Critical Evaluation and Interpretation of Five Machine Learning Models on Three unrelated Datasets

1st Abhishek Sunil Padalkar
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19221576@student.ncirl.ie

***Abstract*—As with global shift towards Artificial Intelligence(AI) and Machine Learning(ML), the need to find patterns and extract knowledge out of ever increasing generation of data on daily basis is more than ever. Today, in all aspects, machine learning models can be applied to find interesting patterns and knowledge. In this project, 3 diverse topics are targeted to implement five ML algorithms, Multiple Linear Regression, Logistic Regression, Naive Bayes Classifier, Support Vector Machine, and Random Forest and are critically evaluated. The project is divided into 3 sub-objective tasks where the algorithms are compared with each other to find which method performs better on a given dataset. It is seen that Multiple linear regression doesn't perform better to predict critical temperature for superconductors with given set of inputs. For an imbalanced data with binary classification, a Logistic Regression model built on balancing the data performs better than on imbalanced data. It also out performs Naive Bayes classifier. When a multi-class classification problem is considered, Random Forest algorithm performs better than Support Vector Machine.**

***Index Terms*—Artificial Intelligence, Machine Learning, patterns, knowledge, Multiple Linear Regression, Logistic Regression, Naive Bayes Classifier, Support Vector Machine, Random Forest**

I. Introduction

With the exponential increase in the technology in the past few years, today Data holds the most value in any space as to find patterns in it for predicting those patterns in the future. Artificial Intelligence, Data Science and Machine Learning are the fields in the technology industry which is reputed the most. As the data grows exponentially with presence of social media and other devices producing data, the necessity to use this data to interpret and extract meaning out of it is now more than ever. This project thus focuses on 3 sub objective tasks with a primary aim to build machine learning models on 3 unrelated diverse datasets and to evaluate and interpret the models on the respective dataset.

Superconductors are fascinating with the magnetic and electrical properties they possess allowing today's scientific research to go further in application of it in the medical field, such MRI machine which can see inside human body. The strong feature of the superconductor is its electromagnetic properties and zero resistance at and below critical temperature. For an element to reach it's superconductivity property, it is necessary for it be below its critical temperature. Critical temperature for any element is very low temperature in Kº. Generally, to know the critical temperature of an element or a compound, it is needed to provide a decreasing temperature to an element until it shows superconducting properties. It is difficult to perform such experiments to maintain such low temperatures to hit a temperature value where the element shows superconductivity properties. Thus, in this project, the first objective is for a superconductor dataset, fit a generalised Multiple Linear Regression model and evaluate it to see if the model can predict critical temperature better of a superconductor based on the element properties.

In Finance Market, today, there is a competition between banks and various financial institutions to have more clients subscribed to their term deposit program. The term deposit is an easy way for banks to have loyal clients as well as make profits by borrowing money from the long term client and invest in the market. To save time and convert more clients into their subscription base, it is important for a bank or any financial institute to target those clients which can be mostly converted as a subscribed client to their program. It is this as the second objective of the project, that given a bank dataset, the aim is to apply, evaluate, and compare two binary classification algorithms, Logistic Regression and Naive Bayes, to predict clients which are more probable to subscribe as "yes" to their term deposit program. Both the algorithms chosen are based on probabilistic classification methods, to know which can be a better model for such prediction.

Forests play an important role in maintaining the ecological balance in the environment and nature. In environmental study, it is necessary to keep track of forest land covers and types in a certain region for to ensure that required management is in place for any changes happening over the years. Thus, the third task of the project is to classify a cover type of forest land accurately given a cartographic parameters of a land. As there are more than 2 tree types in forest areas, this is a multi-class problem for which Support Vector Machine(SVM) and Random Forest(RF) Algorithms are applied. The aim is to evaluate and compare best model within these algorithms and then to compare and contrast the best SVM and best RF model to see which performs best for the given scenario.

For Best understanding of this report, it's better to read the sections to complete one subsection for each dataset and then

move to another dataset to again read those sections. A back and forth method.

In the Methodology section, data set interpretation and transformation according to the required machine learning algorithm is discussed.

In the Implementation section, methods discussed in Methodology are implemented in diverse range of possibilities and a best model possible is chosen for further model comparison and evaluation.

In the Evaluation Section, all the implemented algorithms are evaluated and compared with each other based on the performance metrics and the target objective.

A better understanding can achieved if each of the three objectives , in the below sections, is read individually first completely and then moved to the next objective.

II. Related Work

A study performed by Al-Mukhtar and Mustafa, to predict the suspendend sedimentary element in the water bodies included three different machine learning models, Random Forest, Support Vector Machine(using radial basis function), and Neural Networks. The outcome of the study concluded that Random Forest algorithm outperformed both the other algorithms [1].

Another study, by Zhang, Jin, et. al, was performed to predict CO2 minimum miscibility pressure (MMP) with a given set of inputs such as mole fraction, Tcm, Tr, MWC5, and Vol/Int using three set of algorithms which are Random Forests, Back Propogation Neural Network, and Support Vector Machine. The results for the aim of the study achieved the target by all these methods but the best possible results provided was by Random Forest Algorithm [2].

According to Roy, Sanjiban Sekhar, et. al, intrusion detection systems cannot always detect the constantly changing malware and intrusion activities in a computer network. Thus, to predict an activity they implemented and compared several machine learning algorithms such as Random Forest, Decision Trees, Support Vector Machine and Nearest Centroid Classification. The study concluded that Random Forest outperformed Support Vector Machine classification and other methods based on accuracy and detection rate [3]. These studies indicates that using Random forests for classification purpose provides best possible outcome.

Chemchem, Amine, et. al, performed a study with combination of SMOTE Sampling with Random Forest Algorithm to tackle the issue of poor accuracy for highly imbalanced dataset. It shows that combining this sampling technique improves the accuracy on test data a significant 90% accuracy [4]. This ensures that sampling technique such as SMOTE is necessary for performing a binary classification problem with highly imbalanced data. Thus, for this project purpose, the imbalanced bank dataset trained model is also compared with a model trained on balanced dataset derived from the imbalance dataset.

N. Zurbuchen, et. al, performed a comparison of various machine learning algorithms based on evaluation metrics of Specificity, Sensitivity, ROC, and accuracy showed that gradient boosting algorithm outperformed traditional SVM, Decision Tree and Random forest algorithm. But, the results of Random Forest were close to the gradient boosting method in the Fall Detection System [5].

Another study, by M. Khoshlessan, et. al, compared five ML methods to apply on micro-grid energy management. The comparison was done on the basis of accuracy and it was found that Random Forest with polynomial features performs best for energy management data [6].

Dinesh and Jasmine performed a study where an optimal Support Vector Machine Model was built by tuning the hyperparameter combined with feature selection on image classification of Lung Cancer. It performed models for 20 different parameter set to compare and select the best parameters [7]. Thus, a tuning of hyperparameters is considered for our objective task for Support Vector Machine to produce best possible outcome.

III. Methodology

The sub-objectives of this project follows a Knowledge Discovery in Databases(KDD) Methodology. First, a target data is selected based on our aim for mining and discovery. Then the data is first explored using visualization and performing descriptive analytics to understand the data. Next, according to the machine learning method selected and data understanding, data is pre-processed and transformed to then inputted into the machine learning method to find patterns in it. These models are then evaluated and interpreted to extract knowledge at the final step. The implementation of this methodology for each objective task is discussed below:

A. Superconductors

As a Pre-KDD step, a basic level of understanding is gained and first discussed in the previous sections Introduction and Related Work. Next, dataset is targeted for our research question whether Linear Regression models work better for superconductor dataset. The superconductor dataset collected for this purpose, consists of 9 main properties of element as attributes such as Number of Elements, Atomic Radius, Valence, Atomic Mass, Fusion Heat, First Ionization Energy(fie), Thermal Conductivity, Density, and Electron Affinity. In addition to these variables the dataset consists of other attributes which are statistical measures such as mean, standard deviation, geometric mean, entropy and range of the main features. Also, the data consists of a “wtd” attribute for all these above attributes. The entire dataset consists of 21263 superconductor observations and total of 82 columns. As the aim is to analyze if a linear regression model can better predict the critical temperature of a superconductor based on its properties, it is also necessary to build a generalized model. Thus the dataset is cleaned and transformed according to Ordinary Least Square method requirement. First, all the attributes are numeric in nature is checked. Second, missing values and duplicates are checked in the data. With no missing values, 66 duplicates are removed to move to next step. The

data consists of “wtd” attributes as linear factors of all the other features deriving multicollinearity in the dataset. These “wtd” attributes were dropped to then have 41 features in total with critical temperature variable. As seen in Fig. 1, these 41 variables also demonstrate high correlations between them. Also, 41 variables are still high number of variables increasing complexity in the dataset.

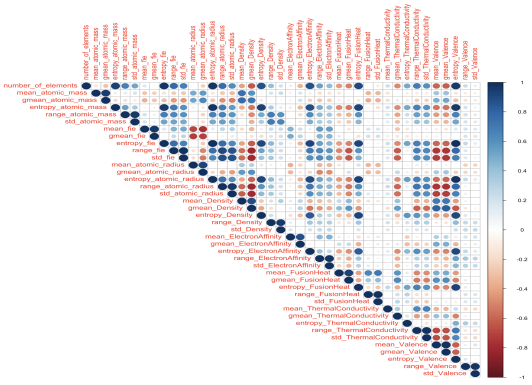


Fig. 1: Correlation Visual plot between variables

Multiple linear regression method has one assumption, absence of multicollinearity between predictors, for the model to be generalised. As high correlation is seen, it violates this assumption. Thus, these selected features are transformed as required for modelling process using Principal Component Analysis technique with axis rotation. Axis rotation is used to have a better understanding of relationship of each component with the response variable. Scree Plot in fig. 2, using Eigen Value principal for component selection indicate 8 components with eigen values >1. Fig. 3 present that collectively 89% of total variability is explained by these 8 components of the dataset. As 89% is reasonable variability to work with, we chose these components for the model building. Linear regression model building and implementation is to achieve a generalized model for this dataset is covered in the following section.

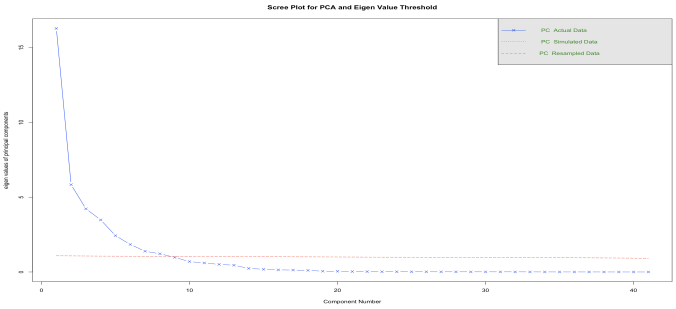


Fig. 2: Scree Plot for Component Selection

	RC1	RC6	RC2	RC3	RC4	RC5	RC8	RC7
SS loadings	9.482	8.362	4.704	3.985	3.735	2.319	2.218	1.882
Proportion Var	0.231	0.204	0.115	0.097	0.091	0.057	0.054	0.046
Cumulative Var	0.231	0.435	0.550	0.647	0.738	0.795	0.849	0.895

Fig. 3: Loadings of the 8 Princpal Components

B. Bank Subscription

As a Pre-KDD step, it is understood that for Banks it is necessary to target the audience which will convert into a subscription for their given program. It means, the focus is to get more correct ”yes” predictions than ”no”. The Bank Dataset present for analysis and prediction consists of 21 attributes in total, where the last one is our dependent variable, i.e a person subscribed “yes” or “no”. Thus, for this binary classification, we chose to model two methods which are Logistic Regression and Naive Bayes. As data selection step for both the algorithms, missing values and duplicates are checked. The dataset consists of 0 missing values and 12 duplicates which are removed. A “duration” variable which is duration of call a client was to discuss about subscription process is removed as per the dataset instructions. It is not considered because this variable is a biased variable which gives very high accuracy. With this selected data, first a visual is plotted to check the balance of ”yes” and ”no” counts in the dataset. Fig. 4 demonstrate a clear imbalance in the dataset with very high ”no” values which is noted for further modeling process. For both the algorithms, continuous “Age” variable is converted into categories of 4 age groups, [15 - 30], [31 - 50], [51 - 70] and [71 - 100]. This transformation is made to increase the performance of the algorithms. This transformed data is used for Naive Bayes algorithm. Where as for Logistic regression, a categorical variable, “client_contacted_previously”, was created based on another numeric variable, “pdays”. Variable “pdays” stores a value “999” for those clients who were not contacted previously of a campaign. And, other values are numbers denoting number of days past, a client was contacted previously. Thus, on the basis of assuming that this information may be useful, this variable is created. Based on these two sets of data for each algorithm, Naive Bayes and Logistic Regression Models are implemented and discussed in the next section.

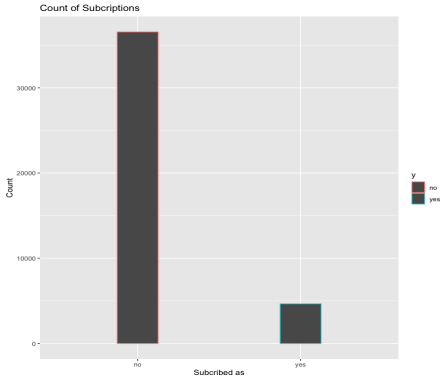


Fig. 4: Check for balance in response variable

C. Forest Land Tree types

As our initial step of Pre-KDD, the aim is to perform an extensive number of models to select a best predicting model to predict a tree type in a forest land with given cartographic features. For the selection process, first missing

values and duplicates are checked which found out to be zero. The dataset consists a total of 581012 observations and 55 features of 7 different tree types in 4 different regions in “Roosevelt National Forest” in Colorado. Fig. 5 shows the distribution of 7 tree type covers in the total Forest regions. Tree type 2, Lodgepole Pine, and 1, Spruce/Fir , are seen to be highest in the given forest land. Cover type 4, Cottonwood/Willow is seen the lowest, being a unique type of tree cover. Other variables in the dataset are: Elevation, Aspect, Slope, Horizontal Distance To Hydrology, Vertical Distance To Hydrology, Horizontal Distance To Roadway, Hillshade 9am, Hillshade Noon, Hillshade 3pm, Horizontal Distance To Fire Points, Wilderness Area, and Soil Type. There are 4 wilderness Area and 40 Soil types in the entire forest land. Soil type and Wilderness Area are split into binary categorical variables where as other are numeric variables. The dataset considered is

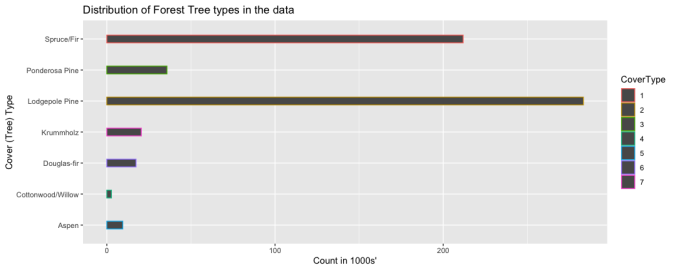


Fig. 5: Distribution of Forest Cover Types

huge, thus, for the purpose of this project it is downsampled to 10000 rows by making sure the ratios of the covertypes remain almost same. As there are 7 Cover types, this is multi-class classification problem. To achieve our target of best possible predicting model, Support Vector Machine and Random Forest machine learning(ML) method are selected for classification purpose. For each of these methods, data is transformed and utilized differently.

1) Random Forests

The aim is to build a large set of Random Forest(RF) models to evaluate the output and select a best model for the prediction. Random forests properties are also evaluated in the process of building these many models. Thus, this high dimension dataset of 55 variables is first converted to 13 variables by combining 4 binary categorical variables of Wilderness area to one with 4 categories and 40 binary categorical variables of Soil Type to one with 40 categories. The dimension of the dataset is thus reduced for faster model building purpose.

2) Support Vector Machines

Support Vector Machine(SVM) is a machine learning algorithm which classify a response based on a best possible hyperplane created with the help of support vector margins. This algorithm classifies a data point in two categories if it is above 1 or below -1 with the hyperplane lying on 0. Thus, the binary variables in the datasets are converted to 1 and -1, where 0 becomes -1. Because of this property of SVM, all the continuous variables which have different scales are rescaled to a range of [-1, 1] using minimum and maximum value of the variable. This transformation is made for svm model building to have correct and better predictions.

The model building and evaluation for SVM and RF is discussed in the next section.

IV. Implementation

Here, we implement each model for the respective dataset to find out the best model using parameter tuning. Each model is first tuned to be the best. In the evaluation and comparison section, we’ll discuss which model works best.

A. Superconductors

In this subsection, implementation of a generalised linear model is discussed. The aim for the research is to fit a linear regression model and check if a generalized model can be a good predictor for the superconductor data. But, if a curvy relationship is found between any predictors and response variables it is noted to add complexity by including polynomial terms for the respective predictor. As discussed in the methodology, the dataset is transformed into rotated principal components for model building and data mining. Before implementation, a correlation plot is checked to see the new relationship between our variables. Fig. 6, indicates that RC1 has an upward and RC5 has a downward linear relationship with critical temperature. RC6 also shows a linear relationship to an extent with a small downward curve in the end. Whereas, the other components show a non-linear relationship. It is seen that RC2 has very less correlation with critical temperature. From the skewness in the distribution in all the variables, we first see that critical temperature is strongly positively skewed. This indicates most superconductors have a very low critical temperature. RC1, RC6, and RC7 show negatively skewed whereas RC5 show a positively skewed distribution. The rest variables reveal to a point normal distribution with mean around 0. Thus, it is expected that RC1, RC6 and RC5 will contribute more to the linear model than rest variables and RC2 might hardly be useful in predicting critical temperature. The data is split into train and test set with 75:25 ratio. Next, a linear model is implemented with all the 8 components as predictors. The results show an adjusted R^2 of 0.548 and RMSE measure score 23.02 on the train set. 4 sets of residual plots are checked to see the fit of the model. Fig. 7 show that residuals are randomly spread only to an extent and scale-location graph indicates a presence of heteroscadasticity at the start. This means the model is violating the assumptions of an ordinary least square method.

To rectify the residual distribution, we first check residual relationship with each of the components of the model. From fig. 8, some curvy relationship is observed in RC5, RC6, RC7, and RC8. Also, for RC1 and RC4 a small curve is present at the end and start respectively. Thus, a polynomial term of these above components is added to fit a new model. Adjusted R^2 0.559 and RMSE measure of 22.63 is noted.

By adding these polynomial terms, we increased the complexity, but not all polynomial terms were significant. Thus, by using best subset regression, set of significant variables

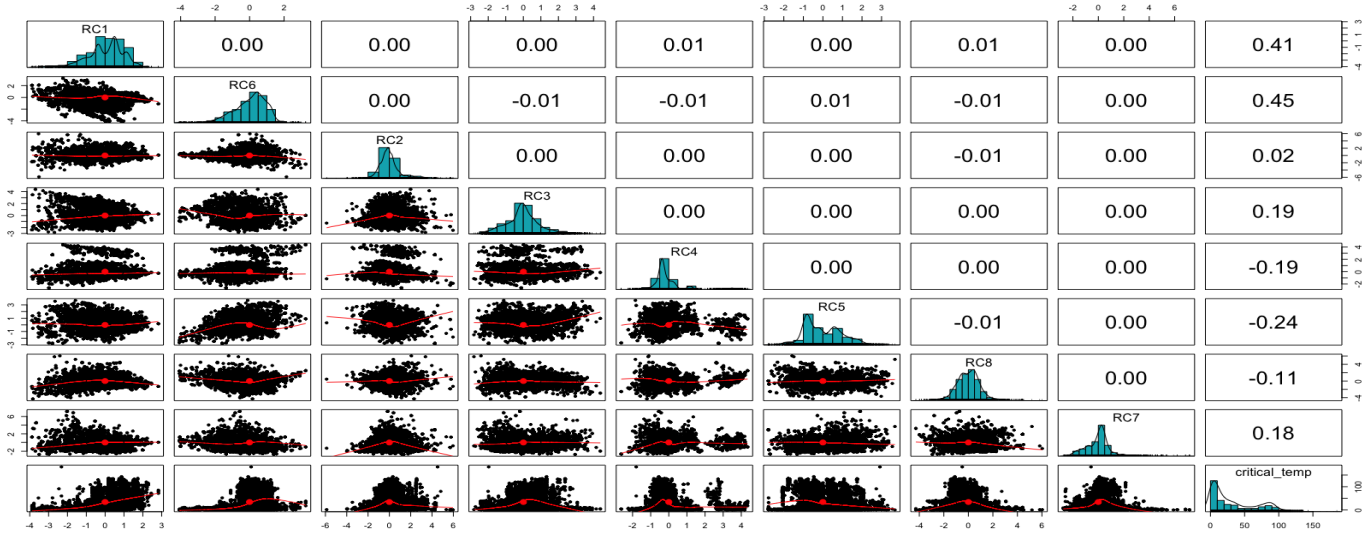


Fig. 6: Correlation plot of the Components with Critical temperature

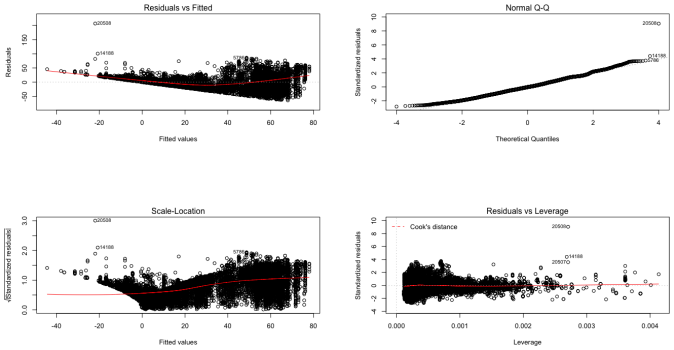


Fig. 7: Residual Plots of Model with all components

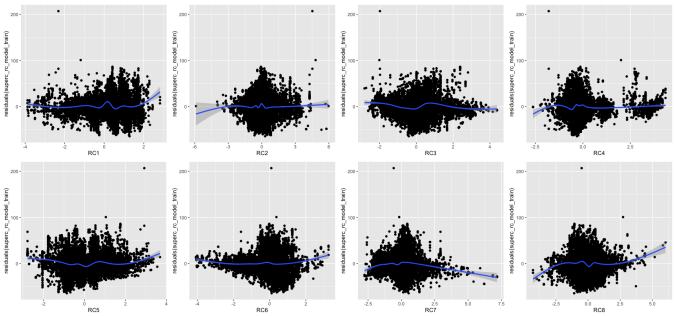


Fig. 8: Component vs Residual Relationship

with lowest AIC and highest R^2 is chosen. It was seen that only RC7 polynomial term is significant and other polynomials were insignificant. The best combination of predictor variables also didn't include RC2. Thus, a new model is built by removing RC2 and polynomial terms except for RC7 polynomial. Adjusted R^2 and RMSE of this model is 0.554 and 22.73 respectively. By removing unnecessary terms and having results closer to the previous model with more variables indicate that we are also following principle of parsimony. Again the plot is checked for this new model, which indicated that residuals randomness improved to a very little extent. The scale-location graph for this model still shows heteroscadasticy presence in the model. To rectify that a new is built now with taking square root of our dependent variable. A significant drop is observed in the RMSE measure with value only 1.834 and an increase in adjusted R^2 with 0.632 score. The fig. 9 show the residual plot of this model indicating that all the assumptions of the linear method is followed. The residual randomness is only to an extent random. The resiudals are normally distributed and there is no influential outlier in the data. For assumption tests, Durbin Watson(DW) test for no autocorrelation, Variance Inflation Factor(VIF) for no multicollinearity and Non-constant Variance Test(NCV) for no heteroscadasticity is checked for this model. DW statistic came out to be 2.01 indicating absence of autocorrelation, and VIF for all the variables are around 1 implying no correlation between the predictors. VIF metrics were expected as above since PCA was performed which produces components with no correlation between each other. The NCV test came out to be significant indicating presence of heteroscadasticity, but the residual plot show a random spread in scale-location graph. It is assumed that because of large size of the data the NCV test is significant even though the plot indicates heteroscadasticity is rectified to most extent. The generalised linear model is thus implemented for the superconductor data, but the performance measure of it indicates a bad fitted model. The interpretation and evaluation is discussed in the next section.

B. Bank Subscription

As discussed in the previous section, the bank dataset is transformed for each algorithm as required. In this section, number of iterations performed for each of the algorithm to achieve higher performance based on metrics such as AIC, Residual Deviance, Detection Rate/Precision, Sensitivity/Recall, Accuracy and Cohen's Kappa is presented. The selection of best model for each algorithm is discussed below:

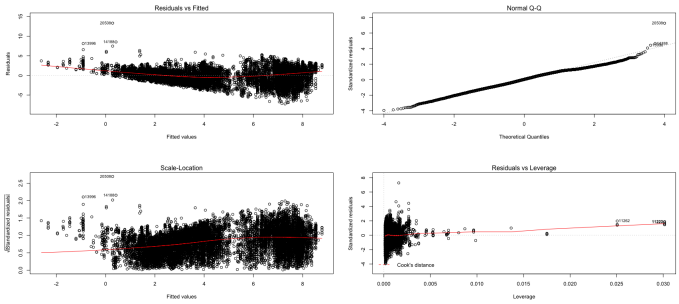


Fig. 9: Residual plot of generalised Linear Model

1) Logistic Regression Model

For logistic regression modeling, the process was divided into two sub parts. From the imbalanced dataset, a balanced dataset was created by random sampling of "no"s to match the number of "yes"es in the dataset. The total rows in the balanced dataset is 9278. The Logistic models were built on these two datasets to understand if a model does better on a balanced or an imbalanced dataset or performs equally well on both.

I. Imbalanced Bank Dataset:

The dataset is first split into train and test set with 75:25 ratio. Ratio for number of nos and yeses in both the train and test set are checked to be approximately same. This implies that model can be correctly trained and tested. First, all variables are selected and a logistic regression model is built. The AIC metric score for this model is 17052, Residual Deviance is 16942 and the Null Deviance is 21656. The summary of this model shows a total of 8 insignificant variables for our desired prediction. Thus, a step-by-step process is followed to remove insignificant variables with highest p-values first. We don't remove all variables at once because one variable might turn significant after removal of another variable. With total of 7 iterations of insignificant variables removal, a logistic model with all significant variables if built. The AIC and Residual deviance of the final model is 17036 and 16966. Thus, this model has a lower AIC and the deviance is increased just by 24. This is the lowest AIC possible for logistic model given this dataset as by further removal of features increased the AIC and Residual deviance by a considerable score. Thus, for imbalanced dataset, this model is finalised as best model. This model is first tested on the test set created. The performance metrics for a "yes" prediction are: Accuracy = 89.74%, Cohen's Kappa = 0.302, Sensitivity = 23.5%, and Detection Rate = 2.7%. A low sensitivity, detection rate is noted. In this imbalanced dataset, a null accuracy is calculated which has a value of 88.5%. By comparing null and model accuracy, it is seen that this model hardly performs any better than just a random null model.

II. Balanced Bank Dataset:

In this part, same procedure is followed as Imbalanced dataset model building with a balanced dataset. First, all the variables are included to build a logistic regression model. For this model, the AIC value is 7534.5, Residual Deviance is 7424.5 and Null Deviance is 9645.7. After step-by-step removal of insignificant variables, the final model for balanced dataset consists of 11 independent variables. The AIC score is 7507.4 and the Residual Deviance is 7437.4. Again, further removal of variable lead to increase in AIC and Residual Deviance, thus this model is finalized for the balanced dataset. This model is tested on the test set created to give following performance scores for a "yes" prediction: Accuracy = 72.8%, Cohen's Kappa = 0.4573, Sensitivity = 62.32%, and Detection Rate = 31.51%. A higher Kappa value, sensitivity and detection rate is noted. Here, the null accuracy is 50% as the dataset is perfectly balanced. By comparing null accuracy with model accuracy, a significant improvement of 22.8% is noted.

These two final models, on imbalanced and balanced dataset, are then also tested on entire bank dataset to finalize the best performing logistic regression model. The table I displays the comparison between these two logistic models for the chosen performance metrics as follows:

TABLE I: Comparison between Logistic models on Bank Data

Performance Metrics	Imbalanced Data Model	Balanced Data Model
Accuracy	90%	81.9%
Cohen's Kappa	0.3031	0.3472
Sensitivity	23.23%	63.97%
Detection Rate	2.6%	7.2%

Even if the accuracy of the model built on the imbalanced dataset is higher than the model built on balanced dataset, the other metrics clearly indicate that model built on the balanced dataset is a better predictor for "yes" as our target goal. The Sensitivity is reasonably high, but the Cohen's Kappa, 0.34, indicates a fair level of agreement. This Kappa measure is context specific, and thus in this case, it is assumed that this value is a good score for prediction of "yes" for bank subscription. Thus, the model built on balanced dataset is selected as best logistic model for further comparison.

2) Naïve Bayes Model

The Naïve Bayes(NB) method classifies the target variable based on simple probabilistic value of chance of an event happening given a set of events. Thus, all the variables are used in modeling a best Naïve Bayes classifier for the bank data. First, the data is split into train and test set with 75:25 ratio. For selection of a good NB classifier, 4 different NB classifiers are built. First model is a simple naive model with all the variables. Second and Third models are built by using resampling methods 10-fold cross validation and 25 reps boostrapping respectively. And, for fourth model, only a set of categorical variables is selected for categorical variable interpretation. The table II displays the performance of these four built models on the test data created.

TABLE II: Comparison between 4 NB classifiers

Performance Metrics	Simple NB	10-fold CV NB	Bootstrap NB	Categorical NB
Accuracy	83.6%	87.58%	87.58%	87.83%
Cohen's Kappa	0.3234	0.368	0.368	0.2767
Sensitivity	51.59%	42.96%	42.96%	27.86%
Detection Rate	5.8%	4.8%	4.8%	3.1%

It is seen that, both the re-sampling techniques produces same results. Considering overall accuracy, the models built

using resampling techniques and the categorical model performs better than the simple naive model. Cohen's Kappa for simple naive and resampling models are above 30 where as for categorical NB is just 0.27. Even though all the models are considered as having fair level of agreement, it is clear that categorical NB performs bad considering a random null model. But, when Sensitivity and Detection Rate is considered, a Simple NB performs reasonably well in predicting correct "yes" subscriptions than the rest of the models. Thus, as our target to achieve higher detection rate and sensitivity, a simple Naïve Bayes classifier is selected as best with the given scenario for bank data.

Further evaluation and comparison between best Logistic Regression and Naive Bayes Model is discussed in the next section.

C. Forest Land Tree types

As discussed in the previous section, all the data transformation is done to perform SVM and RF modeling. In this section, the aim is to select a best possible model to predict the forest cover type for both SVM and RF by tuning the hyper-parameters for each of the model. By building large set of models, the methods properties are also evaluated. The selection of best possible model is based on following performance metrics: Accuracy, Out-of-Bag(OOB) Error rate. For each method, the selection of best model is discussed below:

1) Random Forests

This method uses a bagging technique to classify an outcome by considering different number of variables and different number of trees. Because of this feature, the algorithm never has high variance or a problem of overfitting the dataset for more complexity. As we aim to get a best possible predicting model, a naive approach is taken to build 3300 models in total and select the tuning hyperparameters which result in least OOB error. For all 11 the variables, models are built in increasing number of trees till 300 which are considered in the forest. The maximum number of trees chosen to be 300 as it is assumed to be enough to achieve our aim. As bagging technique uses only 2/3rd of the training set and 1/3rd is left for testing, the model selected is from the train set having least OOB error. Fig. 10 demontrate a clear picture of random forest algorithm having no effect of high variance. As seen, after considering around 100-150 trees in the forest, for all different number of variables considered for each decision tree, the OOB Error fluctuates over a constant mean. It is seen that by just including 5 to 6 variables out of 12 per decision tree in the forest having around 100-150 trees gives optimal result. From the fig. 10, models having number of variables >3 in the decision trees gives a very close error rate around 19%-20%. Thus, even if by choosing any decision tree for number of features >3 and number of trees >100 will give around same result, we select the one which performed the best in the given scenario. The model with number of features 6 and 199 trees in the forest has lowest OOB Error of 19.52%. Thus, this model is selected as the best model for our purpose being

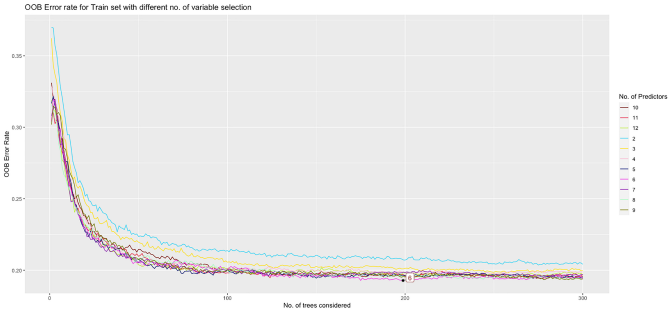


Fig. 10: Error Rate of Random Forest models for 300 trees and all number of variables

cognizant of the fact that others will also perform similarly well. For this model, fig. 11 display elevation to be most important factor in making a decision for the classification. Second is the Soil Type factor. Thus, for tree type prediction, these two factors play important role.

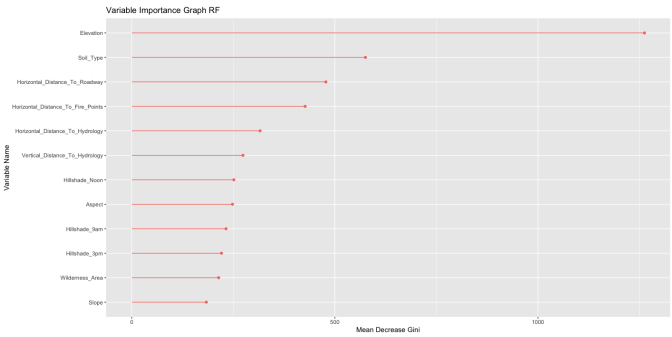


Fig. 11: Variable Importance

2) Support Vector Machine

First, we use two datasets, one with updated dataset and one with raw dataset. Second, 4 models are evaluated on thes 2 datasets to check which kernel, radial or linear, suits the best on which dataset. Based on different relationship in the data, different kernels in the SVM fit the model differently. Thus, this important criteria is checked before tuning the hyperparameters. Table III shows the performance metrics of these 4 models built. The best model, Radial kernel model on transformed data, is chosed based on Cohen's Kappa and Mean AUC metric. The accuracy for this model is also highest of 76.52%.

TABLE III: Comparison between SVM Models

Performance metrics	Raw Data		Transformed Data	
	Linear	Radial	Linear	Radial
Accuracy	66%	50.24%	72.68%	76.52%
Cohen's Kappa	0.0.4322	0	0.5458	0.6129
Mean AUC	0.6656	0.5	0.7209	0.7746

Thus, further analysis is performed by using radial kernel and transformed data. The current Radial SVM on transformed data model was built on gamma = 0.1 and cost = 10. Gamma parameter indicates how many support vectors are considered to fit a hyperplane to classify two categories where as Cost parameter considers how many data points can be allowed to

cross the margin created by support vectors. Higher gamma values means only fewer data points close to the true classification line is considered as support vectors. By selecting higher gamma values, an SVM model tends to overfit the data losing the generalizing ability. Thus, for model tuning process, lower gamma values as 0.01, 0.05, 0.1, 0.25, and 0.5 is chosen with cost ranging from 1 to 100. Cost range is selected in this range as 100 is assumed to be a reasonable value for number of data points crossing the margin. Fig. 12 displays which gamma values performs better for what cost value. It is clear that by increasing gamma above 0.1 for cost value above around 15 fits the data poorly. For cost values 1 to around 12, gamma values 0.05 and 0.1 performs better than the rest gamma values. By increasing the cost value further, it is seen that gamma = 0.1 starts decreasing in accuracy. For gamma values 0.01 and 0.05, the accuracy fluctuates around the same mean for respective gamma. This indicates that considering more data points far from the true classification line as support vectors increases the accuracy of better fitting hyperplane. And for higher support vectors, even after considering higher cost, the SVM model performs roughly the same. Next, for selecting best model, model with highest accuracy is selected. It is seen that gamma = 0.05 gives highest accuracy for 2 cost values. A smaller cost value is selected for this instance. The model with gamma = 0.05 and cost 61 is thus selected. The Kappa measure and AUC for this model are 0.621 and 0.7934 respectively. Thus, by comparing the best model selected from the table III, the new best model has 0.44% higher accuracy, 0.0081 higher Kappa value, and 0.0188 higher AUC. This indicates that hardly any difference is made by reducing the gamma to 0.05 from 0.1 and increasing cost from 10 to 61. Thus, the first model can also be used for the classification purpose. But, as discussed, lower gamma gives a better fit to the data, we select the model with gamma 0.05 as our best model for our classification target.

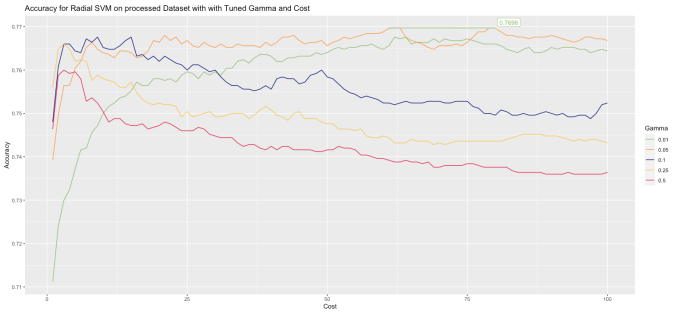


Fig. 12: SVM models with 5 Gamma over Cost 1:100 values

These two best selected, RF and SVM, models are further evaluated and compared in the next section.

V. Evaluation

In this section, the implemented models in the previous sections are evaluated and compared in the sub-project objective manner.

A. Superconductors Linear Model

As discussed in the previous section, a generalised linear model is implemented for the superconductor data. Fig. 13 displays the summary of the final model. In fig. 13a, the adjusted R^2 and RMSE values for the model is 0.632 and 1.834 respectively. In fig. 13b, we see that variable importance all the variables have score above 20 for prediction of critical temperature. It is seen that RC6 and RC1 have highest importance in predicting the value. RC1 is evaluated as to be an entropy statistic of all the variables and RC6 is a combination of density, atomic radius, valence, atomic mass, thermal conductivity and fie with more loading weight-age on density, atomic radius, fie and valence of a superconductor. Thus, these measure can be considered as important predictors for critical temperature. While building this model, a series

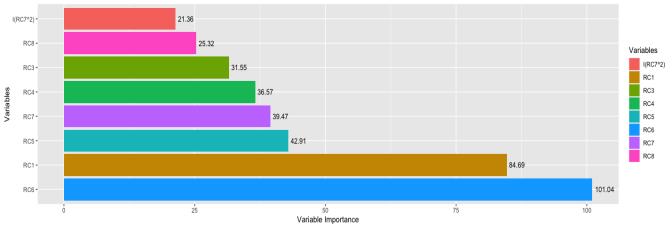
```
Call:
lm(formula = sqrt(critical_temp) ~ RC1 + RC6 + RC3 + RC4 + RC5 +
    RC8 + RC7 + I(RC7^2), data = rc_train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2538 -1.2913  0.0715  1.4050 13.3180

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.161520   0.016140  319.79   <2e-16 ***
RC1          1.289135   0.015223   84.69   <2e-16 ***
RC6          1.488978   0.014737  101.04   <2e-16 ***
RC3          0.460413   0.014594   31.55   <2e-16 ***
RC4         -0.534661   0.014618  -36.57   <2e-16 ***
RC5         -0.635917   0.014819  -42.91   <2e-16 ***
RC8         -0.366678   0.014481  -25.32   <2e-16 ***
RC7          0.611179   0.015486   39.47   <2e-16 ***
I(RC7^2)    -0.149348   0.006992  -21.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.834 on 15889 degrees of freedom
Multiple R-squared:  0.633,     Adjusted R-squared:  0.6328
F-statistic:  3426 on 8 and 15889 DF,  p-value: < 2.2e-16
```

(a) Model metric values



(b) Variable importance for prediction of critical temperature

Fig. 13: Generalized Linear Model Summary

of models were also built and evaluated using a test set as discussed in previous section. Fig. 14 shows that as complexity of the model is increased in the process of generalizing the model, the fit of the model became poor for the test data and overfitted the train set. The adjusted R^2 score increased by a mere 1% at the start and then surged by 7% as the complexity of the model increased. This is a clear indication that the critical temperature of a superconductor don't have a significant linear relationship with these component factors and the a generalised linear model cannot be used as prediction for the same.

B. Bank Subscription Models

As discussed in the previous section, given this scenario, best Logistic Regression model and Naïve Bayes Classifier model is selected based on Sensitivity measure and detection rate with reasonsable Accuracy and Cohen's Kappa. In this section, these 2 best models are compared and evaluated based on the above metrics and also using ROC curve and Area under

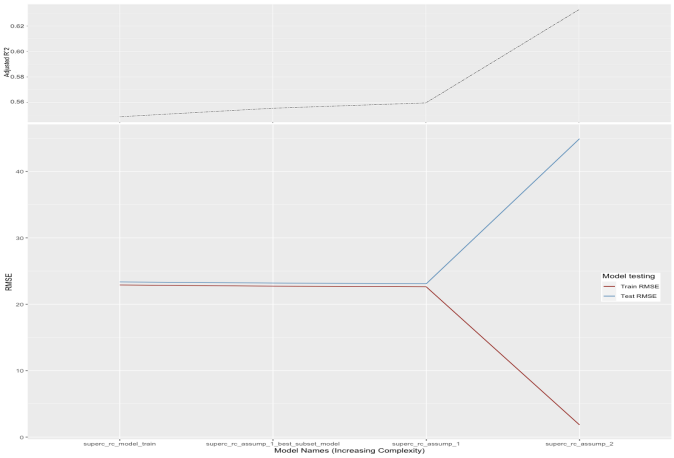


Fig. 14: Bias and Variance Interpretation

the curve(AUC) metrics to find which model performs best for the selected scenario on the bank data. First, from table I and II, Logistic Regression model built on balanced data and Simple Naïve Bayes models are compared. The accuracy of Simple NB is higher by merely 1.7% whereas all the other metrics are higher for Logistic Regression(LR) model. Cohen's Kappa is 0.0238 higher whereas Detection rate is 1.4% higher for LR model than simple NB model. Sensitivity measure for LR model is 63.97% which is 12.38% higher than a simple NB model. From fig 15, a clear indication is seen that the Logistic Regression model is performing better with a greater ROC curve and AUC value 0.7758, 0.0796 higher than a simple NB model.

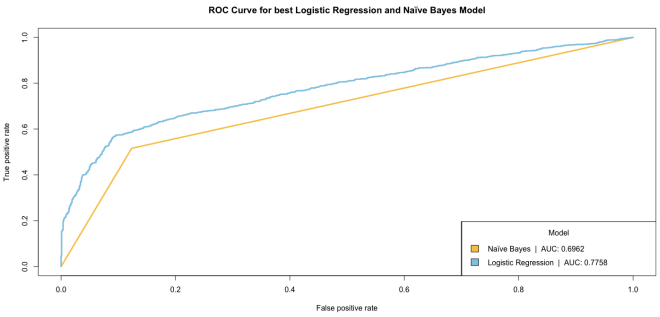


Fig. 15: Comparison between ROC plots of Logistic Regression and Naïve Bayes model

From this evaluation and comparison between the models, it is concluded that given the scenario and selected attributes and methods, Logistic regression model built on a balanced bank dataset performs better than Naïve Bayes classifier. The interpretation of the logistic model is as follows:

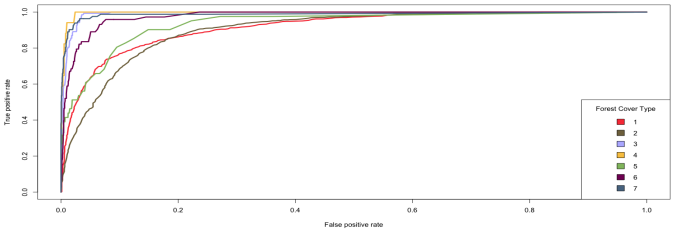
- For all the clients of the bank, this model predicts 81.9% of clients answer for subscription correctly. This is a fairly good accuracy.
- Cohen's Kappa of 0.3472 is considered as fair level of agreement indicating this model is better to some degree than just a random model.
- Sensitivity of 63.97% indicates that 63.97% of fraction true values are predicted as true.
- Detection Rate of 7.2% of the total predicted "yes" clients are true "yes".
- AUC with value of 0.7758 is a fair acceptable measure as seen in the ROC plot.
- ROC plot indicates a conservative property of making a "yes" classification of subscription based on strong evidence to make fewer errors. This is a good indication of the application in real world bank scenario.

C. Forest Land Tree types Models

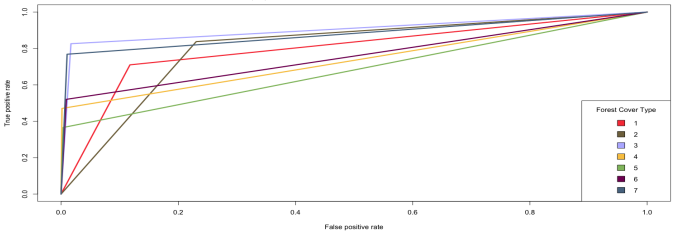
As discussed in the previous section, best models are chosen for the target purpose of the work. In this section, we compare these two best models, SVM and RF, to check which method performs better using Accuracy, Cohen's Kappa, ROC using One-vs-All approach, and Mean Area Under the Curve(AUC) performance metrics. Table IV shows the performance measures of the best SVM and best RF and fig. 16 shows ROC with One-vs-All classification approach, with the given scenario.

TABLE IV: Comparison between Final SVM and RF model

Performance Metrics	Best RF Model	Best SVM Model
Accuracy	80.48%	76.96%
Cohen's Kappa	0.6758	0.621
Mean AUC	0.9581	0.7934



(a) Random Forest



(b) Support Vector Machine

Fig. 16: ROC One-vs-All for final SVM and RF model

From the comparison with the given metrics and ROC plot, as all the metrics are greater, we conclude that Random Forest with 6 features and 199 decision trees model performs best given this scenario. The ROC plot and AUC measure of this model is considerably high to be reliable for correct predictions made. The interpretation of this model is as follows:

- For all the trees in the given forest land, this model predicts 80.48% of trees are correctly predicted. This is a fairly good accuracy.
- Cohen's Kappa of 0.6758 is considered as good level of agreement indicating this model is far better than just a random model.

- AUC with value of 0.9581 is an outstandingly acceptable value as seen in the ROC plot.
- ROC plot indicates that for each class, a conservative nature of making a correct prediction of that class is based on strong evidence to make fewer errors.

VI. Conclusion and Future Work

Thus, five machine learning algorithms are built on 3 completely different types of dataset and are critically compared and evaluated.

For critical temperature prediction of superconductor, it is seen that a linear model performs poor. By including polynomial and square root complexity, the train RMSE drops very low and test RMSE surge high. This result can be due to some non-linear patterns between the principal components and the critical temperature. This is the limitation of linear regression that there must be a linear relationship between dependent and independent variables. As a future study, SVM regression and decision tree regression can be implemented for the superconductor dataset and evaluate them to see if these models can find patterns and better fit the data.

In Banking scenario, practically it is important for banks to know which clients will subscribe to their new campaign, and to target only those clients for leads to save time and increase profits. Thus, it is important to have a higher specificity rate and a conservative ROC for such predictions than just accuracy measure. For our evaluation between Logistic regression and Naive Bayes, even though both the models are reasonably good performing models, Logistic Regression model is chosen as to be the best model with the given scenario between these two methods. In this banking part of the project, due to limitation of time, Maximum Likelihood assumptions for Logistic Regression were not checked. As a future continuation for this study, the assumptions can be checked and also a more robust SVM binary classification model can be compared and evaluated. Also, Naive Bayes algorithm can be evaluated on the balanced bank dataset.

A rigorous evaluation between same methods, SVM and RF, and comparison between these methods is successfully performed on a multi-class classification problem on a Forest Cover type data. The aim to select best possible predicting algorithm for this data in given scenario is achieved. Random Forest with 199 trees and 6 variables inclusion in sub decision trees performed highest. But, as evaluating the graphs it is clear that any value after 100-150 trees with number of variables considered >3 will perform similar with just 0.1 error difference at maximum. Also, SVM performed quite well considering Kappa measure as 0.621 and Mean AUC to be 0.793. Even though SVM is intrinsically more suited for binary classification, it performed considerably well with respect to Random Forest. Also, as time constraint in model building of these two time consuming sophisticated models, no resampling was performed while building them. As a future continuation of this work, we can use resampling to check if it results in better accuracy. And, in comparison to that a more sophisticated neural network models can also be applied and evaluted to see if they perform well than Random Forests on this given scenario.

References

[1] Al-Mukhtar M. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. Environmental Monitoring and Assessment. 2019;(11):1. doi:10.1007/s10661-019-7821-5.

[2] Zhang J, Zhang X, Dong S. Estimation of Crude Oil Minimum Miscibility Pressure During CO2 Flooding: A Comparative Study of Random Forest, Support Vector Machine, and Back Propagation Neural Network. 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Technology and Mechatronics Engineering Conference (ITOEC), 2020 IEEE 5th Information. June 2020:274-284.

[3] Roy SS, Mittal D, Biba M, Abraham A. Random Forest, Support Vector Machine and Nearest Centroid Methods for Classifying Network Intrusion. Annals Computer Science Series. 2016;14(1):9-17. Accessed January 2, 2021.

[4] Chemchem A, Alin F, Krajecki M. Combining SMOTE Sampling and Machine Learning for Forecasting Wheat Yields in France. 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Artificial Intelligence and Knowledge Engineering (AIKE), 2019 IEEE Second International Conference on, AIKE. June 2019:9-14. doi:10.1109/AIKE.2019.00010

[5] N. Zurbuchen, P. Bruegger and A. Wilde, "A Comparison of Machine Learning Algorithms for Fall Detection using Wearable Sensors," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 427-431, doi: 10.1109/ICAIIC48513.2020.9065205.

[6] M. Khoshlessan, B. Fahimi and M. Kiani, "A comparison between Machine learning algorithms for the application of micro-grids Energy management," 2020 IEEE International Conference on Industrial Technology (ICIT), Buenos Aires, Argentina, 2020, pp. 805-809, doi: 10.1109/ICIT45562.2020.9067203.

[7] Valluru D, Jeya IJS. IoT with cloud based lung cancer diagnosis model using optimal support vector machine. Health Care Management Science. 2020;23(4):670-679.

[8] James, G, Witten, D, Hastie, T, and Tibshirani, R 2013, "An Introduction to Statistical Learning With Application in R"