CUSTOMER PATTERNS ANALYSIS USING MULTI-LEVEL ATTRIBUTE SELECTION WITH MACHINE LEARNING MODELS

R.Siva Subramanian

S.A.Engineering College, Chennai, India,
sivasubramanian12@yahoo.com

Dr.M.G.Dinesh,

Assistant Professor, Dept of CSE,
Easa College of Engineering and Technology, Coimbatore
dineshbabu.mg@gmail.com

B.Maheswari

Assistant Professor, Dept of CSE,
Rajalakshmi Institute of Technology, Chennai, India,
mahesasi23@gmail.com

Dr.J.Aswini

Professor, Dept of AI & ML,
Saveetha Engineering College(Autonomous),Affiliated to Anna University
aswini.jayaraman@gmail.com

M.Anita

Assistant Professor, Dept of CSE,
S.A.Engineering College, Chennai, India,
anitam.engg@gmail.com

Abstract

Customer pattern analysis is viewed as a vital parameter for any enterprise's business. Since the business of the enterprises is completely reliant upon the customers, and further, in today's fast-paced business environment, retaining the existing customers and attracting new potential consumers is a tough challenge for each enterprise. With tremendous significance, improvements in artificial intelligence and machine learning algorithms aid in performing efficient analysis of customer pattern. Moreover, in most circumstances, the real-time datasets acquired contain irrelevant, noisy, and correlated factors, which makes it difficult for the ML model to extract the underlying information about the data. So, in order to overcome the problem of real-time customer datasets and perform efficient analysis, the Multi-Level Attribute Selection (MLAS) approach is proposed in this work as a preprocessing mechanism to alleviate the extraneous attributes and select the optimum variable subset to model with the Base Classifier. The proposed approach MLAS composed of three phrases that are used that are used to choose the right set of variables. In first step, consumer dataset related to the problem is collected, and then, second step is to perform data transformation. In third step, FS is performed to choose the right set of variables. Furthermore, the variable sets captured are applied with different machine learning models to perform effective insight about customer behaviour. In this work, two different ML models are studied: Naive Bayes, and K-NN algorithms. Lastly, experimental results obtained using the MLAS approach and without the MLAS approach are evaluated and compared using several performance metrics like accuracy, recall, precision, and F-measure.

Keywords: **Artificial Intelligence; Machine Learning; Customer Pattern; MLAS; Prediction.**

1. Introduction

Customer is considered as the focal point of the enterprises business and gaining a deeper understanding of customer patterns assists the firm in improving its operations. The study of customer behavior analysis assists the marketing professionals in understanding how consumer decisions are influenced and further helps to understand 1. Customers' idea about the enterprises and how they think about other business enterprises products, 2. what causes customers to select from different options, 3. Customer understanding from their point of shopping and researching, 4. How the customer's behavior is influenced through friends, media, and family. The customer behavior analysis is always influenced by different aspects like personal, psychological, and social factors [wen and Ye(2022)]. A greater understanding of consumer patterns aids to improve the marketing strategies and which in turn improves business decisions and makes it possible to retain the customers for a long time. In past years most of the business is based upon the product-centric which resulted in less collection of consumer-related data and analysis of consumer patterns is not performed. However, in today's context, as a result of globalization, enterprises' businesses are shifting from product-centric to customer-centric, which has increased the significance of customer pattern analysis due to rivalry in client acquisition and retention. Furthermore, the expansion of the internet, technology, and enterprise tools led to the collection of a lot of information about clients. These customer data can be used to gain a better understanding of customer patterns and perform analysis. To perform customer patterns analysis with these customer data in this research, different machine models are considered **[**Kim and Kim (2022)**]**. With help of the Machine Learning model better analysis of customers, patterns can be carried out. But in a real-time scenario, the raw customer dataset collected is not in a useful format, and also the customer data collected consists of noisy, missing, irrelevant, and correlated factors. Further with these high dimensional data efficient analyses about the customer pattern cannot be performed and also this dataset increases the computing complexity of the classifier[Bommert *et al*.(2022)**]**. So to perform efficient analysis about customer pattern, first the customer data is preprocessed and further FS technique is performed to choose the variables that are most important to the classifier and in turn helps the classifier make better predictions. Data preprocessing aids to check and eliminate null, missing, and noisy variables in the customer dataset. Next feature selection is applied to pick significant attribute set and remove the insignificant attributes from the dataset and making to obtain better customer prediction without minimizing the underlying structure. FS technique are classified into 3 types : Filtering, wrappering, and embedded. The filter strategy employs statistical analysis to rank the variables based on their association with the class label, and a variable subset is chosen from the ranked variable list. In general, the outcomes obtained using this strategy are poor, but when compared to other methods, it is faster. The wrapper approach employs a different combination of search space and induction algorithm to choose the variable subset. In this approach best variable subset is selected, but compared to time complexity this approach is less optimal. Considering the shortcomings of the preceding methodologies, a Multi-Level attribute selection approach is presented and tested in this study. The basic goal of using Multi-Level attribute selection is to select an efficient variable subset in an efficient time. First, the preprocessed customer dataset is used in conjunction with the Information Gain filter technique. The IG filter approach evaluates customer data and ranks attributes based on their relevance to the class label. The attributes with higher IG values have a stronger correlation with the output class, while the variables with lower IG values have a weaker correction connection respect to target class. Further depending on the cut off value right set of attribute is opted. Next, preselected feature set is processed with the SFS wrapper method. In the SFS wrapper approach, the best variables are selected by exploring the entire combination of search space and employing any induction algorithm[siva subramanian *et al*.(2022)]. Finally, the best variable subset is obtained. The variable subset is then modeled using two distinct machine learning models to see how the prediction results differ when utilizing the MLAS approach versus the filter approach and without MLAS approach. The experimental findings acquired by the aforementioned methodologies are represented using some of the parameters like Accuracy, F-Measure, Precision, and Recall. These metrics aid in calculating how well the ML model performed. Furthermore, time complexity analysis for the various methodologies is calculated and evaluated. From the exploratory analysis, it is clear that using the MLAS strategy helps to get better results than using the filter technique and not using the MLAS strategy. From time study, it is obvious that the K-NN model takes less time than the Naive Bayes model. The remainder of the work is structured as follows: second Chapter is a review of related literature, third Chapter is about theoretical aspects, fourth chapter is about the proposed technique, Fifth chapter is about test findings, and sixth chapter is about the summary.

2. Literature Survey

[Agrawal *et al*.(2021)**]**, emphasizes the importance of the feature selection approach in minimizing the dimension of the dataset as well as improving the performance of the model. The author discusses the significance of metaheuristic algorithms and conducts a study on various research projects carried out to tackle variable selection challenges using metaheuristic algorithms. Moreover based upon the literature survey the author implemented various metaheuristic algorithms to get an optimal variable subset. **[**Rostami *et al.(*2021)**]**, the author implies the advancement of database and computer technologies leads to the generation of an enormous amount of high

dimensional datasets. The use of these large-scale datasets with data mining applications needs high accuracy and speed. So to overcome the problem the use of dimensionality reduction mechanism is considered. The purpose of the dimensional reduction technique is to reduce dataset size without compromising accuracy in the dataset and improving the performance of the model and minimizing the computational complexity. The author performs a competitive study on various variable selection approaches using 6 distinct datasets and the attribute set captured is applied using NB, SVM, and AdaBoost. **[**Ghosh *et al.(*2021)**]**, the author of this research addresses the early diagnosis of cardiovascular diseases aids to minimize the mortality rates. The author has considered different combined datasets to perform CVD decision analysis. Two different feature selection mechanism has been employed to choose the relevant variable subset. One is Relief and the other one is LASSO. Further, the author has proposed a new approach based upon combining with the traditional classifiers. The proposed approaches are DTBM, RFBM, KNNBM, ABBM, ND GBBM. Exploratory analysis captured are represented with different parameter. Based upon the experimental results conducted using proposed approaches it clearly shows RFBM with Relief achieves superior results compared to other approaches. **[**Rostami *et al.*(2021)**]**, the author of this study, discusses the relevance of feature selection mechanisms in selecting significant variable subsets and eliminating correlated variables, which helps improve the model's performance. Various Meta-heuristic approaches to feature selection have been presented in recent years. The fundamental issue with Meta-heuristic techniques is that they ignore the correlation between the collection of selected variables. To address this issue, the author proposes the CGAFS methodology for feature selection. The experimental approach is carried out utilizing three different ML models and three different classification datasets to validate the efficiency of the variable subset derived from the proposed CGAFS methodology. Furthermore, three distinct variable selection procedures are used to compare the proposed CGAFS methodology. **[**Moghaddam *et al.*(2021)**]**, the author address the importance of variable selection preprocessing mechanism to eliminate the redundant and correlated form the dataset, which in turn helps to enhance classifier performance and minimizes the model complexity. In this research, the author proposes the MOFOA approach to perform FS. The experimental procedure of the suggested approach is carried out using eleven datasets. Further experimental results captured are compared with 7 single and 5 multi-objective methods. It clearly shows MOFOA performs better in reducing the classification error and selecting less no of variables compared to others. **[**Tubishat *et al.*2021**]**, author implies the use of different optimization algorithms with feature selection approach is increasing. So based upon the study the author employs an SSA approach with a feature selection technique. But SSA approach holds some drawbacks like local optima and population diversity. To address all these problems improved model of SSA is presented and referred to as DSSA. The experimental approach is performed with 23 datasets and outcome obtained are compared using original SSA, PSO, GA, ALO, and GOA approaches. The experimental result reveals that DSSA methodology performs superior compared to other approaches. **[**Song *et al* (2021)**],** the author address the importance of variable selection preprocessing in pattern recognition and data mining and. Based upon the feature selection approach, in this research, the author proposes BBPOSO with the MI technique. The experimental approach is performed with 16 datasets and outcome obtained is compared with the exiting feature selection approaches. Research shows that the BBPOSO method performs better than other methods. **[**Maleki *et al* (2021)**],** the author address the earlier prediction of lung cancer disease would help to increase the life span of human beings. To conduct an extensive analysis about lung cancer disease, in this research the author applies a genetic algorithm with the K-NN classifier. A genetic algorithm is applied to choose the appropriate attribute set and remove the insignificant attributes from the dataset to minimize the dimensional of the dataset. Further, the attribute set captured from the GA is modeled with a K-NN classifier to perform the diagnosis of the disease. Research performed using lung dataset shows that genetic algorithm with K-NN classifier performs superior in diagnosis of the disease. **[**Thakkar and Lohiya (2020)**]**, the author implies the importance of securing the network and implies the use of IDS helps to extract the network traffic information which is useful for inspection. The further author implies the use of all attributes in the dataset does not contribute to an efficient analysis of the network attack. So for that reason, the author addresses the use of a feature selection mechanism to choose the relevant variable subset which helps in improving the network attack prediction effectively. The author applies different feature techniques like Chi-Square, IG, and RFE to perform feature selection and further variable subset obtained is muddled using different machine learning like SVM, NB, DT, RF, K-NN, LR, and ANN. **[**Bommert *et al.*(2022)**],** the author address the issue in large volume data and implies need for feature selection helps to get the right variable subset and further helps to minimize classifier complexity. Depending upon the study, in this research, the author implement 14 different filter feature selection approaches using 11 gene datasets. The research methodology is conducted using different filter techniques and the outcome obtained are presented and compared. Research shows that simple variance approach performs superior to other approaches. **[**Wang *et al.*(2022)**],** the author implies the importance of feature selection techniques in bio-microarray data and Large scale data, since this dataset consists of high dimensional data which may holds insignificant, and irrelevant attributes. Further, the author implies the use of feature selection aid to choose the relevant attributes which in turn enhance the classifier performance and reduces the computational complexity. In this research, the author proposes the MRMSR approach to perform feature

selection. The experimental procedure is carried out using MRMSR approach and the result obtained are compared using 7 feature selection approaches.

3. Preliminaries

This section explains the fundamentals of feature selection approaches techniques applied. The section goes on to detail the Machine Learning model that was used to analyze customer patterns.

3.1. Feature Selection

FS, also known as attribute selection, is a crucial juncture in ML & pattern recognition. The goal of FS is to discard features that are unnecessary, unrelated, or inappropriate from the dataset to improve model performance. The ultimate focus of FS is to get right attributes from the entire dataset **[**Hu *et al*.(2022)]. FS is used to solve a variety of dataset problems, including image analysis, text mining, biomedical problems, etc. FS can be mathematically described as

Consider the Customer dataset $'D'$ consists $'n'$ number of features. The purpose of feature selection is to choose the significant feature from the $'n'$ features.

Given Consumer dataset $D = \{v_1, v_2, v_3, \dots v_n\}$. The aim is to choose the significant variable from the Consumer dataset D. Extract subset $D = \{v_1, v_2, v_3, \dots v_d\}$ where, $d < n$ *and* $v_1, v_2, v_3, \dots v_d$ represents variables of the dataset.

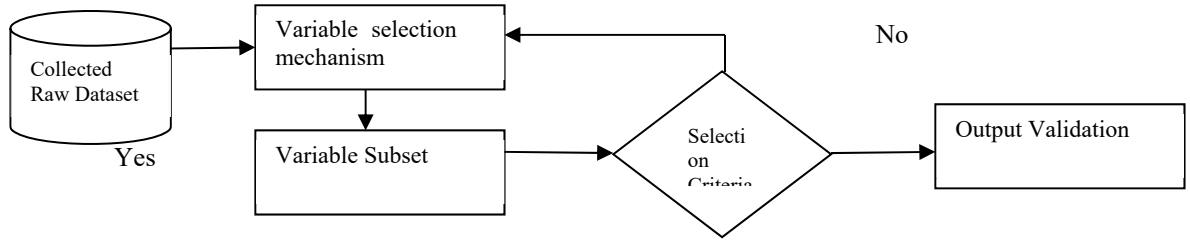

Fig 1: Variable Selection Methodology

The figure 1 represents the overall approach for the feature selection mechanism. In the FS approach, we can see that there are several steps to selecting the feature subset. The first is the collection of the dataset, followed by a selection of variable subset, then the third step is evaluation of variable subset, then selection criteria, and lastly validation. In practice there are different types of feature selection approaches these can be categorized broadly into three types; one filters, the second is a wrapper and the third is embedded approaches **[**Taylor *et al*.(2022)**]**. The filter approach is a clear and concise feature selection technique that employs a statistical method to rank variables based on their correlation with the target label and does not employ any Ml algorithm for learning. The filter method is entirely dependent on the dataset's variables. Variables with a high correlation with the target label are ranked higher, while variables with little or no correlation with the target class are ranked lower. Further based upon the ranking the variables with high scores are selected and processed with the Ml models. The filter approach can be further spitted into univariate and multivariate feature selection. In univariate individual characteristics of the variables are evaluated and ranked accordingly. But in multivariate the relationship between the variables is considered. Advantages of the filter approach are Compatible, variable-based, fast, and applies statistical methods. In the wrapper approach, the model applies some Ml model and search technique to find the superior attribute subset. The approach follows a greedy search mechanism to get right set of attributes from entire dataset using evaluation criteria. The best side of the wrapper approach selects an efficient attribute set. But downside of wrapper is computational suboptimal in contrast to filter approach. Further wrapper is spitted into Exponential search strategy, Sequential Selection strategy, and Randomized Algorithms **[**Alnowami *et al*.(2022)**]**. In the embedded approach, the model integrates the qualities of wrapper & filter approaches. Examples of embedded approaches are LASSO and RIDGE. For the purpose of overcoming the shortcomings of the filter and wrapper approach, the MLAS method is proposed and tested. In MLAS methodology, combines two approaches to select the right attributes set from the complete dataset without minimizing the customer knowledge in the dataset. The main purpose of going through Multi-Level attribute selection is to choose an efficient variable subset in efficient time. Since the filter approach works fast, but variables selected are suboptimal. Similarly, the wrapper approach selects the best variable subset but requires more computational power and time. The MLAS methodology is proposed in order to address the issues raised above in the filter & wrapper approach. In MLAS approach first, the customer data gathered is passed through preprocessing mechanism to handle missing and null data. Following preprocessing, the customer data is processed further by IG feature selection get right subset of attributes from entire dataset. The IG filter approach evaluates customer data and ranks the attributes based on their relevance to the class label. The attributes which exist higher IG values hold a higher correlation with the output class and the variables which exist lesser IG values hold lesser correlation with the output class **[**Chaudhuri and Sahu (2022)**]**.

Further based upon the threshold value right set of attributes is picked up. Next, preselected attribute set is processed with the SFS wrapper approach. In the SFS wrapper approach, the best variables are selected by going through all combination search space and the use of any induction algorithm **[**Thakkar and Lohiya (2020)**]**. Finally, the best variable subset is determined.

3.1.1 Information Gain:

Information gain measures the reduction in the entropy with respect to the transformation from the dataset. IG is applied for feature selection, in which the approach evaluate IG value for each variable with respect to the output class. IG is mathematically defined

$$IG\frac{B}{A} = H(B) - H\frac{B}{A} \quad (1)$$

$$H\frac{B}{A} = \sum_i P(A = v)H(\frac{B}{A=v}) \quad (2)$$

It is the calculation of symmetrical value between the two features. The features with high information gain are selected and for each variable, a rank is assigned accordingly with respect to information regarding the class using the specific variable. The main disadvantage of the approach is biased towards the feature having high values**[**Achin and Vanita(2022)**]**.

3.1.2 Sequential Forward Selection(SFS):

Pseudocode**:**

Input: $Y = \{y_1 y_2 \dots \dots . y_d\}$

Output: $X_k = \{x_j | j = 1,2, \dots, k;\ x_j \in Y\}, where\ k = (0,1,2, \dots, d)$

1.$Begin\ with\ null\ set\ Y_o = \{\emptyset\}$

2.$Select\ the\ next\ variable\ x^+ = \arg max_{x \notin Y_k} J(Y_k + x)$

3.$Update\ Y_{k+1} = Y_k + x^+; k = k + 1$

4. $Goto\ step\ 2$

SFS is a simple wrapper feature selection approach that starts with empty set $\{\emptyset\}$ and adds features one by one accordingly to the criterion function. Further, the procedure is carried out until the stopping condition is satisfied **[**Khurma *et al.* 2022**]**.

The attribute set captured from the MLAS approach is validated by two different machine learning approaches. One is Naive Bayes and the other one is K-NN.

3.2. *Machine Learning Algorithm:*

The variable subset captured from the MLAS approach is further processed with machine learning to examine how the feature subset obtained gets superior results in customer pattern analysis and further to examine the time complexity of the classifier. In this research, two different feature selections are considered.

3.2.1 Naive Bayes(NB)

In ML, NB model belongs to simple probabilistic classifiers which is based bayes theorem and holds important presumption about the dataset, that is attributes should be conditionally independent. Consider a problem of classification a instance which is represented by $I = (i_1, i_2 \dots i_n)$ were n represents the attributes(independent)

$p = (O_k | i_1, i \dots i_n)$ were k represents $O_k$possible outcome.

In case if the number of attributes $n\ is\ large$, then probability tables model is feasible. So with bayes theorem the above equation is decomposed into

$$p\frac{O_k}{I} = \frac{p(O_k)p(I|O_k)}{p(I)} \quad (3)$$

Further with bayesian probability,

$$posterior = \frac{prior * likehood}{evidence}$$

The numerator is similar to joint probability model

$p = (O_k,\ i_1, i_2 \dots i_n)$ can be rewritten as

$$p = (T_k,\ i, i_2 \dots i_n) = p(\ i_1, i_2 \dots i_n, O_k) \quad (4)$$

$$= p(\ i_1 \mid i_2 \dots i_n, T_k)\ p(\ i_2 \dots i_n, O_k) \quad (5)$$

$$= p(\ i_1 \mid i_2 \dots i_n, O_k)\ p(\ i_2 | i_3 \dots i_n, T_k)\ p(\ i_3 \dots i_n, O_k) \quad (6)$$

$$= p(\ i_1 \mid i_2 \dots z_n\ , O_k)\ p(\ i_2 \mid i_3 \dots i_n\ , O_k) \dots p(\ i_{n-1} \dots i_n\ , O_k)\ p(\ i_n \mid\ O_k)\ \ p(\ O_k) \quad (7)$$

Now based upon the "Naive" NB assumption

$$p(\ i_i \mid i_{i+1} \dots i_n\ , O_k) =\ p(\ i_i \mid\ O_k) \quad (8)$$

Joint model expressed as

$$p = (O_k \mid i_1, i_2 \dots i_n) \propto p(O_k,\ i_1, i_2 \dots i_n) \quad (9)$$

$$\propto p(\ O_k) p(\ i_1 \mid O_k) p(\ i_2 \mid O_k) p(\ i_3 \mid O_k) \dots \quad (10)$$

$$\propto p(\ O_k) \prod_{i=1}^{n} p(\ i_i \mid O_k), \quad (11)$$

$\propto refers\ to\ proportionality$

$$p = (O_k \mid i_1, i_2 \dots i_n) = \frac{1}{Y} p(\ O_k) \prod_{i=1}^{n} p(\ i_i \mid O_k) \quad (12)$$

were $Y = p(I) =\ \sum_k p(\ O_k) p(I \mid O_k)$ scaling factor dependent on $i_1, i_2 \dots i_n$

$$= argmax_{k \in \{1,\dots K\}} p(\ O_k) \prod_{i=1}^{n} p(\ i_i \mid O_k) \quad (13)$$

Further based the above equation the given problem instance is classified [16].

3.2.2 K-NN:

K-NN is also referred to as a Lazy learner, Since the model stores all learning data and do not perform learning immediately, and when new instances arrive the model starts its action. K-NN does not make any assumptions about the data**[**Fauziah *et al*.2022**]**.

Let's assume $p = (T_n \mid z_1, z_2 \dots z_n)$ were n represents $T_k$possible outcome and $z_1, z_2 \dots z_n$ represents the feature value. Consider a problem of classification a instance using K-NN which is represented by $I = (z_1, z_2 \dots z_n)$

1. Calculate $D(z_1, z_n)$ were $n = 1,2 \dots .i$ and D represents the distance between two points.
2. Sort the evaluated i distances in non-decreasing position
3.Assume k is a positive integer, pick first k distances from the arrganed list.
4. Look k points corresponding to these k distances
5. k_n denotes no of points to the n^{th} class among k points, $k \geq 0$
6.if $k_n > k_l\ \forall\ n \neq j$ then put z in class n.

Next section briefly explains the proposed MLAS methodology.

4. Multi-Level Attribute Selection(MLAS):

This section provides a high-level overview of the proposed Multi-Level Attribute Selection method. Figure 2 depicts the MLAS technique in its entirety.

4.1. *Description of proposed approach:*

The suggested MLAS approach involves different phases to pick the relevant attribute set to model with ML Model. The following phases are involved: Phase 1: Collecting customer data, Phase 2: Cleaning and filtering(Missing and Null value handling), Phase 3: Feature selection(Information Gain and Sequential Forward Selection), Phase 4: Modeling with Machine Learning Algorithms(Naive Bayes and K-NN models), Phase 5: Performance Evaluation (Accuracy, Recall, Precision, F-Measure).

4.1.1 Phase 1: Customer data Collection

In phase one the data about customer patterns analysis are gathered. The customer dataset gathered in this research were obtained in UCI. The dataset was obtained through bank marketing initiatives, and the goal of the customer dataset is to determine whether or not the client will subscribe to the product. The customer data collected consists of 45211 instances with 17 attributes out of which 16 are input attributes and 1 output variable. Out of 16 input attributes, 8 are numeric and the remaining 8 are categorical attributes. The output attribute consists of two classes (yes & no). From 45211 instance, 39922 instances response belongs to the YES class, and the remaining 5289 instances response belongs to No class.

4.1.2 Phrase 2: Cleaning and filtering

In phase 2 the original customer data collected are preprocessed into a useful format for further processing. First, the customer data collected consists of 79354 instances with 59 attributes. Taking out the non-conclusive findings and then deleting the variables with the same successful results in 45211 instances with 17 characteristics. Then the customer data are further checked for existing of any null, missing, and noisy variables. Further, the data cleaning and filtering are performed to eliminate the irrelevant parameters, missing and null values. Data

processing is considered an important phase, since it helps to improve the computation power and time by reducing the dimension of the customer data.

4.1.3 Phase 3: Feature selection

In phase 3 FS mechanism carried out to choose the relevant variables which aid to improve the ML classifier performance. Feature selection aims get variables which are strongly correlated with the target class and further helps to minimize the time & computational of classifier. In the work MLAS approach is carried out. First, the preprocessed customer dataset is applied with the Information Gain filter approach. In the IG filter approach, the customer data is evaluated and the attributes are ranked accordingly relevant to the class label. The attributes which exist higher IG values hold a higher correlation with the output class and the variables which exist lesser IG values hold lesser correlation with the target class. Further depending on the cut off value attribute set is picked up. Next, preselected attribute set is processed with the SFS wrapper approach. In the SFS wrapper approach, the best variables are selected by going through all combination search space and the use of any induction algorithm. Lastly, the superior attribute set is obtained. The purpose regarding going through MLAS is to choose relevant attribute set to improve classifier performance. Since the filter approach works fast, but variables selected are suboptimal. Similarly, the wrapper approach chooses the best variable subset but requires more computational power and time. The Multi-Level attribute selection strategy is offered to address the aforesaid issue in the filter and wrapper approaches.

4.1.4 Phase 4: Modelling with Machine Learning Algorithms

The variable subset selected from phase 3 should be applied with the ML model to test the efficiency of the variable subset selected through the MLAS approach. In this regard different ML models are being considered to conduct the prediction. In this research two different Ml models are considered: one is Naive Bayes and the other one is K-NN. NB is a straightforward probabilistic model relies on Bayes theorem[Siva subramanian and Prabha (2022)]. NB makes a key presumption about the data, that is attributes present in customer dataset should be independent conditionally & should be considered as equal. K-NN is a simple supervised machine learning. K-NN is also referred to as a Lazy learner, Since the model stores all learning data and do not perform learning immediately, and when new instances arrive the model starts its action. K-NN does not make any assumptions about the data.

K-fold Cross-validation(CV): CV is a resampling technique or procedure to assess the ML on limited sample data. The parameter K represents no of groups into which the given sample is spitted. In this research, here K represents 10.

4.1.5 Phase 5: Evaluation of results

The results of ML models is projected using different parameters like Accuracy, Recall, Precision, F-Measure. These parameters aid in comprehending the performance of the ML models. The further result obtained using MLAS strategy, filter approach, and without any feature selection approach are compared and presented.

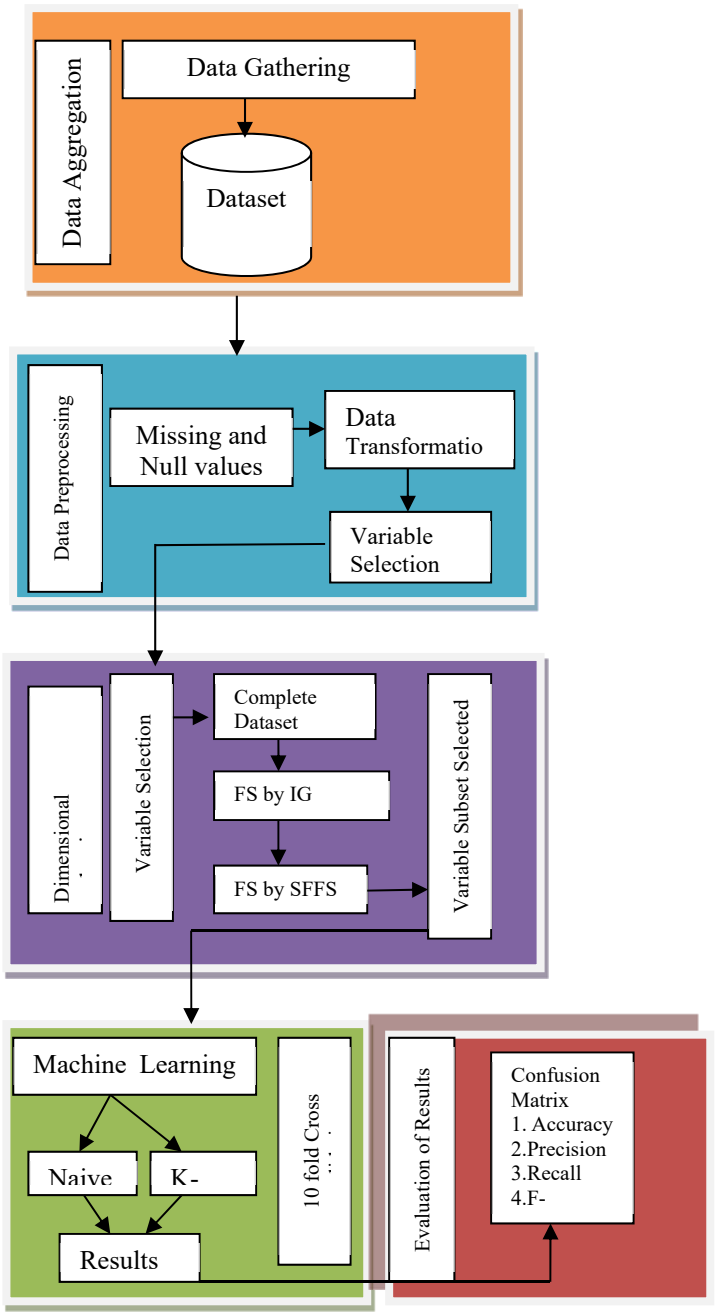

Fig 2: Description of proposed approach

4.2 Pseudocode of MLAS approach:

Input: $T = (v_1, v_2, v_3, \ldots\ldots, v_n \mid C_n)$ V_n represents variables with n instances and C_n class label
Output: Predicted Labels(Yes or No)

1. Customer data Collection related to the problem.
2. Filtering and Cleaning(Missing and Null value handling)
3. Feature selection using Multi-Level approach(Information Gain followed by SFS)
4. Modeling with Machine Learning Algorithms Naive Bayes and K-NN.
5. Results are projected using different metrics.

5. Performance Analysis

5.1 Confusion Matrix

To understand efficiency of ML model using the MLAS approach different metrics are applied to evaluate namely Accuracy, Recall, Precision, F-Measure. These metrics help to calculate how the ML model has performed [Siva Subramanian *et al.*2022]. Further, these metrics values can be calculated using the information obtained using the confusion matrix. Furthermore, the confusion matrix is denoted by four terms:
True Positive(TP): Represents a number of customers in YES Category and Model has predicted correctly.
True Negative(TN): Represents a number of customers in NO Category and Model has predicted correctly.
False Positive(FP): Represents a number of customers who are in the NO category and the model has predicted them as YES Category.
False Negative(FN): Represents the number of customers who are in YES category and model has predicted them as NO Category

5.1.1 Accuracy:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (14)$$

Accuracy represents no of predicted classes (positive and Negative) correctly

5.1.2 Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (15)$$

Recall represents no of positive classes predicted correctly from the complete positive classes.

5.1.3 Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (16)$$

Precision represents no of positive classes predicted correctly from all the classes we have predicted as positive

5.1.4 F-Measure:

$$\text{F} - \text{Measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision+Recall})} \quad (17)$$

6. Experimental Results:

The experimental strategy is performed using bank Marketing dataset acquired from UCI respiratory and further the outcome captured are compared using MLAS approach, filter approach and without MLAS approach.

6.1 Experimental Procedure(MLAS approach):

1. The customer dataset is first acquired and is preprocessed into a usable format.
2. Using the proposed MLAS process the relevant variables are selected. In the MLAS approach, first, the variables are ranked accordingly to their importance with class label using information gain filter approach and by using suitable threshold value attribute subset is selected. Moreover, the preselected attribute set is then processed using the SFS approach to choose the significant attribute subset to model with the Ml classifier.
3. The variable subset acquired using the MLAS procedure is modeled using two different classifiers one is Naive Bayes and the other one is K-NN.
4. Experimental outcomes captured are projected using different performance criteria like accuracy, F-Measure, recall, and precision.
5. Moreover experimental findings are compared using filter approach and without MLAS approach.
6. Time complexity analysis is carried out for all methodologies.

6.2 Results MLAS approach, filter approach and without MLAS approach

S.No	Model	Accuracy	Recall	Precision	F-Measure
1	NB	88.0073	0.528	0.488	0.507
3	K-NN	86.9678	0.355	0.431	0.389

Table 1: Results of different machine learning models without use of feature selection

Tab.1 summarizes the outcomes of different machine learning models without using feature selection methodology. In this case, the Naive Bayes model gets higher accuracy of 88.0073 with respect to K-NN approach. Correspondingly, the NB model has a greater recall of 0.528 than the K-NN model. Similarly, the Naive Bayes

model has greater precision of 0.488 than the K-NN model. Similarly, the Naive Bayes model has a greater F-measure of 0.507 than the K-NN model. It is obvious from the preceding experimental results that Naive Bayes performs better than K-NN classifiers without employing feature selection.

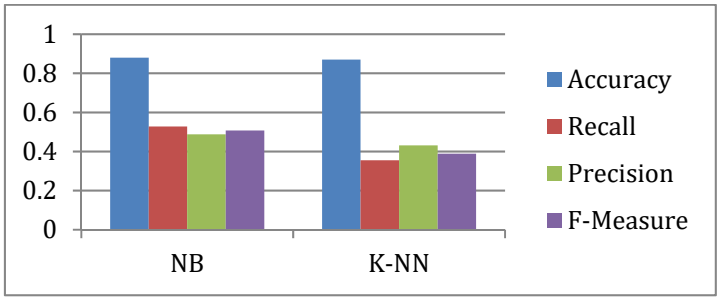

Fig 3:Comparsion of different machine learning models without use of feature selection

S.No	Model	Time in Seconds
1	NB	0.39
3	K-NN	0.08

Table 2: Time complexity analysis of different ML models without employing feature selection

Tab. 2 shows the computing time of various machine learning classifiers. In this case, we can see that K-NN takes less time in processing than the Naive Bayes classifier.

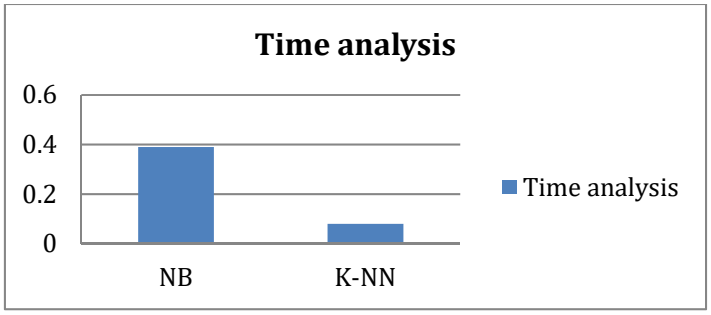

Fig 4:Comparsion of time analysis different ML models without the use of feature selection

S.No	Model	Accuracy	Recall	Precision	F-Measure
1	NB	88.8235	0.459	0.526	0.490
3	K NN	87.6225	0.455	0.470	0.462

Table 3: Results of different machine learning models using filter approach

Tab. 3 illustrates the outcomes of various machine learning models utilizing filter methodology. In this case, the Naive Bayes model has a greater accuracy of 88.8235 with respect K-NN approach. Correspondingly, the NB model has a greater recall of 0.459 than the K-NN model. Similarly, the Naive Bayes model has a greater precision of 0.526 than the K-NN model. Similarly, the Naive Bayes model has a greater F-measure of 0.490 than the K-NN model. From on the above experimental results, it is obvious that when employing filter feature selection, Naive Bayes outperforms the K-NN classifier. In addition, when compared to the experimental results in table 1, the results in table 3 are better. Since this is due to the use of relevant feature set to model with ML models.

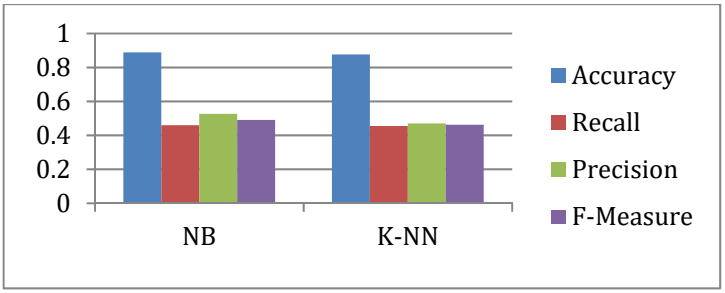

Fig 5: Comparison of different machine learning models using filter approach

S.No	Model	Time in Seconds
1	NB	0.31
3	K-NN	0.06

Table 4: Time complexity analysis of different ML models using filter feature selection approach

Tab. 4 shows the computing time of various machine learning classifiers. In this case, we can see that K-NN takes less time in processing than the Naive Bayes classifier. Similar to the computing time in table 2, computational time in table 4 are improved. Because of the usage of a relevant feature set and the removal of an uncorrelated variable subset, the processing time has been reduced.

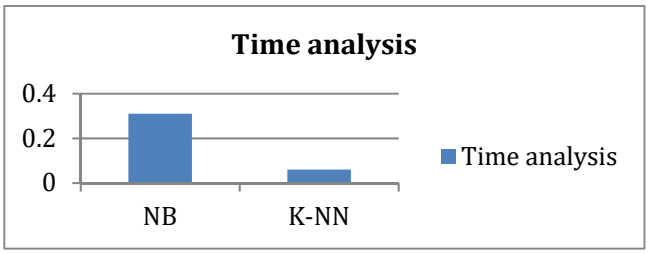

Fig 6: Comparison of time analysis different ML models using filter feature selection approach

S.No	Model	Accuracy	Recall	Precision	F-Measure
1	MLAS-NB	89.7237	0.356	0.603	0.448
3	MLAS-K NN	88.9983	0.294	0.556	0.384

Table 5: Results of different machine learning models using MLAS approach

Tab. 5 represents the outcomes of various machine learning models using the MLAS approach. In this case, the Naive Bayes model gets higher accuracy of 89.7237 with respect to K-NN approach. Correspondingly, the NBs model has a greater recall of 0.356 than the K-NN model. Similarly, the Naive Bayes model has a greater precision of 0.603 than the K-NN model. Similarly, the Naive Bayes model has a greater F-measure of 0.448 than the K-NN model. From the aforementioned experimental results, it is obvious that when utilizing the MLAS technique, Naive Bayes outperforms the K-NN classifier. In addition, as compared to the experimental findings in tables 1 and 3, the results projected using the MLAS approach are enhanced due to the application of Multi-level feature selection. Since in the filter method, variables are ranked and selected based upon using some threshold value due to this optimal variable subset is not captured. Similarly, in the wrapper approach the use of search space and the use of induction algorithm aids to get the best variable subset, but computational time is high. However, in the MLAS approach, both the inefficiencies in the filter and the wrapper are eliminated, and the best variable subset is captured in a short amount of time.

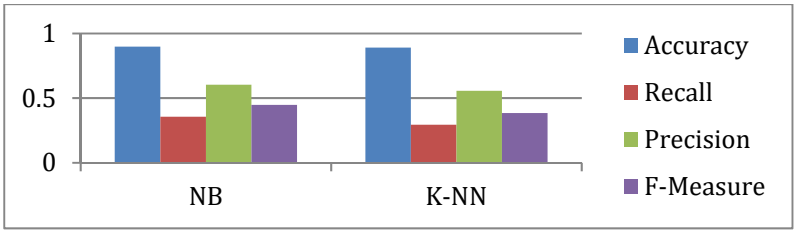

Fig 7: Comparison of different ML models using MLAS approach

S.No	Model	Time in Seconds
1	NB	0.06
3	K-NN	0.05

Table 6: Time complexity analysis of different machine learning models using MLAS approach

Tab. 6 shows the computing time of various machine learning classifiers. In this case, we can see that K-NN takes less time to process than the Naive Bayes classifier. Similar to the computing times in tables 2 and 4, the computational time in table 6 is improved. Because of the usage of a relevant feature set and the removal of an uncorrelated variable subset, the processing time has been reduced.

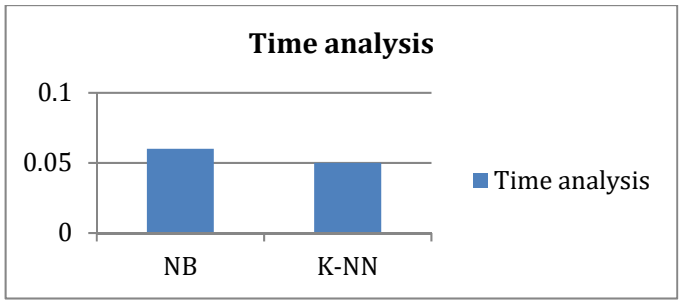

Fig 8 : Comparison of time analysis different machine learning models using MLAS approach

6.3 Results Discussion:

Tab.1–6 present the experimental data obtained using various methodologies. Tab. 1,3,5 shows experimental results of the ML model without MLAS approach, filter approach, and MLAS approach. Similarly, tab. 2,4,6 represents computational time analysis of without MLAS approach, filter approach and MLAS approach. Table 1 describes the results of the ML model without using MLAS approach. Based on the results, it is evident that Naive Bayes outperforms K-NN classifiers that do not employ the MLAS technique. Tab.2 shows the computational time analysis ML model without using the MLAS approach. From the time analysis, it is obvious that the K-NN model gets less time than the Naive Bayes model. Table 3 represents the results of the ML model using the filter approach. From the results, it is obvious that Naive Bayes outperforms better compared to the K-NN classifier using the filter approach. Tab.4 represents the computational time analysis of the ML model using filter strategy. From the time analysis, it is obvious that the K-NN model takes less time than the Naive Bayes model. Tab. 5 represents the results of the ML model using the MLAS approach Based on the results, it is obvious that the Naive Bayes classifier outperforms the K-NN classifier utilizing the MLAS technique. Tab 6 represents the computational time analysis of the ML model using the MLAS approach. From the time analysis, it's clearly understand K-NN model takes less time than the Naive Bayes model. From the experimental outcome, it is obvious that using the MLAS strategy helps to achieve superior outcomes with respect to the filter technique & without the MLAS methodology. Likewise, the filter approach helps to improve the prediction results, but the results obtained are suboptimal.

6.4 Research Findings:

1. The customer datasets collected from the real world have a highly significant of holding missing, noisy, irrelevant, and correlated variables.
2. The adoption of preprocessing techniques aids to deal with the missing, noisy null values from the customer dataset.

3. The use of feature selection techniques strategies allows to choose significant attribute subset and enhance the ML model prediction performance.

4. Results obtained using the filter approach are suboptimal since the filter technique sort the attributes accordingly to correlation with the class label. Further, there needs a cuttoff value to choose the attribute set from the ranked variable ranked list.

5. Similarly, results captured using without MLAS approach are poor compared to other approaches. Since this could be owing to the usage of all features that are irrelevant and correlated to the model with ML classifier.

6. When compared to other methodologies, the results achieved utilizing the MLAS approach are superior. Since the removal of irrelevant and correlated variables helps to achieve better prediction results in less processing time.

7. Conclusion

Customer is believed as a focal point of the enterprise business and gaining a greater understanding of customer patterns aids to increasing the customer satisfaction and retention of customers for a long time with the intent of improving the enterprise's business. However, as a result rapid proliferation of systems, the technology, and internet the amount of consumer data collected in real-time is immense. With these advancements, the customer data generated are high dimensional. To analyze these customer data and to discover the interesting patterns in the data machine learning techniques are explored. However, the real-time customer data complied comprises missing, noisy, irrelevant, and correlated factors. Further, the use of these customer data directly with ML models leads to poor prediction. To address the issue, preprocessing and feature selection procedures are explored before feeding the data into ML models. Data preprocessing strategies enable dealing with missing, noisy null values from the customer dataset. Further preprocessed customer data is passed through the suggested MLAS methodology to identify the significant attribute set and eliminate irrelevant variable subset to model with ML models. In this research two different Ml models are explored: one is Naive Bayes and the other one is K-NN. A subsequent experimental procedure is carried out using variable subset acquired using MLAS approach, filter approach and without MLAS approach. The results captured are presented from tab 1 to tab 6. Experimental results captured are projected using different performance criteria like accuracy, F-Measure, recall, and precision. Also, time complexity analysis is carried out for all methodologies. The results analysis unequivocally shows the use of the MLAS approach helps to get superior results when compared with the filter approach and without the MLAS approach. Similarly, in terms of computational time, the MLAS approach performs superior, since this is due to utilization of relevant feature set and exclusion of uncorrelated variable subset has minimized its processing time. As a future work use more high dimensional data and different ML models are encouraged.

Funding Statement: The authors received no specific funding for this study

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References:

[1] Achin J and J.Vanita (2022), Sentiment classification using hybrid feature selection and ensemble classifier, *Journal of intelligent & Fuzzy systems*, vol.42, no.2,pp.659-668,2022.

[2] Agrawal.P, H.Abutarboush, T.Ganesh and A.W.Mohamed (2021),Metaheuristic ALgorithms on Feature Selection: A Survey of One Decade of Research (2009-2019), *IEEE Access*, vol.9, pp.26766-26791.

[3] Alnowami M R, F.A. Abolaban and E.Taha (2022), A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer, *Journal of Radiation Research and Applied Sciences*, Volume 15, Issue 1, pp 104-110,2022.

[4] Bommert, T.Welchowski, M.Schmid and J.Rahnenfuhre (2022), "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," Briefings in Bioinformatics, vol.23, no.1, pp.354.

[5] Siva Subramanian, R., Prabha, D., Aswini, J., Maheswari, B, "Evaluation of Different Variable Selection Approaches with Naive Bayes to Improve the Customer Behavior Prediction". In: Smys, S., Balas, V.E., Palanisamy, R. (eds) Inventive Computation and Information Technologies. Lecture Notes in Networks and Systems, vol 336. Springer, Singapore. https://doi.org/10.1007/978-981-16-6723-7_14,(2022)

[6] Chaudhuri A and T.P.Sahu(2022), Multi-objective feature selection based on quasi-oppositional based Jaya algorithm for microarray data, *knowledge-Based Systems*, vol.236,pp.107804.

[7] Fauziah, M.A.Tiro and Ruliana(2022), Comparison of k-Nearest Neighbor (k-NN) and support vector machine(SVM) methods for classification of poverty data in papua, *Journal of Mathematics and Applied Science*, vol.2, no.2, pp.83-91.

[8] Ghosh.P, S.Azam, MJonkman, A.Karim, J.M.Shamrat, et al.(2021), Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques, *IEEE Access*, vol.9,pp.19304-19326.

[9] Hu L, L.Gao, Y.Li, P.Zhang and W.Gao (2022), Feature-specific mutual information variation for multi-label feature selection, *Information Sciences*, vol.593, pp.449-471.

[10] Kim.Y.J and H.S.Kim,(2022), The Impact of hotel customer experience on customer satisfaction through online reviews, *Sustainability*, vol.14, no.2,pp.848,

[11] Khurma R A, I.Aljarah, A.Sharieh, M.A.Elaziz, R.Damaševičius et al.(2022), A Review of the Modification Strategies of the Nature Inspired Algorithms for Feature Selection Problem, *Mathematics*, vol.10, no. 3, pp.464.
[12] Moghaddam.B.N, M.Ghazanfari and M.Fathian (2021), A novel multi-objective forest optimization algorithm for wrapper feature selection, *Expert Systems with Applications*, vol.175, pp.114737.
[13] Maleki N, Y.Zeinali and S.T.ANiaki (2021), A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection, *Expert systems with Applications*, vol.164.
[14] Rostami.M, K.Berahmand, E.Nasiri and S.Forouzandeh (2021), "Review of swarm intelligence-based feature selection methods, *Engineering Applications of Artificial Intelligence*, vol.100, pp.104210.
[15] Rostami.M, K.Berhmand and S.Forouzandeh (2021), A novel community detection based genetic algorithm for feature selection, *J big Data*, vol.4, pp.2.
[16] Rrmoku K, B.Selimi, and L.Ahmedi (2022), Application of Trust in Recommender Systems—Utilizing Naive Bayes Classifier, *Computation* vol.10, no. 1pp. 6.
[17] Song X.F., Y.Zhang, D.w.Gong and X.Y. Sun (2021), Feature selection using bare-bones particle swarm optimization with mutual information, *Pattern Recognition*, vol.112, pp.107804.
[18] Wen X and Y. Ye, (2022), An analysis of customer change, government support, and cash holdings, *Internal journal of engineering business management*, vol.14, pp. 1-11.
[19] Tubishat M, S.Jaafar, M.Alswaitti, S.Mirijalili, N.Idris, et al (2021)., Dynamic Salp swarm algorithm for feature selection," *Expert Systems with Applications*, vol.164, pp.113873.
[20] Thakkar A and R.Lohiya (2020), Attack classification using feature selection techniques: a comparative study, *Journal of Ambient Intelligence and Humanized Computing*, vol.12, pp-1249-1266.
[21] Wang.Z,X.Zhang,Z.Zhang and D.Sheng (2022), Credit portfolio optimization: A multi-objective genetic algorithm approach," *Bora Istanbul Review*, vol.22, no.1, pp.69-76.
[22] Wang Y, X.Li and R.Ruiz (2022), Feature Selection with Maximal Relevance and Minimal Supervised Redundancy , *IEEE Transactions on Cybernetics*.
[23] Taylor P , N.Griffiths , V.Hall , Z.Xu ,and A.Mouzakitis(2022), Feature Selection for supervised learning and compression, *Applied Artificial Intelligence*, pp.1-35,2022.
[24] Wang R and H.Y.B, A predictive model for Chinese children with developmental dyslexia-Based on a genetic algorithm optimized back-propagation neural network, *Expert Systems with Applications*, vol.187,pp.115949
[25] R.Sıva Subramanıan, R & Dr.D.Prabha(2022), Ensemble variable selection for naive bayes to improve customer behaviour analysis, " Computer Systems Science and Engineering, Vol.41, no.1, pp.339-355.

Authors Profile

R.Siva Subramanian is an Assistant Professor in department of Computer Science and Engineering at S.A. Engineering College, Chennai. He has completed his under graduate and Post Graduate in Information Technology. He has completed his Ph.D in Information and communication Engineering from Anna University, Chennai. His research interests include Data mining, Big Data Analytics and Customer Analysis. He has published many papers in International journals and conferences

Dr.M.G.Dinesh is an Assistant Professor in department of Information technology at EASA College of Engineering and Technology, Coimbatore. He has completed his under graduate and Post Graduate in Information Technology. He has completed his Ph.D in Information and communication Engineering from Anna University, Chennai. His research interests include Data mining, Big Data Analytics and Network Security. He has published many papers in International journals and conferences.

B.Maheswari working as an Assistant Professor In Rajalakshmi Institute of Technology, Chennai, TamilNadu, India, in the department of Computer Science and Engineering. She received the B.E degree in Computer Science and Engineering and M.E degree in Software Engineering from Anna University, India. She is currently pursuing PhD in computer science and engineering from Anna University, Chennai. Her area of interest is Internet of things and Big Data.

Dr. Aswini.J received the M.E. degree in Computer Science and Engineering from Anna University, Chennai in 2011. She has completed her Ph.D. in Computer Science and Engineering at Meenakshi Academy of Higher Education and Research, Chennai during the year 2020 and is a Professor in Department of AIML at Saveetha Engineering College, Autonomous, Affiliated to Anna University, Chennai. Her research interests include Cloud Computing, IoT and Machine Learning. She has published 20 papers in International journals and conferences.

Mrs.Anita.M working as a Assistant Professor in S.A.Engineering, Chennai, TamilNadu, India. She is currently pursuing PhD in computer science and engineering under the guidance of Dr.Meena Kowshalya from Government College of Technology, Coimbatore. Her area of interest is signal processing in Big data mining using medical and health related dataset.