Frame Work to Classify Data in Interactive System to Enhance Decision Making

M. Chandrakumar Peter[a], and Dr. A.B. KarthickAnandBabu[b]

[a] Research Scholar, Department of Computer Science, Tamil University, Thanjavur, Tamilnadu, India
[b] Assistant Professor, Department of Computer Science, Tamil University Thanjavur, Tamilnadu, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: Decision-making is a process of choosing among alternative courses of action for the purpose of attaining a goal or goals. The ultimate objective of data analytics is to ease the decision making process but this has lot of challenges and proper planning is the only way to overcome. The idea behind this research work is to propose a novel framework for data analytics to make effective decisions in an organization by aiding the various stages of the decision making mechanism. According to the design science methodology, the research has been formulated and used in the frame work design process. A novel framework was proposed that combines different aspects of data analytics, needed architectures and tools are incorporated in the various stages of decision making process. Based on the Simons the decision-making process, a new framework was designed with 4 phases namely Data, analytical, model deployment and visualization. The decisive objective of the proposed framework is to ease the process of decision making and also to take effective decision.

In the process of future planning by the organization, it needs simple accurate estimation techniques for predictions to make effective decisions. Predictions always deal with the future events based on past incidents or records. Different kinds of predictions have been done regularly in many fields for the benefit of an individual, a group of people, an organization or a country. Support Vector Machines can be used to create a powerful prediction model because of its capability in classification and regression. The purpose of this research work is to develop a decision support system model was developed using novel algorithm. The newly developed framework has been proposed for the purpose of data analytics and for prediction. In this work, the machine learning algorithms Support Vector Machines (SVM), Random Forest, Decision Tree, Naïve Bayes and the newly proposed algorithm has been analyzed and results are compared. The outcome of this research work proves that the proposed framework model provides better result than other model.

The objective of designing and developing the proposed framework is to ease the process of decision making in the scenario of interactive system. A student performance assessment is used to evaluate the proposed framework using real data. To test the correctness of proposed framework, an experiment was done with the student data to predict student performance using newly proposed machine learning framework. The result justifies the proposed framework for the decision making process, gives added value.

Keywords: Classification, Overlapping, Over fitting, Feature Selection

1. Introduction

Machine learning is a form of Artificial Intelligence that enables a system to learn from data rather than through explicit programming. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine-learning model is the output generated when you train one of your machine-learning algorithms with data. After training the machine, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model. Machine learning applied in various fields like Education, Medicine, Retail, Finance, Manufacturing, Bioinformatics, agriculture and Telecommunication [1]. Here we are considering the supervised learning methods to classify the given data. The training dataset to get better boundary conditions that could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is calling it as classification.

Predictive analytics reveals the relationships and patterns within huge volumes of data that can be used to predict events and behaviour, and it is spread vastly. Predictive analytics is majorly used in business areas, education, politics, agriculture etc., where it helps to reach conclusions about customer behaviour and helps to understand patterns to create new sales and reduce churn to the competition [2].

1.1Types of Machine Learning Algorithms

Machine learning divides into three different groups based on their learning style: They are 1.Supervised learning, 2.Unsupervised learning and 3. Reinforcement learning

1.1.1 Supervised Learning:

It takes place when an algorithm learns from input data also known as training data which has known labels .A model is prepared through the training or learning process which predicts the correct response when given a new example. The supervised approach is further divided into two: Classification and Regression. In classification, the algorithm predicts the class to which the given tests data fall into whereas regression predicts a numeric value for target variable [3].

1.1.2 Unsupervised Learning:

This type of machine learning occurs when an algorithm learns from input data without any labels and does not have a definite result, leaving the algorithm to decide the data patterns on its own. A model is equipped by learning the features present in the input data to extract general rules. It is done through a mathematical process to reduce redundancy or to organize data by similarity. Unsupervised learning is again majorly used in two different formats: Clustering in which we group similar items together and density estimation which is used to find statistical values that describe the date.

1.1.3 Reinforcement Learning:

Reinforcement learning allows machines to determine automatically its behaviour within a specific context to maximize its performance. Simple reward feedback is required to learn its behaviour known as reinforcement signal. This learning occurs when you present the algorithm with examples that lack labels, as in unsupervised learning. Reinforcement learning is linked to applications for which the algorithm must make decisions unlike unsupervised learning it is treated as learning by trial and error method.

2. Literature Survey

The amount of data growing within the last few decades is enormous which leads to variety of different Machine Learning techniques that have been developed to analysis the available data. All of them could be divided into two; Supervised and Unsupervised according to whether they have labelled instances or not. If they have, then these techniques are called supervised, if not they are referred to as unsupervised [1].

Information Technology plays vital role in Machine learning to collect huge amount of data and which is used smart data analysis for technological progress with great accuracy whereas, It is a quite nature the manual work has some mistakes during analyses or chance to establish relationships between multiple features [2].

ML is used to make prediction and better understand the system. Machine learning has a wide range of application such as Education, Retail, Finance, Manufacturing, Medicine, Telecommunication, and Bioinformatics [4].

In the existing systems of machine learning model, initial dataset will be imported and it is pre-processed. Pre-processed data will be split into training set with 80% of data and the rest of the data (20%) will be taken as testing set. Building models by using various classifications models such as Support Vector Machines (SVM), Random Forest, Decision Tree and Naïve Bayes. In SVM classification algorithm kernel used to transform the given dataset [5].

Predicted value from the trained model will evaluate the performance such as accuracy, sensitivity and specificity of the build model using classification and regression algorithm. Various types of functions such as linear, polynomial and radial basis function (RBF) where polynomial and RBF kernels are used for non-linear function [6].

Existing framework model requires high training time. It works poorly with overlapping classes and it is also sensitive to the types of kernel used. So, the existing system suffers with overlapping and overfitting problem, moreover it also consumes more time.

3. Methodology

3.1 Proposed System

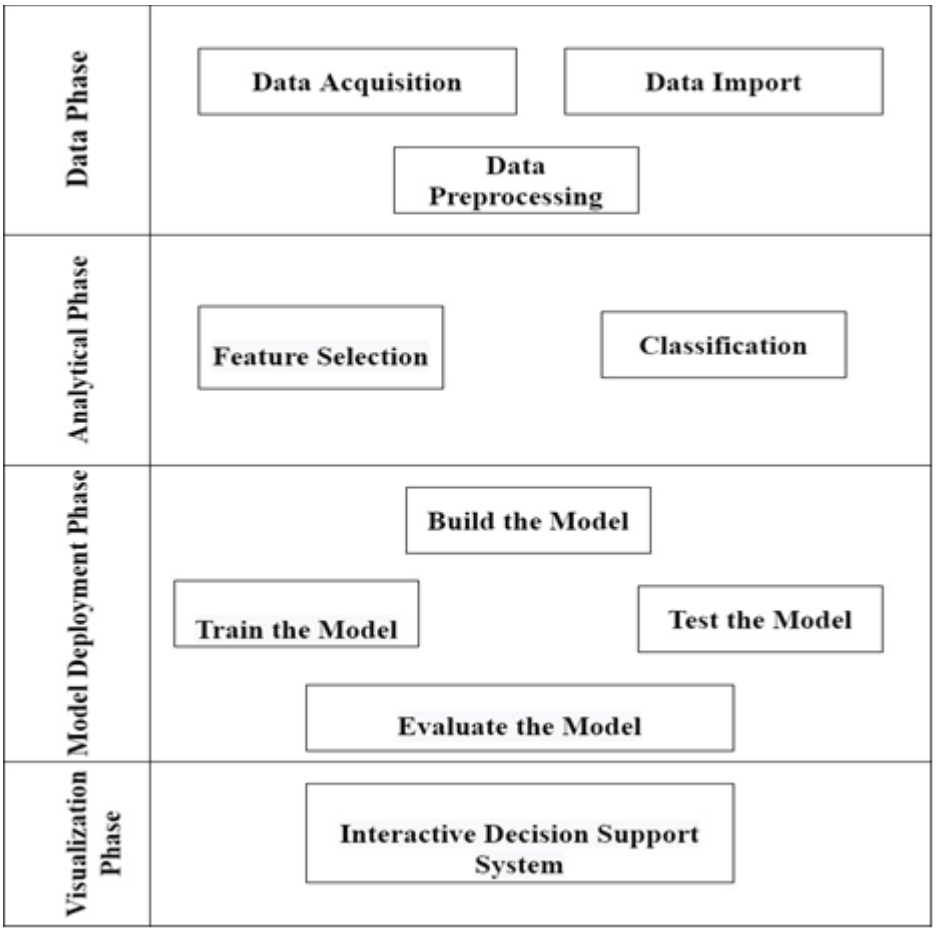
3.1.1 Data Acquisition:

The data can be numerical, Categorical and Ordinal. The datasets are collected from UCI Machine Learning Repository. Datasets can also be retrieved from several data sources such as Kaggle Datasets, Datasets via AWS, Google's Dataset Search Engine, Microsoft Datasets, Awesome Public Dataset Collection, Government Datasets, Computer Vision Datasets and Scikit-learn dataset.

3.1.2 Data Pre-processing:

Data Pre-processing plays a vital role in designing and developing the classification framework. The misleading data reduces the accuracy of the framework. Misleading data refers to irrelevant, missing, redundant and noise data. It is necessary to give importance in pre-processing the data before the process of classification.

Figure 1: Architectural Diagram of Framework



3.1.3 Classification:

3.1.3.1 Feature Selection:

Feature Selection is the first step of any machine learning model. The performance of the model highly depends on the process of feature selection. The features of the data that are used to train the model have a huge impact on the performance of the model. If the irrelevant or partially relevant features of the data are taken into consideration then it is resulted in a negative impact on the performance and reduces the accuracy of the model.

This process selects the features that can contribute more to the prediction or output of the framework. In this framework, the recursive feature rejection algorithm is proposed for the feature selection process. This algorithm aims to identify the best performing feature subset. It frequently creates models and keeps aside the best or the worst performing feature at every iteration. It constructs the next model with the left features until all the features are drained. It then ranks the features based on the order of their removal.

1: Add randomness to the specified data set by making shuffled copies of all features.

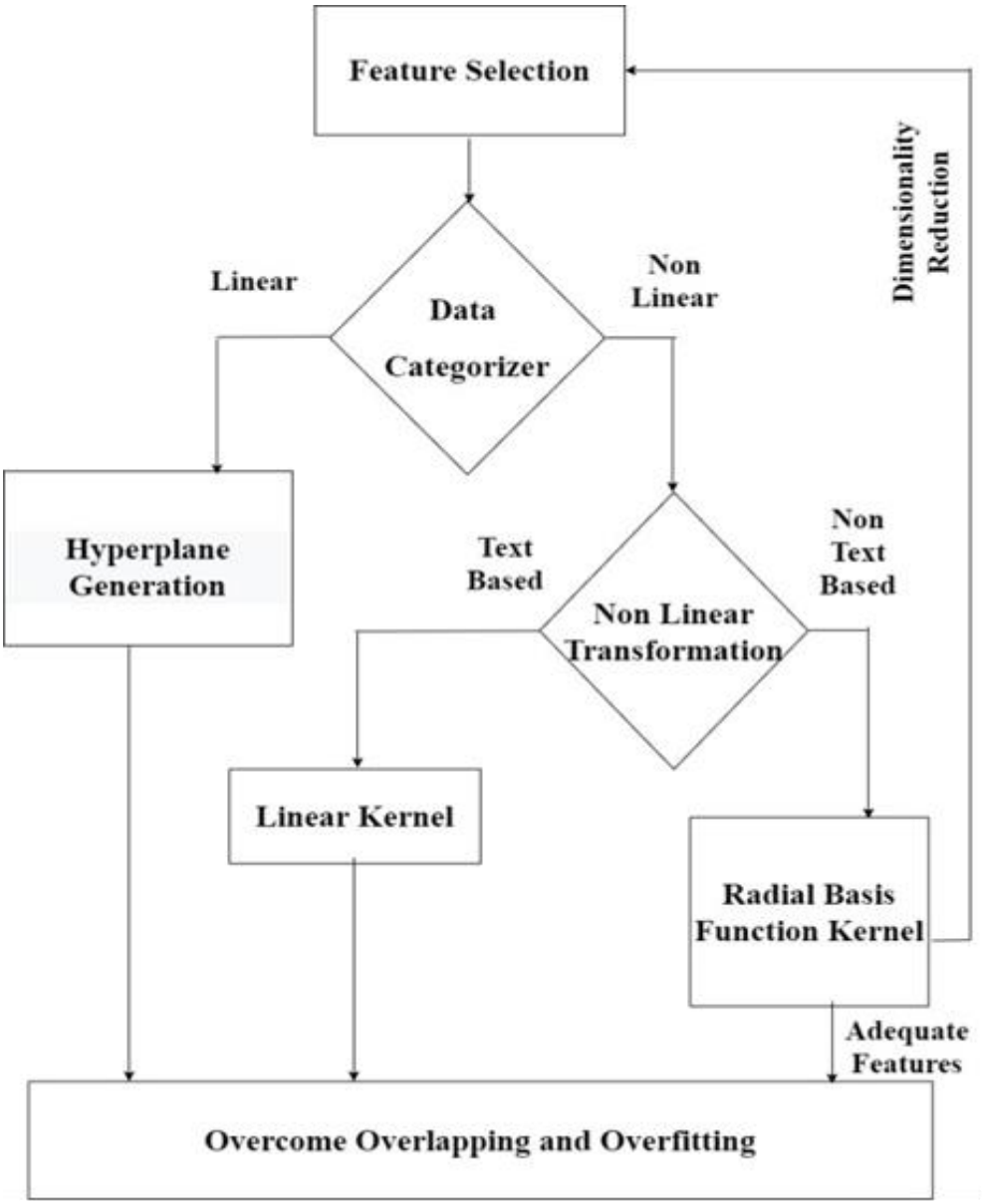
2: Applies the feature importance measure to assess the importance of every feature where higher means more vital. Mean Decrease Accuracy is used for the feature importance measure.

3: At every iteration, it checks whether an actual feature has a higher importance than the best of its supportive features (i.e. whether the feature has a higher Z-score than the maximum Z-score of its supportive features) and constantly removes features which are deemed highly unimportant.

4: The algorithm stops either when all features get confirmed or rejected or it reaches a specified limit of runs.

The square-root function is used for the identification of the specific limit of runs. This function will take square root of the total number of features in individual run.

Figure 2 : Flow Diagram of Analytical Phase



3.1.3.2 LNL Data Categorizer:

The LNL Data Categorizer is used to categorize the data as Linear or Non Linear Data. The LNL consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages – (1) analyzing the correlation and directionality of the data, (2) estimating the model, i.e., fitting the line, and (3) evaluating the validity.

3.1.3.3 Hyperplane Generation:

If the Data Set belongs to Linear Data Category, the process of generating Hyperplane is to be activated. During the generation of the Hyperplane, the equation of the line is considered. The equation of the line is

$Y = aX + b$

The straight line is generated using the above equation. The generated straight line is referred as Hyperplane. The data points on either side of the hyperplane that are closest to the hyperplane are called

Support Vectors which is used to plot the boundary line.

3.1.3.4 Non Linear Transformations:

If the Data Set belongs to Non Linear Data Category, the process of generating Non Linear Transformations is to be activated. The Non Linear Transformations is used to project data points to higher dimensional space.

Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single line. It is one of the most common kernels to be used. It is mostly used when there are a large number of features in a particular data set. One of the examples where there are a lot of features is Text Classification, as each alphabet is a new feature. So we mostly use Linear Kernel in Text Classification.

1: Calculate dot product of two vectors to check how much they make an effect on each other

2: If Data = Text based then

3: Use Linear Kernel (with More Features) to reduce training time else

4: Use Gaussian Radial Basis Function (RBF) Kernel for less Features to reduce training time

i: P = No. of input features/values.

ii: M = No. of transformed vector dimensions (hidden layer width). So M ≥ P usually be.

iii: Each node in the hidden layer, performs a set of non-linear radian basis function.

F(x, xj) = exp(-gamma * ||x - xj||^2), The value of gamma varies from 0 to 1. The most preferred value for gamma is 0.1.

5: Repeat the process of Feature Selection

3.1.3.5 Overfitting Problem Support (OFPS):

The redundant data received from the process of data collection prone to the opportunity to make decisions based on noise. Noise refers to the irrelevant information or randomness in the dataset and also interferes with signal. It is necessary to separate the signal and the noise to avoid the overfitting problem. If the dataset has too many features and the model is not properly regularized, then the model will take up the noise into consideration instead of signal. This overfit model will then make predictions based on that noise. The model will perform badly on the unseen data. It is essential that the model has to obey the concept of goodness of fit. Goodness of fit refers to how closely a model's predicted values match the observed or true values.

To overcome the problem of overfitting, the following algorithm is proposed.

1: Split the dataset into separate training and test subsets

2: Generate multiple mini train-test splits from the training and test subsets

3: Train the model iteratively

4: For each iteration, measure the performance of the model by comparing the accuracy obtained on the train and test subsets

5: If Accuracy (Train)>Accuracy (Test) then goto Step 3

6: If Accuracy (Train) == Accuracy (Test) then stop the training at that point.

This algorithm overcomes the problem of overfitting by ending the process of training when the accuracy level of training and test dataset is equal.

3.1.3.6 Overlapping Problem Support (OLPS):

Class overlap is caused due to ambiguous regions in the data where the prior probabilities of two or more classes are approximately equal. Hence the Class imbalance and overlap are strongly coupled with each other; the importance has to be given in addressing the challenge of class overlap in the presence of imbalanced classes.

OLPS is a clustering-based under sampling technique that identifies data regions where minority class samples are embedded deep inside majority class. By removing majority class samples from these regions, OLPS pre-processes the data in order to give more importance to the minority class during classification. Experiments show that OLPS achieves improved performance over an existing method for handling class imbalance.

1: Let S be the original training set.

2: Form clusters on S denoted by Ci such that 1 <i< |C|.

3: Find the degree of minority class dominance and denote it by r.

4: r = No. of minority class samples / Size of the cluster

5: If r = 0, then all the samples of the cluster belong to the majority class

6: If r = 1, then all the samples belong to the minority class

7: For clusters that satisfy 0 <ri< 1 and r ≥ τ (where, τ = f(r) is an empirically determined threshold for r and is uniform over all the clusters), remove all the majority class samples and retain the minority class samples.

First, the entire training data is clustered ignoring the class attribute using Euclidean distance as the distance measure. The degree of minority class dominance of each cluster, denoted by r, is calculated as the ratio of number of minority class samples to the size of the cluster. Therefore, r = 0 indicates that all the samples of the

cluster belong to the majority class, and r = 1 indicate that all the samples belong to the minority class. For clusters with $0 \leq r \leq 1$, the majority class samples are removed if r is equal to or greater that an empirically determined threshold τ. Clearly, if the τ is low, more majority class examples would be removed as compared to when τ is high. This method creates a "vacuum" around the minority class samples in each cluster and thus helps the machine learning classifiers learn the decision boundary more efficiently.

3.1.4 Evaluating the Model:

The main aim of the evaluation is to fit the best line within a threshold value t. The threshold value refers to the distance between the hyperplane and boundary line. Boundary lines are the two lines apart from hyperplane, which creates a margin for data points. The points that are located within the boundary value satisfies the condition

$-t < Y-aX+b < t$

are used for predicting the value. The error between the real and predicted value is minimized using the proposed model.

3.1.4.1 Feature Scaling:

Normally, the data gathered during the data collection process are of varying ranges and magnitudes which make building the model difficult. Thus, the range of the data needs to be normalized to a smaller range which enables the model to be more accurate in the training phase itself.

3.1.4.2 Training the Model:

Generally, the data set needs to be split into the training set and the test set in any machine learning model. The proposed model is trained with the values of the training set and the predictions are tested on the test set.

3.1.4.3 Predicting the Results:

The values are predicted using the proposed model. The real values and the predicted values are compared. If there is a significant deviation between the predicted values and the real values of the test set then it is concluded that the proposed model is not the perfect fit for the tested data set.

4. Results and Discussions:

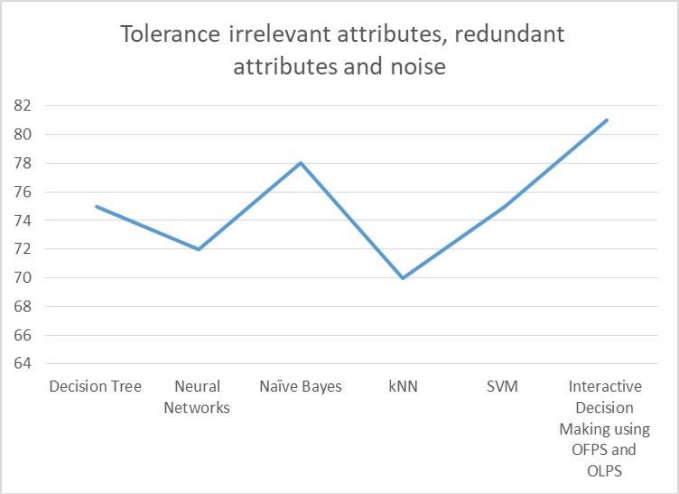
The Proposed Framework "Interactive Decision Making using OFPS and OLPS" concentrates in reducing training time, overcoming the problem of overlapping and overfitting. This research work contributes the following concepts which are made as a part of the proposed framework.

i. Feature Selection
ii. LNL Data Categorizer
iii. Non Linear Transformations
iv. Overfitting Problem Support
v. Overlapping Problem Support
vi. Evaluation the Model

The proposed framework is compared with the various existing algorithms and the result shows that the efficiency of the proposed methodology is improved. The following tables and graphs show that there is a significant improvement in the performance of the proposed methodology based on the following parameters: (1) Tolerance irrelevant attributes, redundant attributes and noise, (2) Learning Speed, (3) Classification Speed, (4) Dealing with Overfitting, (5) Dealing with Overlapping and (6) Accuracy

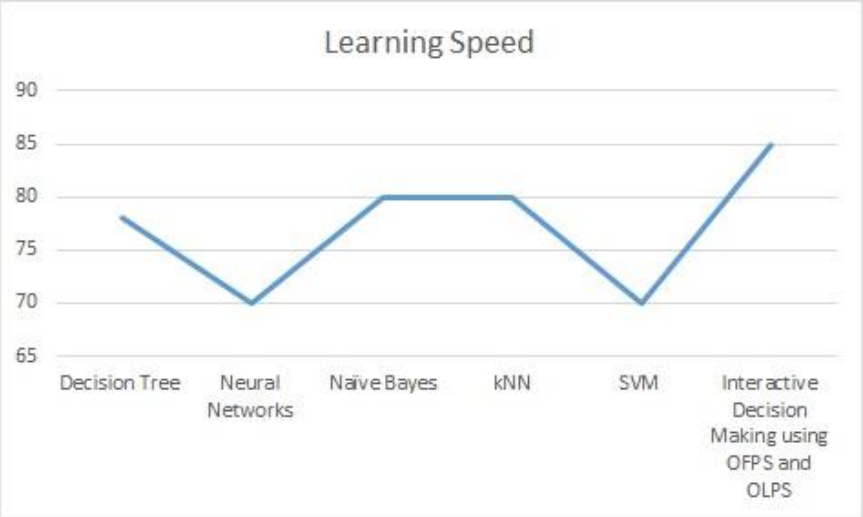
4.1. Feature Selection:

Tolerance irrelevant attributes, redundant attributes and noise	**(in %)**
Decision Tree	75
Neural Networks	72
Naïve Bayes	78
kNN	70
SVM	75
Interactive Decision Making using OFPS and OLPS	81



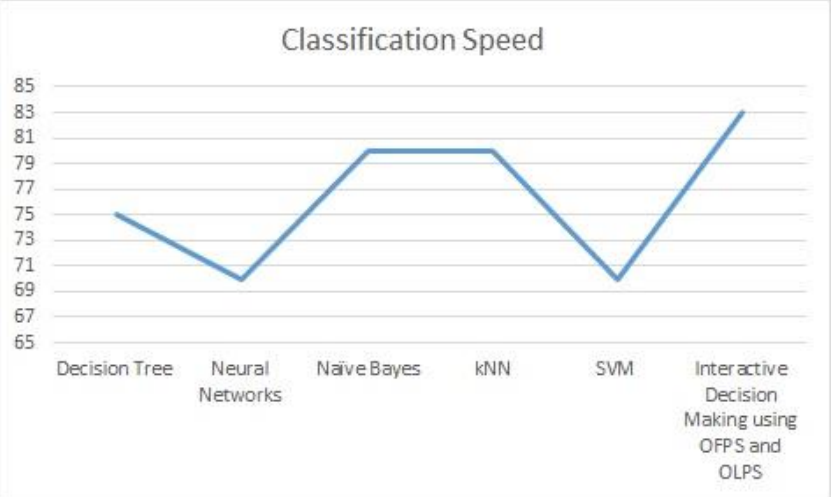
4.2. LNL Data Categorizer:

Learning Speed	**(in %)**
Decision Tree	78
Neural Networks	70
Naïve Bayes	80
kNN	80
SVM	70
Interactive Decision Making using OFPS and OLPS	85



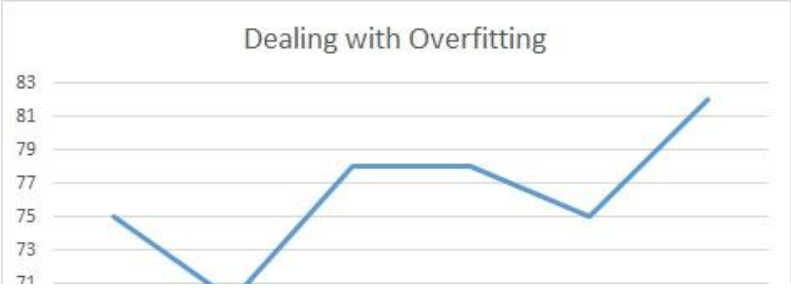
4.3. Non Linear Transformations:

Classification Speed	**(in %)**
Decision Tree	75
Neural Networks	70
Naïve Bayes	80
kNN	80
SVM	70
Interactive Decision Making using OFPS and OLPS	83



4.4. Overfitting Problem Support:

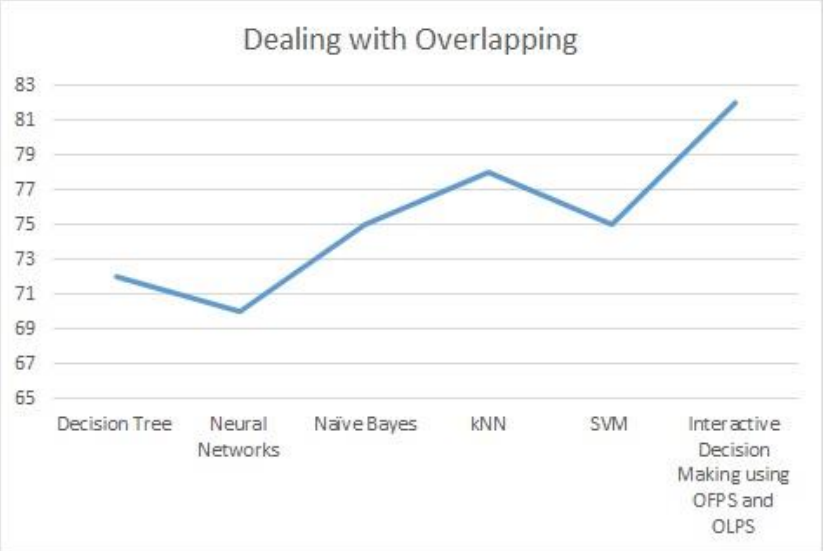
Dealing with Overfitting	**(in %)**
Decision Tree	75
Neural Networks	70
Naïve Bayes	78
kNN	78
SVM	75



Interactive Decision Making using OFPS and OLPS	82

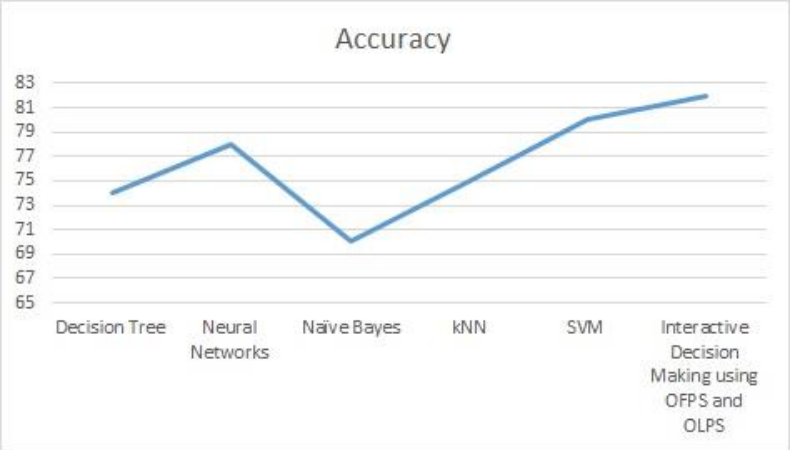
4.5. Overlapping Problem Support:

Dealing with Overlapping	**(in %)**
Decision Tree	72
Neural Networks	70
Naïve Bayes	75
kNN	78
SVM	75
Interactive Decision Making using OFPS and OLPS	82



4.6. Evaluating the Model:

Accuracy	**(in %)**
Decision Tree	74
Neural Networks	78
Naïve Bayes	70
kNN	75
SVM	80
Interactive Decision Making using OFPS and OLPS	82



5. Conclusion:

In this research work, it shows that proposed framework model with various machine learning algorithms such as Support Vector Machines, Random Forest, Decision Tree and Naïve Bayes. Beside this a newly proposed algorithm will also be used for classification and regression to enhance the decision making. Experimental result shows that the accuracy of the new algorithm used in the proposed framework is more in terms of accurate prediction. Meanwhile, the experiment result also shows that it performs better with different classifiers. The proposed framework improves the test speed and avoids the overlapping and overfitting problems. Experimental results support the above statement.

As future works, researcher plans to extend the proposed framework with different combination of classifiers to achieve better performance and to suit any kind of prediction application.

References

1. KavitaPabreja, , Maharaja Surajmal, "Comparison of Different Classification Techniques for Educational Data" India International Journal of Information Systems in the Service Sector Volume 9 • Issue 1 • January-March 2017- DOI: 10.4018/IJISSS.2017010104.
2. Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007). Pp. 249 – 268. Retrieved from IJS website: http://wen.ijs.si/ojs-2.4.3/index.php/ informatica/article/download/148/140.
3. Nguyen N., Paul J., and Peter H. (2007). A Comparative Analysis of Techniques for Predicting Academic Performance. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12.
4. Kumar S. Anupama and Dr.Vijayalakshmi M.N. (2011). Efficiency of Decision Trees in Predicting Students Academic Performance. Computer Science& Information Technology 02, pp. 335–343.
5. Bharadwaj B.K., Pal S. - Data Mining A Prediction for Performance Improvement Using Classification. International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.
6. Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in HigherEducation," IEEE Access , Volume: 4 ,2016.
7. G. Gray, C. McGuinness, P. Owende, An Application of Classification Models to Predict Learner Progression in Tertiary Education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554
8. S. T. Jishan, R. I. Rashu, N. Haque, R. M. Rahman, Improving accuracy of student's final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, Decision Analytics 2 (1) (2015)
9. AmirahMuhammedShahiri, A Review on Predicting Student's Performance Using Data Mining Techniques (2013)
10. M. Sharmila Begum And Dr. A. George An Analysis on Data Mining Processes on Big Data Framework, International Journal of Computer Sciences and Engineering, Vol.2, Issue 12, January 2015.