accuracy score: 0.8675078864353313 confusion matrix: [[185 25] [17 90]] precision recall f1-score support 0 0.92 0.88 0.90 210 1 0.78 0.84 0.81 107 accuracy 0.87 317 macro avg 0.85 0.86 0.85 317 weighted avg 0.87 0.87 0.87 317

20		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion matrix: [[187 23] [22 85]] precision recall f1-score support 0 0.89 0.89 0.89 210 1 0.79 0.79 0.79 107 accuracy 0.86 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317	
25		Mean absolute error : 0.12933753943217666 Mean squared error : 0.12933753943217666 r2 score : 0.4215843346684468 The max error value : 1 accuracy score: 0.8706624605678234 confusion matrix: [[189 21] [20 87]] precision recall f1-score support 0 0.90 0.90 0.90 210 1 0.81 0.81 0.81 107 accuracy 0.87 317 macro avg 0.85 0.86 0.86 317 weighted avg 0.87 0.87 0.87 317	
30		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion matrix: [[190 20] [25 82]] precision recall f1-score support 0 0.88 0.90 0.89 210 1 0.80 0.77 0.78 107 accuracy 0.86 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317	

35		Mean absolute error : 0.13249211356466878 Mean squared error : 0.13249211356466878 r2 score : 0.4074766355140187 The max error value : 1 accuracy score: 0.8675078864353313 confusion matrix: [[192 18] [24 83]] precision recall f1-score support 0 0.89 0.91 0.90 210 1 0.82 0.78 0.80 107 accuracy 0.87 317 macro avg 0.86 0.84 0.85 317 weighted avg 0.87 0.87 0.87 317
40		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion matrix: [[181 29] [16 91]] precision recall f1-score support 0 0.92 0.86 0.89 210 1 0.76 0.85 0.80 107 accuracy 0.86 317 macro avg 0.84 0.86 0.85 317 weighted avg 0.86 0.86 0.86 317

4.4.3 Picking the best Perceptron model.

Now that we have obtained the results for both Perceptron Forward SFS and Perceptron Backward SFS, the analysis of the best model can be performed. The model that we find the best is the model with 40 selected features. It has a high accuracy score as well as an extremely high f1-score, making it a strong and reliable model when compared to the other Perceptron models. The parameters recorded for this model are ['alpha': 0.01, 'loss': 'hinge', 'penalty': 'l2']

```
Mean absolute error : 0.11356466876971609
Mean squared error : 0.11356466876971609
r2 score : 0.49212283044058747
The max error value : 1

accuracy score:  0.886435331230284
confusion matrix: [[193  17]
 [ 19  88]]
              precision    recall  f1-score   support

           0       0.91      0.92      0.91       210
           1       0.84      0.82      0.83       107

    accuracy                           0.89       317
   macro avg       0.87      0.87      0.87       317
weighted avg       0.89      0.89      0.89       317
```

4.5 Logistic Regression (LR) model
4.5.1 Logistic Regression Forward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'nCp', 'nCIR', 'SpMax_A', 'nN', 'nX'	{'C': 10.0, 'penalty': 'l2'}	0.8224932249322494
10	'nHM', 'nCp', 'nCRX3', 'nCIR', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'nN', 'nArCOOR', 'nX'	{'C': 100.0, 'penalty': 'l2'}	0.8401084010840109
15	'J_Dz(e)', 'nHM', 'F01[N-N]', 'nCp', 'LOC', 'nCRX3', 'nCIR', 'N-073','SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'TI2_L', 'nN', 'nArCOOR', 'nX'	{'C': 1.0, 'penalty': 'l2'}	0.8414634146341463
20	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'NssssC', 'nCp', 'F03[C-N]', 'LOC', 'SM6_L', 'nCRX3', 'nCIR', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'TI2_L', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'C': 100.0, 'penalty': 'l2'}	0.8550135501355015
25	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'NssssC', 'nCp', 'F03[C-N]', 'LOC', 'SM6_L', 'nN-N', 'nCRX3', 'nCIR', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'C': 10.0, 'penalty': 'l2'}	0.8482384823848239
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'NssssC', 'nCp', 'nO','F03[C-N]', 'SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'C': 1.0, 'penalty': 'l2'}	0.8577235772357724
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'C': 10.0, 'penalty': 'l2'}	0.8631436314363143
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'NssssC', 'nCb-','C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'C': 1.0, 'penalty': 'l2'}	0.8658536585365854

TRAINING SET RESULT FOR LR MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Training set Mean absolute error : 0.17479674796747968 Mean squared error : 0.17479674796747968 r2 score : 0.2181240298617786 The max error value : 1 accuracy score: 0.8252032520325203 [[436 53] [76 173]] precision recall f1-score support 0 0.85 0.89 0.87 489 1 0.77 0.69 0.73 249 accuracy 0.83 738 macro avg 0.81 0.79 0.80 738 weighted avg 0.82 0.83 0.82 738
10	Training set Mean absolute error : 0.15311653116531165 Mean squared error : 0.15311653116531165 r2 score : 0.3151008943750464 The max error value : 1 accuracy score: 0.8468834688346883 [[439 50] [63 186]] precision recall f1-score support 0 0.87 0.90 0.89 489 1 0.79 0.75 0.77 249 accuracy 0.85 738 macro avg 0.83 0.82 0.83 738 weighted avg 0.85 0.85 0.85 738
15	Training set Mean absolute error : 0.14769647696476965 Mean squared error : 0.14769647696476965 r2 score : 0.33934511050336336 The max error value : 1 accuracy score: 0.8523035230352304 [[446 43] [66 183]] precision recall f1-score support 0 0.87 0.91 0.89 489 1 0.81 0.73 0.77 249 accuracy 0.85 738 macro avg 0.84 0.82 0.83 738 weighted avg 0.85 0.85 0.85 738

20		Training set Mean absolute error : 0.12601626016260162 Mean squared error : 0.12601626016260162 r2 score : 0.43632197501663106 The max error value : 1 accuracy score: 0.8739837398373984 [[453 36] [57 192]] precision recall f1-score support 0 0.89 0.93 0.91 489 1 0.84 0.77 0.81 249 accuracy 0.87 738 macro avg 0.87 0.85 0.86 738 weighted avg 0.87 0.87 0.87 738	
25		Training set Mean absolute error : 0.13143631436314362 Mean squared error : 0.13143631436314362 r2 score : 0.4120777588883141 The max error value : 1 accuracy score: 0.8685636856368564 [[452 37] [60 189]] precision recall f1-score support 0 0.88 0.92 0.90 489 1 0.84 0.76 0.80 249 accuracy 0.87 738 macro avg 0.86 0.84 0.85 738 weighted avg 0.87 0.87 0.87 738	
30		Training set Mean absolute error : 0.11246612466124661 Mean squared error : 0.11246612466124661 r2 score : 0.49693251533742344 The max error value : 1 accuracy score: 0.8875338753387534 [[456 33] [50 199]] precision recall f1-score support 0 0.90 0.93 0.92 489 1 0.86 0.80 0.83 249 accuracy 0.89 738 macro avg 0.88 0.87 0.87 738 weighted avg 0.89 0.89 0.89 738	

35		Training set Mean absolute error : 0.10975609756097561 Mean squared error : 0.10975609756097561 r2 score : 0.5090546234015819 The max error value : 1 accuracy score: 0.8902439024390244 [[455 34] [47 202]] precision recall f1-score support 0 0.91 0.93 0.92 489 1 0.86 0.81 0.83 249 accuracy 0.89 738 macro avg 0.88 0.87 0.88 738 weighted avg 0.89 0.89 0.89 738	
40		Training set Mean absolute error : 0.11924119241192412 Mean squared error : 0.11924119241192412 r2 score : 0.46662724517702725 The max error value : 1 accuracy score: 0.8807588075880759 [[450 39] [49 200]] precision recall f1-score support 0 0.90 0.92 0.91 489 1 0.84 0.80 0.82 249 accuracy 0.88 738 macro avg 0.87 0.86 0.87 738 weighted avg 0.88 0.88 0.88 738	

TEST SET RESULT FOR LR MODEL FORWARD SFS

Number of selected features	The model output when the best parameter is fitted
5	Mean absolute error : 0.1829652996845426 Mean squared error : 0.1829652996845426 r2 score : 0.1817534490431687 The max error value : 1 accuracy score: 0.8170347003154574 confusion_matrix [[181 29] [29 78]] precision recall f1-score support 0 0.86 0.86 0.86 210 1 0.73 0.73 0.73 107 accuracy 0.82 317 macro avg 0.80 0.80 0.80 317 weighted avg 0.82 0.82 0.82 317
10	Mean absolute error : 0.17034700315457413 Mean squared error : 0.17034700315457413 r2 score : 0.23818424566088114 The max error value : 1 accuracy score: 0.8296529968454258 confusion_matrix [[184 26] [28 79]] precision recall f1-score support 0 0.87 0.88 0.87 210 1 0.75 0.74 0.75 107 accuracy 0.83 317 macro avg 0.81 0.81 0.81 317 weighted avg 0.83 0.83 0.83 317
15	Mean absolute error : 0.14511041009463724 Mean squared error : 0.14511041009463724 r2 score : 0.35104583889630625 The max error value : 1 accuracy score: 0.8548895899053628 confusion_matrix [[192 18] [28 79]] precision recall f1-score support 0 0.87 0.91 0.89 210 1 0.81 0.74 0.77 107 accuracy 0.85 317 macro avg 0.84 0.83 0.83 317 weighted avg 0.85 0.85 0.85 317

20		Mean absolute error : 0.14826498422712933 Mean squared error : 0.14826498422712933 r2 score : 0.33693813974187803 The max error value : 1 accuracy score: 0.8517350157728707 confusion_matrix [[186 24] [23 84]] precision recall f1-score support 0 0.89 0.89 0.89 210 1 0.78 0.79 0.78 107 accuracy 0.85 317 macro avg 0.83 0.84 0.83 317 weighted avg 0.85 0.85 0.85 317
25		Mean absolute error : 0.14511041009463724 Mean squared error : 0.14511041009463724 r2 score : 0.35104583889630625 The max error value : 1 accuracy score: 0.8548895899053628 confusion_matrix [[187 23] [23 84]] precision recall f1-score support 0 0.89 0.89 0.89 210 1 0.79 0.79 0.79 107 accuracy 0.85 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.85 0.85 0.85 317
30		Mean absolute error : 0.14195583596214512 Mean squared error : 0.14195583596214512 r2 score : 0.36515353805073436 The max error value : 1 accuracy score: 0.8580441640378549 confusion_matrix [[186 24] [21 86]] precision recall f1-score support 0 0.90 0.89 0.89 210 1 0.78 0.80 0.79 107 accuracy 0.86 317 macro avg 0.84 0.84 0.84 317 weighted avg 0.86 0.86 0.86 317

35		Mean absolute error : 0.12933753943217666 Mean squared error : 0.12933753943217666 r2 score : 0.4215843346684468 The max error value : 1 accuracy score: 0.8706624605678234 confusion_matrix [[188 22] [19 88]] precision recall f1-score support 0 0.91 0.90 0.90 210 1 0.80 0.82 0.81 107 accuracy 0.87 317 macro avg 0.85 0.86 0.86 317 weighted avg 0.87 0.87 0.87 317	
40		Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 accuracy score: 0.886435331230284 confusion_matrix [[190 20] [16 91]] precision recall f1-score support 0 0.92 0.90 0.91 210 1 0.82 0.85 0.83 107 accuracy 0.89 317 macro avg 0.87 0.88 0.87 317 weighted avg 0.89 0.89 0.89 317	

The LR forward SFS model is completed, and the results are observed in terms of the number of features selected. The accuracy ratings of each model hovers between 0.82 and 0.89. The most promising model is the one with 40 numbers of features selected. It has an accuracy rating of 0.89 and an f1-score of 0.91. We will now proceed with LR backward SFS.

4.5.2 Logistic Regression Backward SFS

Number of features selected	Features that were selected	Best parameters	Training score
5	'SpMax_L', 'nHM', 'SdO', 'nCrt', 'nN'	{'C': 10.0, 'penalty': 'l2'}	0.8523035230352303
10	'SpMax_L', 'nHM', 'NssssC', 'nCb-', 'Mi', 'nArNO2', 'SdO', 'nCrt', 'nHDon', 'nN'	{'C': 10.0, 'penalty': 'l2'}	0.8523035230352303
15	'SpMax_L', 'nHM', 'NssssC', 'nCb-', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'Mi', 'nArNO2', 'SdO', 'nCrt', 'nHDon', 'nN', 'nArCOOR'	{'C': 10.0, 'penalty': 'l2'}	0.8604336043360433
20	'SpMax_L', 'J_Dz(e)', 'nHM', 'NssssC', 'nCb-', 'SdssC', 'HyWi_B(m)','LOC', 'SM6_L', 'Mi', 'nArNO2', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'nCrt','F02[C-N]', 'nHDon', 'Psi_i_A', 'nN', 'nArCOOR'	{'C': 10.0, 'penalty': 'l2'}	0.8631436314363143
25	'SpMax_L', 'J_Dz(e)', 'nHM', 'NssssC', 'nCb-', 'C%', 'nCp', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Mi', 'nArNO2', 'SpPosA_B(p)', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon', 'Psi_i_A', 'nN', 'nArCOOR'	{'C': 1.0, 'penalty': 'l2'}	0.8631436314363143
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Mi', 'nN-N', 'nArNO2', 'SpPosA_B(p)', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'C': 1.0, 'penalty': 'l2'}	0.8712737127371274
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Mi', 'nN-N', 'nArNO2', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]','nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'nArCOOR', 'nX'	{'C': 1.0, 'penalty': 'l2'}	0.8672086720867208
40	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'NssssC', 'nCb-', 'C%', 'nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCRX3', 'SpPosA_B(p)', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt', 'C-026', 'F02[C-N]', 'nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	{'C': 1.0, 'penalty': 'l2'}	0.8658536585365854

TRAINING SET RESULT FOR LR MODEL BACKWARD SFS

Number of selected	The model output when the best parameter is fitted
5	Training set Mean absolute error : 0.15311653116531165 Mean squared error : 0.15311653116531165 r2 score : 0.3151008943750464 The max error value : 1 accuracy score: 0.8468834688346883 [[447 42] [71 178]] precision recall f1-score support 0 0.86 0.91 0.89 489 1 0.81 0.71 0.76 249 accuracy 0.85 738 macro avg 0.84 0.81 0.82 738 weighted avg 0.84 0.85 0.84 738
10	Training set Mean absolute error : 0.13414634146341464 Mean squared error : 0.13414634146341464 r2 score : 0.39995565082415563 The max error value : 1 accuracy score: 0.8658536585365854 [[446 43] [56 193]] precision recall f1-score support 0 0.89 0.91 0.90 489 1 0.82 0.78 0.80 249 accuracy 0.87 738 macro avg 0.85 0.84 0.85 738 weighted avg 0.86 0.87 0.86 738
15	Training set Mean absolute error : 0.12330623306233063 Mean squared error : 0.12330623306233063 r2 score : 0.44844408308078954 The max error value : 1 accuracy score: 0.8766937669376694 [[448 41] [50 199]] precision recall f1-score support 0 0.90 0.92 0.91 489 1 0.83 0.80 0.81 249 accuracy 0.88 738 macro avg 0.86 0.86 0.86 738 weighted avg 0.88 0.88 0.88 738

20		Training set Mean absolute error : 0.12330623306233063 Mean squared error : 0.12330623306233063 r2 score : 0.44844408308078954 The max error value : 1 accuracy score: 0.8766937669376694 [[447 42] [49 200]] precision recall f1-score support 0 0.90 0.91 0.91 489 1 0.83 0.80 0.81 249 accuracy 0.88 738 macro avg 0.86 0.86 0.86 738 weighted avg 0.88 0.88 0.88 738	
25		Training set Mean absolute error : 0.12737127371273713 Mean squared error : 0.12737127371273713 r2 score : 0.4302609209845518 The max error value : 1 accuracy score: 0.8726287262872628 [[447 42] [52 197]] precision recall f1-score support 0 0.90 0.91 0.90 489 1 0.82 0.79 0.81 249 accuracy 0.87 738 macro avg 0.86 0.85 0.86 738 weighted avg 0.87 0.87 0.87 738	
30		Training set Mean absolute error : 0.11924119241192412 Mean squared error : 0.11924119241192412 r2 score : 0.46662724517702725 The max error value : 1 accuracy score: 0.8807588075880759 [[450 39] [49 200]] precision recall f1-score support 0 0.90 0.92 0.91 489 1 0.84 0.80 0.82 249 accuracy 0.88 738 macro avg 0.87 0.86 0.87 738 weighted avg 0.88 0.88 0.88 738	

35		Enter the penalty : 35 Training set Mean absolute error : 0.11924119241192412 Mean squared error : 0.11924119241192412 r2 score : 0.46662724517702725 The max error value : 1 accuracy score: 0.8807588075880759 [[451 38] [50 199]] precision recall f1-score support 0 0.90 0.92 0.91 489 1 0.84 0.80 0.82 249 accuracy 0.88 738 macro avg 0.87 0.86 0.87 738 weighted avg 0.88 0.88 0.88 738	
40		Training set Mean absolute error : 0.11924119241192412 Mean squared error : 0.11924119241192412 r2 score : 0.46662724517702725 The max error value : 1 accuracy score: 0.8807588075880759 [[451 38] [50 199]] precision recall f1-score support 0 0.90 0.92 0.91 489 1 0.84 0.80 0.82 249 accuracy 0.88 738 macro avg 0.87 0.86 0.87 738 weighted avg 0.88 0.88 0.88 738	

TEST SET RESULT FOR LR MODEL BACKWARD SFS

Number of selected	The model output when the best parameter is fitted
5	Mean absolute error : 0.15772870662460567 Mean squared error : 0.15772870662460567 r2 score : 0.2946150422785937 The max error value : 1 accuracy score: 0.8422712933753943 confusion_matrix [[184 26] [24 83]] precision recall f1-score support 0 0.88 0.88 0.88 210 1 0.76 0.78 0.77 107 accuracy 0.84 317 macro avg 0.82 0.83 0.82 317 weighted avg 0.84 0.84 0.84 317
10	Mean absolute error : 0.12618296529968454 Mean squared error : 0.12618296529968454 r2 score : 0.4356920338228749 The max error value : 1 accuracy score: 0.8738170347003155 confusion_matrix [[189 21] [19 88]] precision recall f1-score support 0 0.91 0.90 0.90 210 1 0.81 0.82 0.81 107 accuracy 0.87 317 macro avg 0.86 0.86 0.86 317 weighted avg 0.87 0.87 0.87 317
15	Mean absolute error : 0.12618296529968454 Mean squared error : 0.12618296529968454 r2 score : 0.4356920338228749 The max error value : 1 accuracy score: 0.8738170347003155 confusion_matrix [[189 21] [19 88]] precision recall f1-score support 0 0.91 0.90 0.90 210 1 0.81 0.82 0.81 107 accuracy 0.87 317 macro avg 0.86 0.86 0.86 317 weighted avg 0.87 0.87 0.87 317

20		Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 accuracy score: 0.886435331230284 confusion_matrix [[191 19] [17 90]] precision recall f1-score support 0 0.92 0.91 0.91 210 1 0.83 0.84 0.83 107 accuracy 0.89 317 macro avg 0.87 0.88 0.87 317 weighted avg 0.89 0.89 0.89 317
25		Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 accuracy score: 0.886435331230284 confusion_matrix [[192 18] [18 89]] precision recall f1-score support 0 0.91 0.91 0.91 210 1 0.83 0.83 0.83 107 accuracy 0.89 317 macro avg 0.87 0.87 0.87 317 weighted avg 0.89 0.89 0.89 317
30		Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 accuracy score: 0.886435331230284 confusion_matrix [[192 18] [18 89]] precision recall f1-score support 0 0.91 0.91 0.91 210 1 0.83 0.83 0.83 107 accuracy 0.89 317 macro avg 0.87 0.87 0.87 317 weighted avg 0.89 0.89 0.89 317

35		Mean absolute error : 0.1167192429022082 Mean squared error : 0.1167192429022082 r2 score : 0.47801513128615936 The max error value : 1 accuracy score: 0.8832807570977917 confusion_matrix [[191 19] [18 89]] precision recall f1-score support 0 0.91 0.91 0.91 210 1 0.82 0.83 0.83 107 accuracy 0.88 317 macro avg 0.87 0.87 0.87 317 weighted avg 0.88 0.88 0.88 317
40		Mean absolute error : 0.11356466876971609 Mean squared error : 0.11356466876971609 r2 score : 0.49212283044058747 The max error value : 1 accuracy score: 0.886435331230284 confusion_matrix [[191 19] [17 90]] precision recall f1-score support 0 0.92 0.91 0.91 210 1 0.83 0.84 0.83 107 accuracy 0.89 317 macro avg 0.87 0.88 0.87 317 weighted avg 0.89 0.89 0.89 317

The LR backward SFS model is now completed, the results were obtained and are now compared. The accuracy ratings for each number of feature floats in the range of 0.84 and 0.89, with the most promising model being the one with 20 features selected. It has an accuracy rating of 0.89 and an f1-score of 0.91.

The best Logistic Regression model is determined to be one with backward feature selection with 20 selected features. It has an accuracy rating of 0.89 which is high and an f1 score of 0.91.

4.5.3 Picking the best Logistic Regression model.

Optimal model with features

1) Forward
   Number of features = 40
2) Backward
   Number of features = 20

Best Feature selected for LR.

The LR model with 20 features where the features are obtained through Backward SFS.

```
Mean absolute error : 0.11356466876971609
Mean squared error : 0.11356466876971609
r2 score : 0.49212283044058747
The max error value : 1

accuracy score:  0.886435331230284
confusion_matrix [[191  19]
 [ 17  90]]
              precision    recall  f1-score   support

           0       0.92      0.91      0.91       210
           1       0.83      0.84      0.83       107

    accuracy                           0.89       317
   macro avg       0.87      0.88      0.87       317
weighted avg       0.89      0.89      0.89       317
```

4.6 Neural Network

For the neural network, we performed feature selection as well. Feature selection for 'Forward' and 'Backward' were done. The feature selection is done in increment of 5. But since this is a neural network, it is started from 20 features first and we increment by 5 from there. Hence, the sequence would be 20,25,30 and 35. The selected feature will be used to train the model later. The layer consists of Linear, ReLU and Sigmoid functions. The model is trained for 500 epochs. After that, the model is evaluated. The results are stored in the table below both for 'Forward SFS' and 'Backward SFS'. After this is done, the best neural network model among the set of neural network model created is picked and evaluated.
The layers used in the neural network is a follows:

```
When the number of features
ranges from 20 to 35.
neural_model = nn.Sequential(
    nn.Linear(30,15),
    nn.ReLU(),
    nn.Linear(15,7),
    nn.ReLU(),
    nn.Linear(7,1),
    nn.Sigmoid()
)
```

4.6.1 Neural Network Forward SFS

TRAINING SET RESULT

Number of features selected	Selected feature	Training Set Output
20	'SpMax_L', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'SdssC', 'LOC', 'SM6_L','F03[C-O]', 'Mi', 'nN-N', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'nCrt','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	For training set Mean absolute error : 0.22397892 Mean squared error : 0.22397892 r2 score : -0.005925072641626006 The max error value : 1.0 (759, 1) accuracy score: 0.7760210803689065 [[461 44] [126 128]] precision recall f1-score support 0.0 0.79 0.91 0.84 505 1.0 0.74 0.50 0.60 254 accuracy 0.78 759 macro avg 0.76 0.71 0.72 759 weighted avg 0.77 0.78 0.76 759
25	'SpMax_L', 'J_Dz(e)', 'F01[N-N]', 'nCb-', 'C%', 'nO', 'F03[C-N]','SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'Mi', 'nN-N', 'nArNO2','B01[C-Br]', 'B03[C-Cl]', 'SdO', 'nCrt', 'F02[C-N]', 'SpMax_B(m)','Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	For training set Mean absolute error : 0.1949934 Mean squared error : 0.1949934 r2 score : 0.12425346617081967 The max error value : 1.0 (759, 1) accuracy score: 0.8050065876152833 [[487 18] [130 124]] precision recall f1-score support 0.0 0.79 0.96 0.87 505 1.0 0.87 0.49 0.63 254 accuracy 0.81 759 macro avg 0.83 0.73 0.75 759 weighted avg 0.82 0.81 0.79 759

30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'nCb-', 'C%', 'nO', 'F03[C-N]','SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2','nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'N-073', 'SpMax_A', 'SdO', 'nCrt','F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	For training set Mean absolute error : 0.21343873 Mean squared error : 0.21343873 r2 score : 0.04141257783562702 The max error value : 1.0 (759, 1) accuracy score: 0.7865612648221344 [[501 4] [158 96]] precision recall f1-score support 0.0 0.76 0.99 0.86 505 1.0 0.96 0.38 0.54 254 accuracy 0.79 759 macro avg 0.86 0.69 0.70 759 weighted avg 0.83 0.79 0.75 759
35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'nCb-', 'C%','nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Me', 'Mi', 'nN-N', 'nArNO2', 'nCIR', 'B01[C-Br]','B03[C-Cl]', 'N-073', 'SpMax_A', 'SdO', 'TI2_L', 'nCrt', 'F02[C-N]','nHDon', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	For training set Mean absolute error : 0.21080369 Mean squared error : 0.21080369 r2 score : 0.05324699045494019 The max error value : 1.0 (759, 1) accuracy score: 0.7891963109354414 [[382 123] [37 217]] precision recall f1-score support 0.0 0.91 0.76 0.83 505 1.0 0.64 0.85 0.73 254 accuracy 0.79 759 macro avg 0.77 0.81 0.78 759 weighted avg 0.82 0.79 0.79 759

TEST SET RESULT FOR NN MODEL FORWARD SFS

Number of selected features	The model output for test set
20	For testing set Mean absolute error : 0.25592417 Mean squared error : 0.25592417 r2 score : -0.14627767840624073 The max error value : 1.0 (211, 1) accuracy score: 0.7440758293838863 [[118 22] [32 39]] precision recall f1-score support 0.0 0.79 0.84 0.81 140 1.0 0.64 0.55 0.59 71 accuracy 0.74 211 macro avg 0.71 0.70 0.70 211 weighted avg 0.74 0.74 0.74 211

25	For testing set Mean absolute error : 0.22748815 Mean squared error : 0.22748815 r2 score : -0.018913491916658254 The max error value : 1.0 (211, 1) accuracy score: 0.7725118483412322 [[129 11] [37 34]] precision recall f1-score support 0.0 0.78 0.92 0.84 140 1.0 0.76 0.48 0.59 71 accuracy 0.77 211 macro avg 0.77 0.70 0.71 211 weighted avg 0.77 0.77 0.76 211
30	For testing set Mean absolute error : 0.21327014 Mean squared error : 0.21327014 r2 score : 0.04476860132813276 The max error value : 1.0 (211, 1) accuracy score: 0.7867298578199052 [[137 3] [42 29]] precision recall f1-score support 0.0 0.77 0.98 0.86 140 1.0 0.91 0.41 0.56 71 accuracy 0.79 211 macro avg 0.84 0.69 0.71 211 weighted avg 0.81 0.79 0.76 211
35	For testing set Mean absolute error : 0.19905214 Mean squared error : 0.19905214 r2 score : 0.10845069457292389 The max error value : 1.0 (211, 1) accuracy score: 0.8009478672985783 [[102 38] [4 67]] precision recall f1-score support 0.0 0.96 0.73 0.83 140 1.0 0.64 0.94 0.76 71 accuracy 0.80 211 macro avg 0.80 0.84 0.80 211 weighted avg 0.85 0.80 0.81 211

The table above shows the models that has been created by using the features selected by 'Forward SFS'. If we study the table, we can say that the model started classifying the data correctly when the number of selected features is 30. But when the number of features selected is more than 30, the model has started to misclassify the data severely which affects its performance. Moving on to 'Backward SFS' to gain more outputs.

4.6.2 Neural Network Backward SFS

Number of features selected	Selected feature	Training Set Output
20	'SpMax_L', 'J_Dz(e)', 'nHM', 'nCb-', 'C%', 'nO', 'F03[C-N]', 'LOC','SM6_L', 'F03[C-O]', 'Me', 'nArNO2', 'B01[C-Br]', 'B03[C-Cl]', 'nCrt','SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	For training set Mean absolute error : 0.64426875 Mean squared error : 0.64426875 r2 score : -1.893513885422089 The max error value : 1.0 (759, 1) accuracy score: 0.3557312252964427 [[16 489] [0 254]] precision recall f1-score support 0.0 1.00 0.03 0.06 505 1.0 0.34 1.00 0.51 254 accuracy 0.36 759 macro avg 0.67 0.52 0.29 759 weighted avg 0.78 0.36 0.21 759
25	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'nCb-', 'C%', 'nO', 'F03[C-N]','SdssC', 'LOC', 'SM6_L', 'F03[C-O]', 'Me', 'nN-N', 'nArNO2','B01[C-Br]', 'B03[C-Cl]', 'SdO', 'nCrt', 'F02[C-N]', 'SpMax_B(m)','Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR'	For training set Mean absolute error : 0.33465084 Mean squared error : 0.33465084 r2 score : -0.5029704026527824 The max error value : 1.0 (759, 1) accuracy score: 0.6653491436100132 [[505 0] [254 0]] precision recall f1-score support 0.0 0.67 1.00 0.80 505 1.0 0.00 0.00 0.00 254 accuracy 0.67 759 macro avg 0.33 0.50 0.40 759 weighted avg 0.44 0.67 0.53 759
30	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'nCb-', 'C%', 'nO','F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L', 'F03[C-O]', 'Me','nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B01[C-Br]', 'B03[C-Cl]', 'SpMax_A','SdO', 'nCrt', 'F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)','nArCOOR'	For training set Mean absolute error : 0.33465084 Mean squared error : 0.33465084 r2 score : -0.5029704026527824 The max error value : 1.0 (759, 1) accuracy score: 0.6653491436100132 [[505 0] [254 0]] precision recall f1-score support 0.0 0.67 1.00 0.80 505 1.0 0.00 0.00 0.00 254 accuracy 0.67 759 macro avg 0.33 0.50 0.40 759 weighted avg 0.44 0.67 0.53 759

35	'SpMax_L', 'J_Dz(e)', 'nHM', 'F01[N-N]', 'F04[C-N]', 'nCb-', 'C%','nCp', 'nO', 'F03[C-N]', 'SdssC', 'HyWi_B(m)', 'LOC', 'SM6_L','F03[C-O]', 'Me', 'nN-N', 'nArNO2', 'nCRX3', 'nCIR', 'B01[C-Br]','B03[C-Cl]', 'N-073', 'SpMax_A', 'B04[C-Br]', 'SdO', 'TI2_L', 'nCrt','F02[C-N]', 'SpMax_B(m)', 'Psi_i_A', 'nN', 'SM6_B(m)', 'nArCOOR', 'nX'	For training set Mean absolute error : 0.14492753 Mean squared error : 0.14492753 r2 score : 0.3491073059377714 The max error value : 1.0 (759, 1) accuracy score: 0.855072463768116 [[455 50] [60 194]] precision recall f1-score support 0.0 0.88 0.90 0.89 505 1.0 0.80 0.76 0.78 254 accuracy 0.86 759 macro avg 0.84 0.83 0.84 759 weighted avg 0.85 0.86 0.85 759

TEST SET RESULT FOR NN MODEL BACKWARD SFS

Number of selected features	The model output for test set
20	For testing set Mean absolute error : 0.6445498 Mean squared error : 0.6445498 r2 score : -1.8869215604305318 The max error value : 1.0 (211, 1) accuracy score: 0.35545023696682465 [[4 136] [0 71]] precision recall f1-score support 0.0 1.00 0.03 0.06 140 1.0 0.34 1.00 0.51 71 accuracy 0.36 211 macro avg 0.67 0.51 0.28 211 weighted avg 0.78 0.36 0.21 211
25	For testing set Mean absolute error : 0.33175355 Mean squared error : 0.33175355 r2 score : -0.4859155090451268 The max error value : 1.0 (211, 1) accuracy score: 0.6682464454976303 [[140 0] [70 1]] precision recall f1-score support 0.0 0.67 1.00 0.80 140 1.0 1.00 0.01 0.03 71 accuracy 0.67 211 macro avg 0.83 0.51 0.41 211 weighted avg 0.78 0.67 0.54 211

30	For testing set Mean absolute error : 0.3364929 Mean squared error : 0.3364929 r2 score : -0.5071428734600572 The max error value : 1.0 (211, 1) accuracy score: 0.6635071090047393 [[140 0] [71 0]] precision recall f1-score support 0.0 0.66 1.00 0.80 140 1.0 0.00 0.00 0.00 71 accuracy 0.66 211 macro avg 0.33 0.50 0.40 211 weighted avg 0.44 0.66 0.53 211
35	For testing set Mean absolute error : 0.15165877 Mean squared error : 0.15165877 r2 score : 0.3207243387222277 The max error value : 1.0 (211, 1) accuracy score: 0.8483412322274881 [[124 16] [16 55]] precision recall f1-score support 0.0 0.89 0.89 0.89 140 1.0 0.77 0.77 0.77 71 accuracy 0.85 211 macro avg 0.83 0.83 0.83 211 weighted avg 0.85 0.85 0.85 211

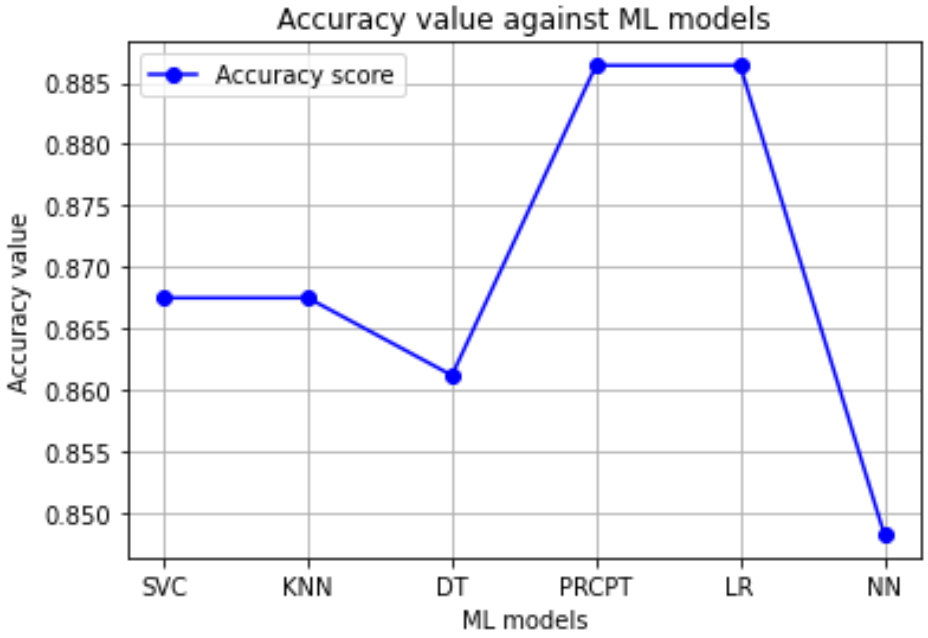
4.6.3 Picking the best Neural Network model.

After obtaining the results for both 'Forward SFS' and 'Backward SFS', the analysis of the best model can be done. We must pick a model that has high accuracy score which means it can successfully classify the data. With that said, the best model will be the model with 35 feature that was selected via 'Backward SFS' with the accuracy score of 0.848. This is because this model has a high accuracy score and the score for error is the lowest which makes it suitable. The result of the model is below:

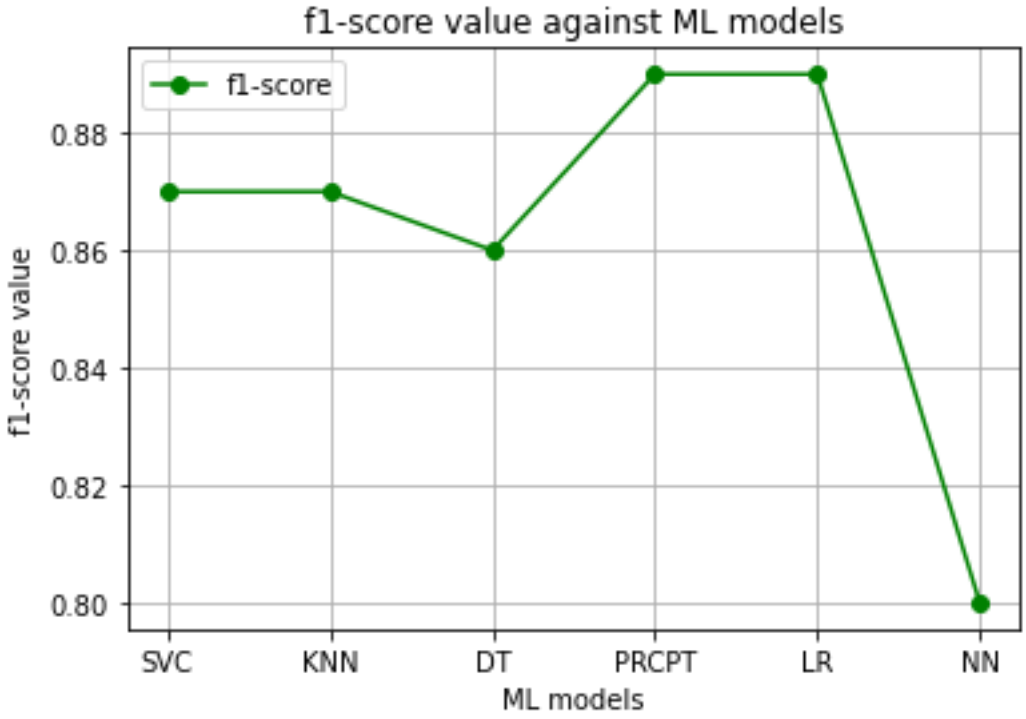
```
Mean absolute error : 0.15165877
Mean squared error : 0.15165877
r2 score : 0.3207243387222277
The max error value : 1.0

(211, 1)
accuracy score:  0.8483412322274881
[[124  16]
 [ 16  55]]
              precision    recall  f1-score   support

         0.0       0.89      0.89      0.89       140
         1.0       0.77      0.77      0.77        71

    accuracy                           0.85       211
   macro avg       0.83      0.83      0.83       211
weighted avg       0.85      0.85      0.85       211
```

4.7 Overall Best Machine Learning Model

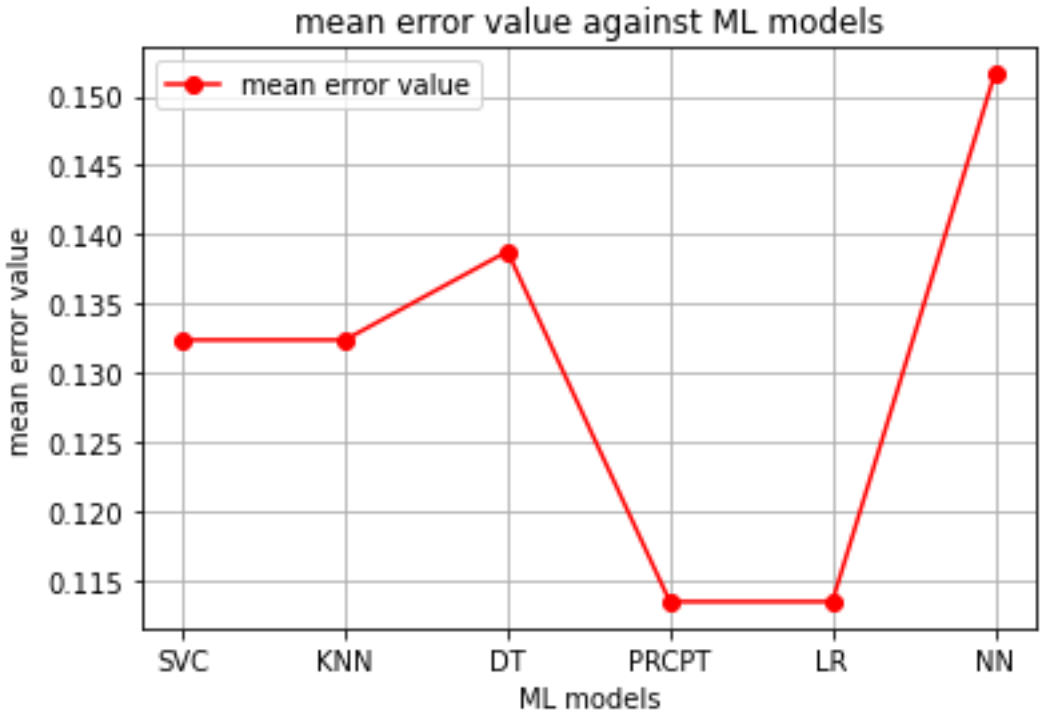
In order to determine the overall machine learning model for this dataset, we have decided to plot a few graphs that will help us to determine which model is the best. We focused on the following performance metric which are the accuracy score, f1-score, mean error value, recall and precision value, and number of feature used by the model. The analysis on this metrics will help us to determine which machine learning model is the best for the dataset. The graphs are shown below:



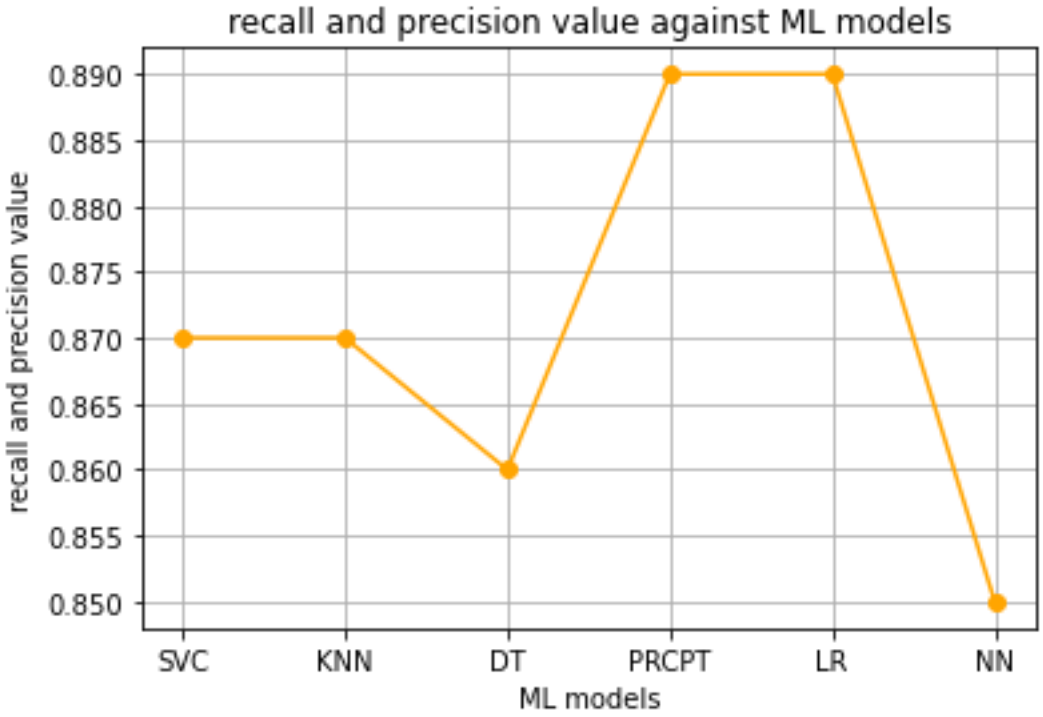
The graph above shows the accuracy value against each ML models.



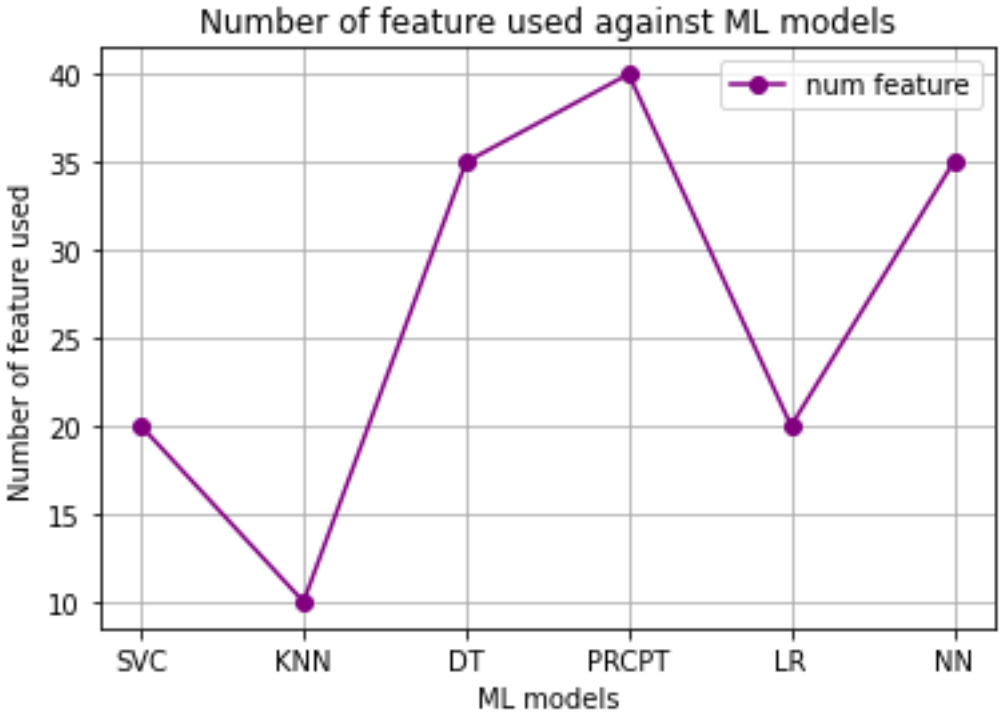
The graph above shows the f1-score value against each ML models.



The graph above shows the mean error value against each ML models.



The graph above shows the recall and precision value against each ML models.



The graph above shows the number of feature used against each ML models.

The best model should have high accuracy, f1-score, precision and recall value. It also must also have smaller mean error value and the preferred number of feature used should be minimum. So, by looking at the graphs, we can say that the top performing models are the Perceptron and the Logistic Regression model. These 2 models did outperform the other models. But in terms of the number of features used, it varies where the Logistic Regression model uses 20 features only while the

Perceptron model uses 40 features in order to get that result. Thus, with that our team has decided to pick the Logistic Regression model with 20 features where the 20 features are picked via Backward SFS, as our overall best machine learning model for this QSAR dataset. Perceptron is ruled out because it uses a greater number of features to get the same result as the Logistic regression model. Plus, using a lot of features will lead to higher usage of computational power which is not efficient. Hence, we picked the LR model. The parameter and the feature of the best model which is LR model is given below.

```
Index(['SpMax_L', 'J_Dz(e)', 'nHM', 'NssssC', 'nCb-', 'SdssC', 'HyWi_B(m)',
       'LOC', 'SM6_L', 'Mi', 'nArNO2', 'Psi_i_1d', 'B04[C-Br]', 'SdO', 'nCrt',
       'F02[C-N]', 'nHDon', 'Psi_i_A', 'nN', 'nArCOOR'],
      dtype='object')
Feature:  20
The best parameters : {'C': 10.0, 'penalty': 'l2'}
Training score : 0.8631436314363143

Mean absolute error : 0.11041009463722397
Mean squared error : 0.11041009463722397
r2 score : 0.5062305295950156
The max error value : 1

accuracy score:  0.889589905362776
confusion_matrix [[193  17]
 [ 18  89]]
              precision    recall  f1-score   support

           0       0.91      0.92      0.92       210
           1       0.84      0.83      0.84       107

    accuracy                           0.89       317
   macro avg       0.88      0.88      0.88       317
weighted avg       0.89      0.89      0.89       317
```

4.8 Fuzzy Logic Model

The fuzzy logic model is created using MATLAB software. The figure below shows the overall model.

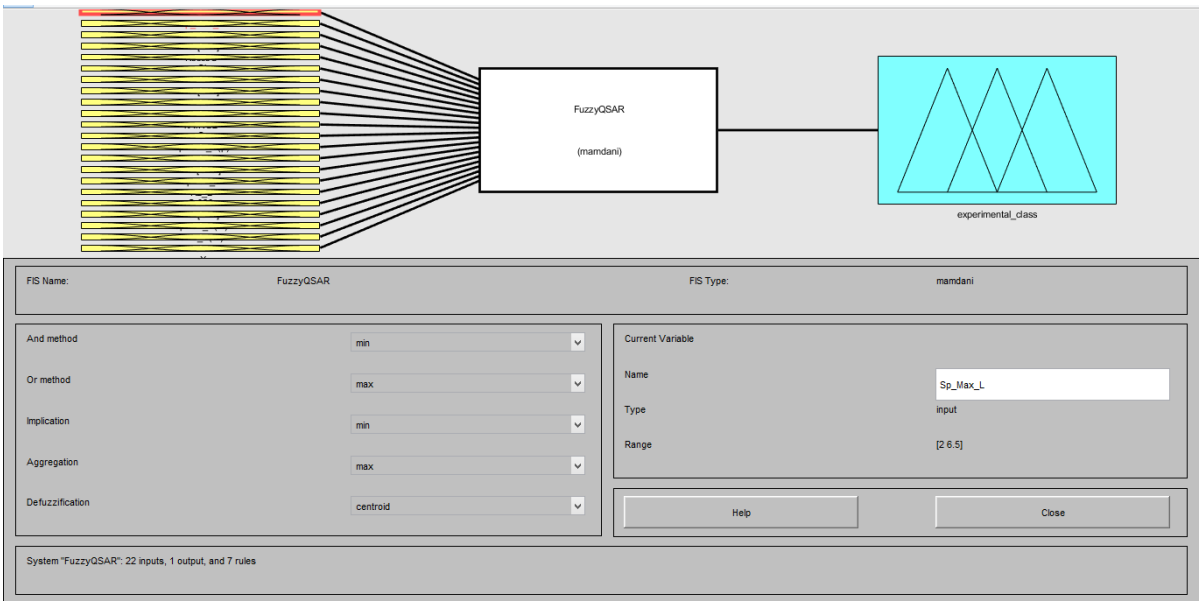


Figure 4.8.1

The first step is to identify and determine the relevant features from the 41 features available. This is because not all the features will be used in predicting the target variable. So, to select the relevant features, some background studies has been done and we have decided to only use 22 features out of 41 features. The 22 features are as follows:

- Sp_Max_L
- nHM
- F04[C-N]
- NssssC
- nCb
- n0
- F03[C-N]
- LOC
- Mi
- nArNO2
- C%
- SpPosA_B(p)
- nCIR
- B03[C-Cl]
- SpMax_A
- TI2_L
- C-026
- F02[C-N]
- SpMax_B(m)
- SM6_B(m)
- nArCOOR
- nX

Those 22 features stated above are concluded to be the features that influences the output the most in predicting the target variable. Then, for fuzzy inference system, Mamdani model has been chosen instead of Sugeno model. This is because Mamdani model is much more intuitive, and it is suitable for human inputs. Basically, we must input the data from the dataset into the model so, it is suitable to choose Mamdani model. Other than that, it also has a much more interpretable rule base and a widespread acceptance. After selecting the suitable model, the membership functions must be set up for both inputs and outputs. The range for all the inputs is set by using minimum and maximum value from the given dataset. Then, the membership functions range as shown in the example below, the range for high, medium, and low membership function is set to an equal range at first, but the result turns out to be inaccurate.

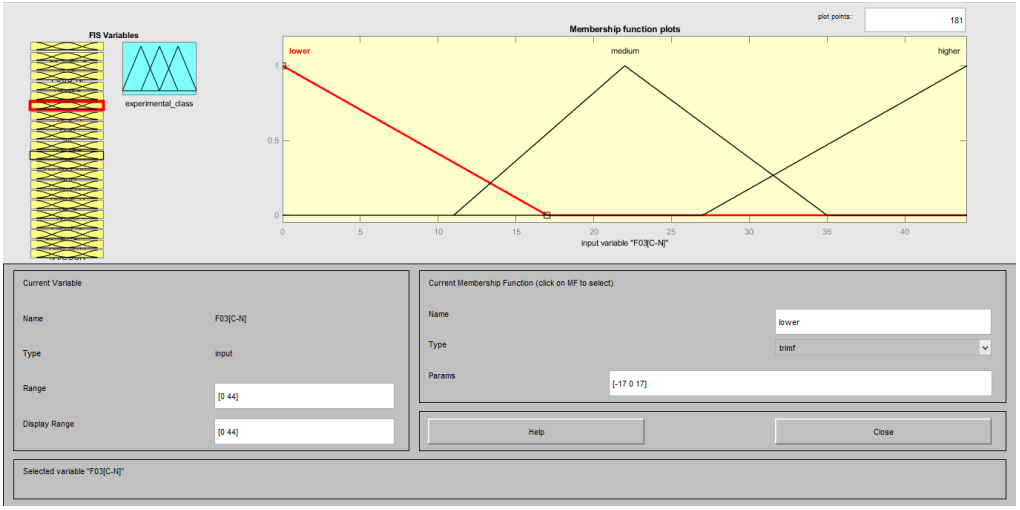


Figure 4.8.2 Fine tuning

So, the membership values have been fine-tuned after analyzing the QSAR dataset file. The type of membership function used are triangular and trapezoidal only. Furthermore, there are a total of seven rules that had been made for this fuzzy logic model. Out of these rules, two of the rules have used 'or' connection. This is because these rules are created especially for the 'NRB' class data. So, whenever a data that have been input falls on these rules, it will take the highest value for the output because of the 'or' connection. By this, the probability of getting the 'NRB' class as an overall output is high. Other rules are created using 'and' connection because data that falls on these rules might have both 'RB' and 'NRB' classes. So, it will take the lower value for the output.

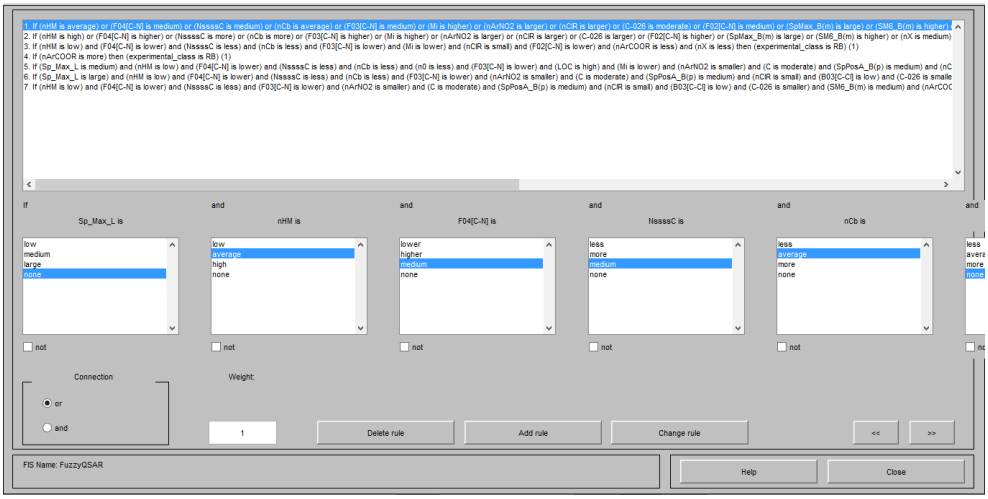


Figure 4.8.3 Rules

The rules are as follows:

- If (nHM is average) or (F04[C-N] is medium) or (NssssC is medium) or (nCb is average) or (F03[C-N] is medium) or (Mi is higher) or (nArNO2 is larger) or (nCIR is larger) or (C-026 is moderate) or (F02[C-N] is medium) or (SpMax_B(m) is large) or (SM6_B(m) is higher) or (nX is medium) then (experimental_class is NRB) (1)
- If (nHM is high) or (F04[C-N] is higher) or (NssssC is more) or (nCb is more) or (F03[C-N] is higher) or (Mi is higher) or (nArNO2 is larger) or (nCIR is larger) or (C-026 is larger) or (F02[C-N] is higher) or (SpMax_B(m) is large) or (SM6_B(m) is higher) or (nX is medium) then (experimental_class is NRB) (1)
- If (nHM is low) and (F04[C-N] is lower) and (NssssC is less) and (nCb is less) and (F03[C-N] is lower) and (Mi is lower) and (nCIR is small) and (F02[C-N] is lower) and (nArCOOR is less) and (nX is less) then (experimental_class is RB) (1)
- If (nArCOOR is more) then (experimental_class is RB) (1)
- If (Sp_Max_L is medium) and (nHM is low) and (F04[C-N] is lower) and (NssssC is less) and (nCb is less) and (n0 is less) and (F03[C-N] is lower) and (LOC is high) and (Mi is lower) and (nArNO2 is smaller) and (C is moderate) and (SpPosA_B(p) is medium) and (nCIR is small) and (B03[C-Cl] is low) and (SpMax_A is medium) and (TI2_L is high) and (C-026 is smaller) and (SpMax_B(m) is low) and (SM6_B(m) is medium) and (nArCOOR is less) and (nX is less) then (experimental_class is NRB) (1)
- If (Sp_Max_L is large) and (nHM is low) and (F04[C-N] is lower) and (NssssC is less) and (nCb is less) and (F03[C-N] is lower) and (nArNO2 is smaller) and (C is moderate) and (SpPosA_B(p) is medium) and (nCIR is small) and (B03[C-Cl] is low) and (C-026 is smaller) and (SM6_B(m) is medium) and (nArCOOR is less) and (nX is less) then (experimental_class is NRB) (1)
- If (nHM is low) and (F04[C-N] is lower) and (NssssC is less) and (F03[C-N] is lower) and (nArNO2 is smaller) and (C is moderate) and (SpPosA_B(p) is medium) and (nCIR is small) and (B03[C-Cl] is low) and (C-026 is smaller) and (SM6_B(m) is medium) and (nArCOOR is less) and (nX is less) then (experimental_class is NRB) (1)

Finally, the fuzzy logic model has been created and several data have been randomly picked from the QSAR dataset file to be tested in the model. Before testing the dataset, it must be sorted and arranged according to the inputs that have been made. If the overall output is above 0.5, it means that the data belongs to ‘RB’ class and below 0.5 means, it belongs to ‘NRB’ class.

The results are as shown below with the selected data:
i) Output for data with class of Ready Biodegradable (RB)

a) Row 2 from dataset file

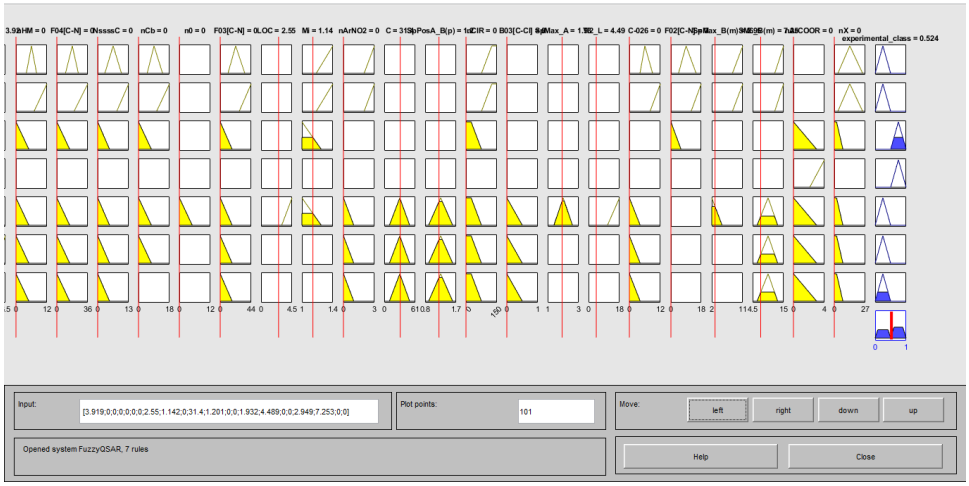


Figure 4.8.4

b) Row 21 from dataset file

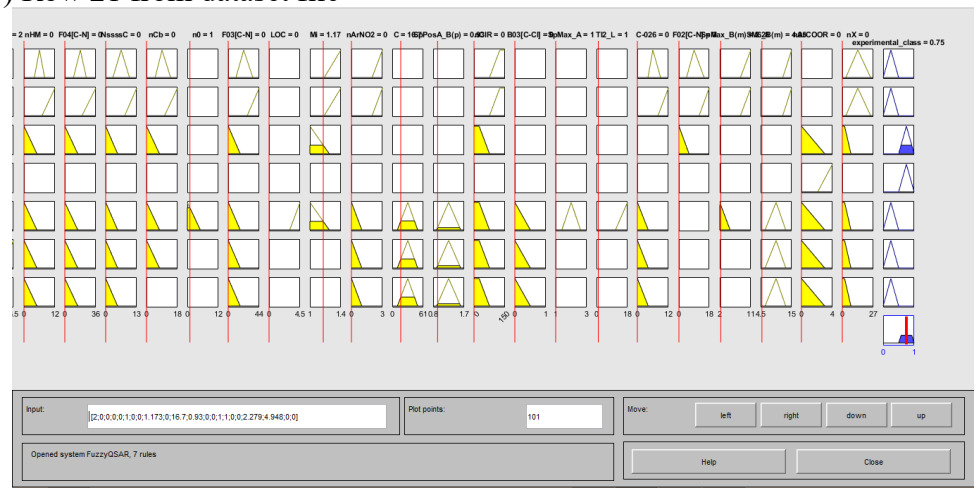


Figure 4.8.5

c) Row 154 from dataset file

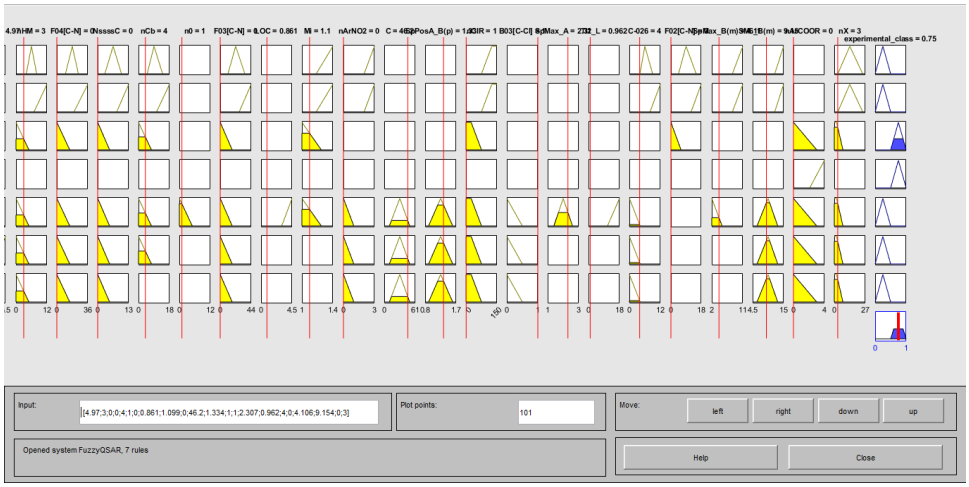


Figure 4.8.6

d) Row 842 from dataset file

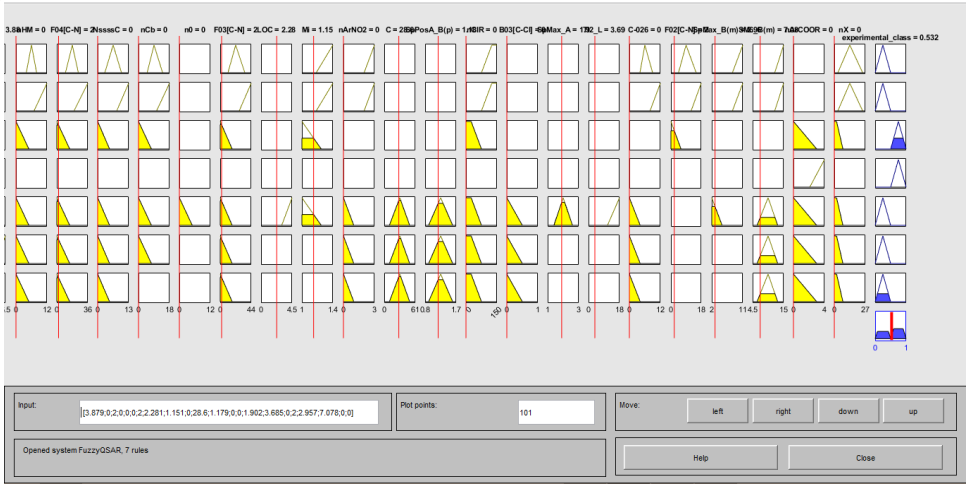


Figure 4.8.7

e) Row 909 from dataset file

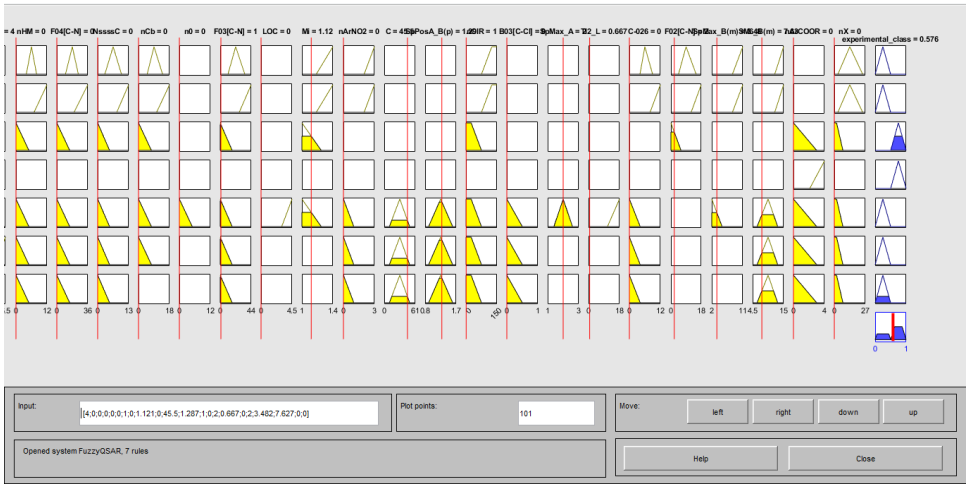


Figure 4.8.8

ii) Output for data with class of Non-Ready Biodegradable (NRB)

a) Row 596 from dataset file

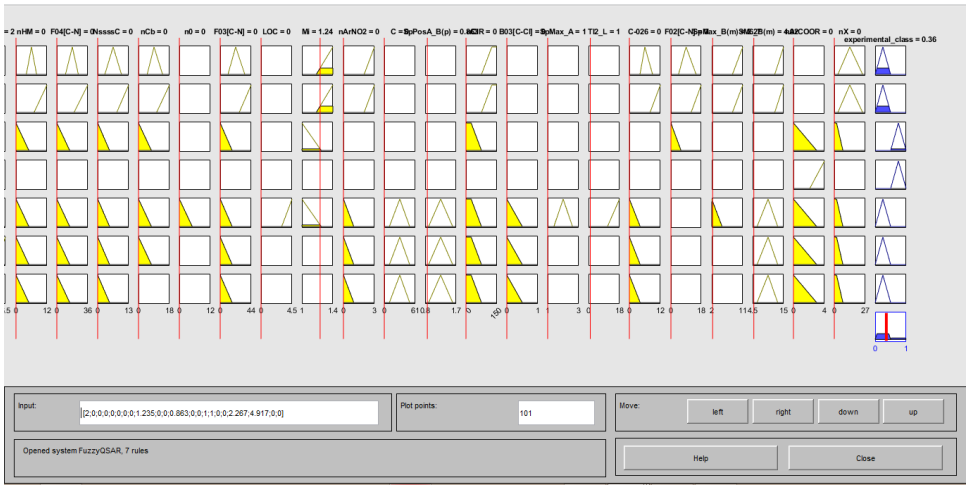


Figure 4.8.9

b) Row 615 from dataset file

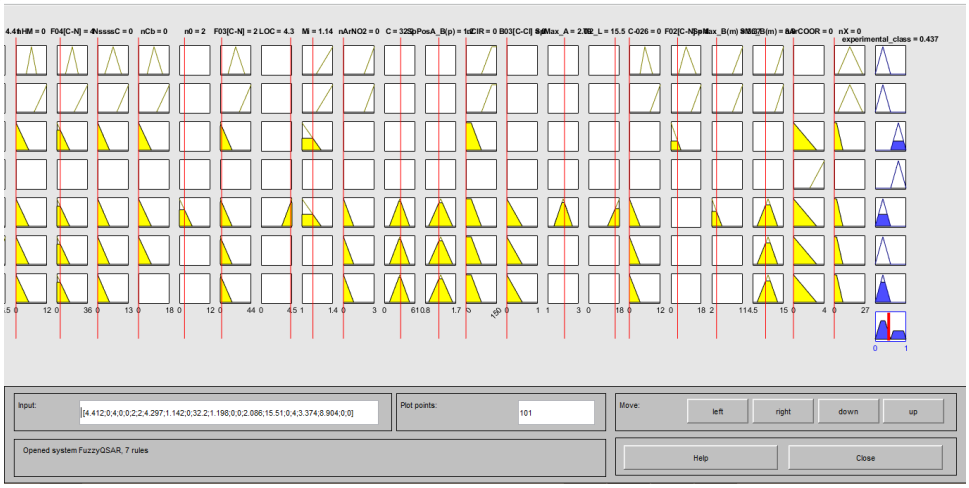


Figure 4.8.10

c) Row 802 from dataset file

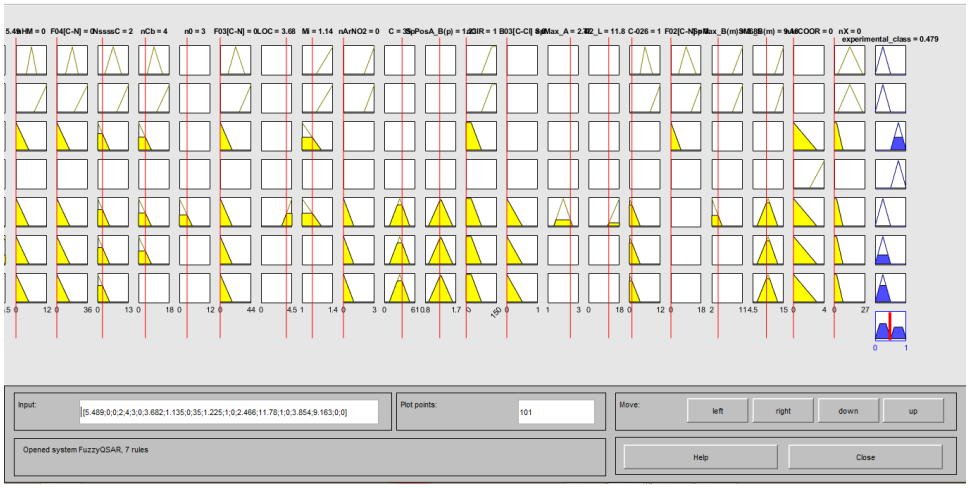


Figure 4.8.11

d) Row 1045 from dataset file

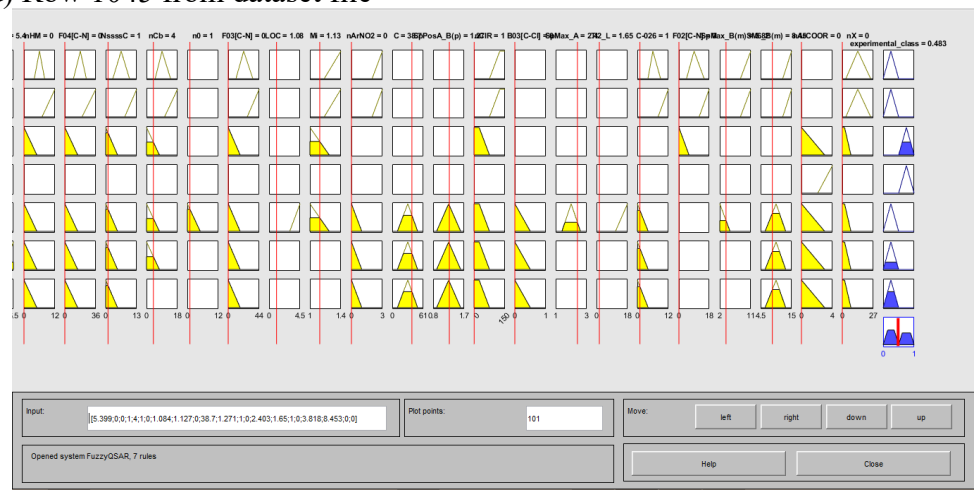


Figure 4.8.12

e) Row 1055 from dataset file

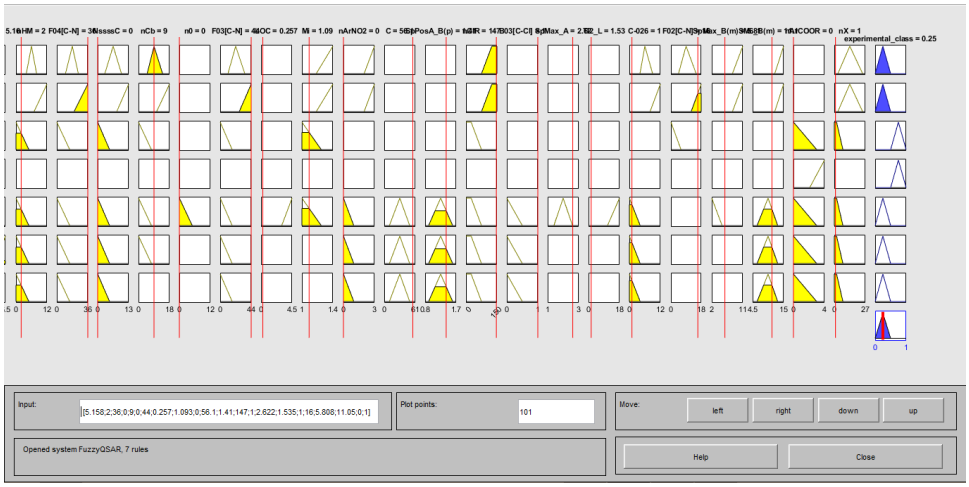


Figure 4.8.13

4.9 Comparison between Machine Learning Model and Fuzzy Logic Model

If we compare both machine learning model and fuzzy logic model, they both have pros and cons in terms of performance. Machine learning model is better at identifying patterns than fuzzy logic model. This is because machine learning model requires less human intervention since it is mostly automated. But fuzzy logic model requires a high human intervention since the inputs and rules has to be created and set up by humans.

Other than that, machine learning does takes up more computer resources and requires more time to process huge dataset compared to fuzzy logic model where fuzzy logic model uses a set of rules to make the prediction of the target variable. Besides that, fuzzy logic does not occupy a lot of space. Fuzzy logic model also is much more flexible because the rules can be modified anytime and contains a simple structure for it.

So, when its performance drops, the rules can be updated from time to time easily. But, for machine learning model, it does take more time since we have to repeat the process of feature selection and training in order to get a better result. All in all, both models are unique in their own ways at predicting the target variable. If we have to choose one from these models, then it would be the machine learning model because we can expect a higher performance in just a few trials while fuzzy logic model requires a lot of trials and errors testing because humans can make a lot of errors while machine cannot.

5.0 Conclusion

To conclude everything that has been stated, we have successfully managed to finish both tasks. We built 2 models. A machine learning model as well as a fuzzy logic model to compare the two in terms of performance, accuracy, and reliability. For the machine learning model, we implemented the feature selection method, specifically the wrapper method which utilized both types. The forward sequential feature selection and the backward sequential feature selection for SVM, Perceptron, KNN, etc.

We achieved results for each model and selected the best model out of each one. Then, we compared them in order to get the overall best machine learning model for the QSAR dataset. We then proceeded to work on the fuzzy logic model where we used the Mamdani model. We picked 22 features out of 41 features to implement the fuzzy logic model. Then, the rules are created by testing it out and fine tuning it accordingly. In short, we followed the three main steps: fuzzification, inference and defuzzification in order to create the fuzzy logic model.

5.1 Challenges

For most of this project, the biggest challenge was time. Working on those models took a lot of effort, but it also took a lot of time to run the program multiple times for every number of features for the QSAR biodegradation dataset, specifically the machine learning models where the 'LassoCV' was implemented for some of them so running the forward sequential feature selection and backward sequential feature selection took over 2 hours for each major model. Fuzzy logic model requires a lot of testing for validation of datasets and need to frequently change, add, and delete some rules. So, it requires a lot of patience and focus to create and fine tune the rules thoroughly.

5.2 Errors

As of recent, there are no known errors in the results. Regarding the programs, some errors were observed, but most were resolved.

5.3 Suggestions

In order to accurately determine which model is the best model to be implemented one a dataset, you need to take the time to use multiple methods in order to compare them and rank them in terms of performance, accuracy, reliability, and precision. Comparing the models allows you to determine the best one suited for your task/project and increases the chances of success in line with any form of work/project you are tasked with.

6.0 References

1. Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Quantitative Structure - Activity Relationship models for ready biodegradability of chemicals. *Journal of Chemical Information and Modeling*, *53*, 867–878.
2. Rocha, W. F. C., & Sheen, D. A. (2016). Classification of biodegradable materials using QSAR modelling with uncertainty estimation. *SAR and QSAR in Environmental Research*, *27*(10), 799–811. https://doi.org/10.1080/1062936x.2016.1238010
3. Verma, A. (2021, February 11). *PyTorch [Tabular] — Binary Classification - Towards Data Science*. Medium. https://towardsdatascience.com/pytorch-tabular-binary-classification-a0368da5bb89
4. *QSAR Biodegradation machine learning example*. (n.d.). Neural Designer Is a Registered Trademark of Artificial Intelligence Techniques, S.L. Retrieved May 10, 2021, from https://www.neuraldesigner.com/learning/examples/QSAR-biodegradation.
5. *QSAR Biodegradation machine learning example*. QSAR biodegradation| Example | Neural Designer. (n.d.). https://www.neuraldesigner.com/learning/examples/QSAR-biodegradation.