Computer Systems Science & Engineering
DOI:10.32604/csse.2022.020043
Article

Tech Science Press

Ensemble Variable Selection for Naive Bayes to Improve Customer Behaviour Analysis

R. Siva Subramanian[1,*] and D. Prabha[2]

[1]Anna University, Chennai, 600025, India
[2]Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, 641008, India
*Corresponding Author: R Siva Subramanian. Email: sivasubramaniyants@gmail.com

Received: 07 May 2021; Accepted: 20 June 2021

Abstract: Executing customer analysis in a systemic way is one of the possible solutions for each enterprise to understand the behavior of consumer patterns in an efficient and in-depth manner. Further investigation of customer patterns helps the firm to develop efficient decisions and in turn, helps to optimize the enterprise's business and maximizes consumer satisfaction correspondingly. To conduct an effective assessment about the customers, Naive Bayes(also called Simple Bayes), a machine learning model is utilized. However, the efficacious of the simple Bayes model is utterly relying on the consumer data used, and the existence of uncertain and redundant attributes in the consumer data enables the simple Bayes model to attain the worst prediction in consumer data because of its presumption regarding the attributes applied. However, in practice, the NB premise is not true in consumer data, and the analysis of these redundant attributes enables simple Bayes model to get poor prediction results. In this work, an ensemble attribute selection methodology is performed to overcome the problem with consumer data and to pick a steady uncorrelated attribute set to model with the NB classifier. In ensemble variable selection, two different strategies are applied: one is based upon data perturbation (or homogeneous ensemble, same feature selector is applied to a different subsamples derived from the same learning set) and the other one is based upon function perturbation (or heterogeneous ensemble different feature selector is utilized to the same learning set). Furthermore, the feature set captured from both ensemble strategies is applied to NB individually and the outcome obtained is computed. Finally, the experimental outcomes show that the proposed ensemble strategies perform efficiently in choosing a steady attribute set and increasing NB classification performance efficiently.

Keywords: Naive bayes or simple bayes; variable selection; homogeneous ensemble; heterogeneous ensemble; customer prediction

1 Introduction

Recent developments in technology and the enhancement of Customer Relationship Management (CRM) systems have resulted in the collection of vast amounts of customer-related business data [1]. The customer data generated in the recent 2–5 years is higher than the customer data collected in the past years. The use of CRM systems by enterprises helps to get 360-degree information about customer interaction with business and powerful analyze of consumer data helps to acquire better insight into the consumer behavior patterns [2,3]. Moreover, enterprise businesses rely upon customers, the need of sorting out the probable customers and improving customer contentment towards the enterprise's business is the needed. To conduct efficient customer investigation and to develop impressive decision strategies, in this work, a Simple Bayes or Naive Bayes ML model is performed. Simple Bayes is a straightforward, and powerful probabilistic approach that helps to conduct a better customer prognosis. NB is based upon the Bayes rule and makes a strong belief that the attributes present in the customer dataset should conditionally independent and equal [4]. Contravention of these NB presumptions in the customer dataset enables the NB classifier to observe a poor performance in prognosis and increases the complexity of the simple Bayes model. Due to this presumption, simple Bayes stands exceptional compared to other ML models. However, the premise stated by NB is mostly not true in many real-time data. Since the consumer data captured by CRM holds redundant and uncertain attributes. Moreover, the captured consumer data are in large dimensional. These customer data should not be processed with the NB model for customer analysis without employing an appropriate preprocessing procedure.

To optimize the NB prognosis on these customer datasets, the need for an attribute selection approach is applied to elect an attribute subgroups to apply to the Naive Bayes model. The intent of applying attribute selection procedure is to take away the redundant and uncertain variables from the dataset and improve the efficiency of the ML model [5]. The attribute selection procedure is widely grouped into three different forms, one is the filter approach, second is the embedded approach, and third is the wrapper approach. In these feature approaches, the filter method works fast compared to the other two approaches, and for a high-dimensional dataset, the filter approach is the best option [6]. However, the attribute subgroup captured by the filter procedure is not satisfactory. However, the outcome obtained from the filter approach changes when a different subsample is applied, which is derived from the same learning. Furthermore, the drawback of picking the finest threshold value to pick the pertinent attribute set from the sorted attribute list, obtained after evaluation using filter methods is also one of the major concerns. To address the drawback with the filter method and to pick a steady attribute set for modeling with NB, in this work, an ensemble approach is used with the attribute selection method. The ensemble approach's intention is to obtain steady results by integrating several weak models into a single model. By the use of ensemble learning, it helps to acquire better prediction outcomes compared to a single method [7]. In this work, two different ensemble learning is utilized with an attribute selection approach: one is data perturbation (DP) and the other one is Function Perturbation (FP). The DP method intent to split the learning set into different subsamples and apply the same variable selection method to different subsamples and combine them by using some aggregation method. Next, the FP method intent to apply different variable selection mechanisms to the learning set and combine the output captured from different methods into single methods. In this way, a steady attribute set is generated and further, the attribute set is utilized to NB to improve the prediction in the customer dataset. The work focuses was to decide the finest attribute set, which trends to satisfy NB presumption, and increases the simple Bayes prediction in the consumer data which contains uncertainty in the data. The experimental procedure is conducted using two different proposed methodologies for selecting an attribute set: one is based on data perturbation (homogeneous ensemble) and the other one is function perturbation (heterogeneous ensemble). Furthermore, the selected attribute sets are evaluated using Simple Bayes to assess the results achieved.

2 Literature Survey

The authors address the use of ensemble learning in electricity consumption to predict how much electricity will be used in the office building. In this work, the author develops three ensemble approaches: one is GBR, the second is Adaboost, and the last one is RF. The suggested approaches are experimented using real data and the outcome captured reveals Adaboost performance effectively compared to other approaches [8]. The author executes a study on accurate estimation of software development estimation methods and found that most of the methods utilized are suboptimal in computational intelligence and lack in performance. To acquire accurate estimation in the software development effort, ensemble learning-based methods are carried out. In this study, various homogeneous and heterogeneous methods are developed and the outcome attained is concerted using various nonlinear and linear combiners. The suggested approach has experimented with five software datasets and the outcome captured reveals that ensemble learning performs better compared to individual models [9]. The author enables the use of ensemble learning in landslide susceptibility prediction. The author applies four different ensemble learning approaches to understand the landslide susceptibility prediction: one is stacking, the second is weighted averaging, the third one is simple averaging, and four one is blending.

These approaches apply different classifiers like CNN, RNN, SVM, and LR to avoid the drawbacks of these models and aim to attain reliable results. The approach works in three steps: one is the spatial database, the second is splitting training and test data, and the third one is applying different ensemble learning approaches. The outcome captured reveals that ensemble learning achieves superior results compared to the individual method [10]. The use of Deep Neural Networks (DNN) in various fields like speech, text, and visual has achieved good performance and success. Due to this, the same principles are applied in different subfields of ML, which includes the ensemble learning method. In some recent works, a deep homogeneous ensemble approach is proposed with a large number of models in each layer, and the proposed approach holds high computational costs in classification. To solve the classification problem in the above model, the MULES framework is proposed. In this framework, a small number of heterogeneous models are applied so that resources are used efficiently and Evolutionary Algorithm (EA) based selection approach (NSGA-II algorithm) is applied to enhance the MULES framework performance. The suggested approach has experimented with 33 datasets and the outcome captured reveals the MULES framework performs better [11]. The author addresses the problem of creating fake news and how it will affect political scenarios and other worldwide impacts. To understand the fake news detection in social media, the authors address the use of ensemble learning approaches. This work applies five different ensemble learning approaches and three different ML models to understand fake news in the Urdu language. The suggested approach has experimented on two Urdu news datasets and the outcome captured reveals that ensemble learning achieves superior results compared to the individual method [12]. The author addresses the use of ensemble learning in cryptocurrency prediction. To acquire better prediction in this research, RF and SGBM ensemble learning approach is proposed. The suggested approach is experimented with using BTC, ETH, and XRP coins and the outcome captured reveals the proposed ensemble learning approach performs better in cryptocurrency prediction [13]. The authors address the importance of CAD disease and the prediction of early CAD disease helps to reduce the life risk. To understand CAD disease, the author proposes a heterogeneous ensemble approach. In a heterogeneous ensemble, three different ML models are applied K-NN, RF, and SVM, and the outcomes captured are integrated using a technique like AVEn, MVEn, and WAVEn. The suggested approach has experimented with Z-Alizadeh data and the outcome captured reveals WAVEn achieves superior results compared to other approaches [14]. To build an effective bankruptcy method in this investigation, the author applies two different variable selection techniques (one is based upon the filter approach and the other one is based upon the wrapper approach) and two different classification techniques applied to uncover the better one. The experimental outcome reveals that the wrapper approach (Genetic algorithm) performs intelligently compare to the filter procedure (Information Gain) to adopt the attribute set and the genetic algorithm to NB & SVM acquires the lowest error rate without ensemble learning methods (Boosting and bagging) [15].

3 Proposed Methodology

The work intent to boost the NB performance in the consumer dataset which comprised of redundant and uncertain attributes and further to choose the uncorrelated attributes which satisfy NB premises and improve the performance of the simple Bayes model. To pick the finest attribute set from these consumer databases, an attribute selection procedure is conducted. The attribute selection procedure is performed using: subset assessment and individual assessment [16]. In the individual method based upon dependencies with the target class, attributes are evaluated and ranked. The attributes are ranked high if hold higher dependency and ranked down if hold no dependency on output predictors. In the subset method, attribute subsets are preferred based on the integrating of the ML model and search process. Compare with other methods, subset assessment performs superior and chooses the finest attribute subset, but the problem is the subset assessment is computationally suboptimal. For that reason, the individual assessment approach is carried out for attribute selection. The attribute selection procedure is grouped into three forms: one is filter, the other is the wrapper, and the third is embedded [17]. In this research, a filter approach is carried out, since compared to the other two approaches, this method is computationally optimal and best suitable for huge-dimensional data. However, the problem with filter methods is they do not take out the correlated variables, instead of that, the filter methods sort the attributes intelligently by correlation with the target label. Moreover, the need for choosing the optimal threshold value is an essential one to get the top attribute set from the ranked attribute list captured by the evaluation of the filter method. Furthermore, the outcome captured from the filter approach changes when a different subsample is applied, which is derived from the same learning set. To select the finest steady attribute set for evaluation with NB, ensemble methods are utilized for variable selection. The benefit of using the ensemble method makes to attain a steady attribute set by integrating several weak models into one strong model. Therefore, that the outcome captured is stable with varying subsamples generated from the same learning set. Two different ensemble methods are applied: one is data perturbation (homogeneous ensemble) and the other one is function perturbation (heterogeneous ensemble).

3.1 Ensemble Approaches

In this research, two different ensemble approaches are applied: one is data (homogeneous ensemble) perturbation and the other one is function (heterogeneous ensemble) perturbation.

3.1.1 Data Perturbation (Homogeneous Ensemble)

In the homogeneous ensemble, the identical attribute selector is utilized on dissimilar N subsamples captured from the same learning set L and ends in different V_n same attribute selector outputs. Here, the output captured from the attribute selectors ends in a ranking format, which means the attributes are sorted accordingly to the correlation with the output predictors. The attributes are ranked high if hold higher dependency and ranked down if hold no dependency on output predictors. Then using the aggregation method, the outcomes captured from N models are combined into a single model. Furthermore, using a dissimilar cut off, the best attribute set is captured from the final single model. Here N represents 25 subsamples (which means from the learning set 25 subsamples are generated). The procedure for data perturbation is given in Algorithm 1 and Fig. 1.

Algorithm 1:

```
Input: L-Learning (Training) set, N-number of subsamples, $\tau$-Threshold value
Output: P-Prediction results
Step 1. N subsamples $(s_1 \ldots s_{n)}$ is generated from the learning set L
Step 2. for each n from 1 to N do
Step 3. Generating $V_n$ ranking list by using the same attribute selector on different subsamples n
Step 4. end
Step 5. VL = aggregation of ranking $V_n$ by using some suitable aggregation methods.
Step 6.$VL_\tau$ = select the finest attribute set using the threshold value $\tau$ from the VL
Step 7. Apply NB model to the selected attribute set obtained from $VL_\tau$
Step8. Obtain prediction results P
```

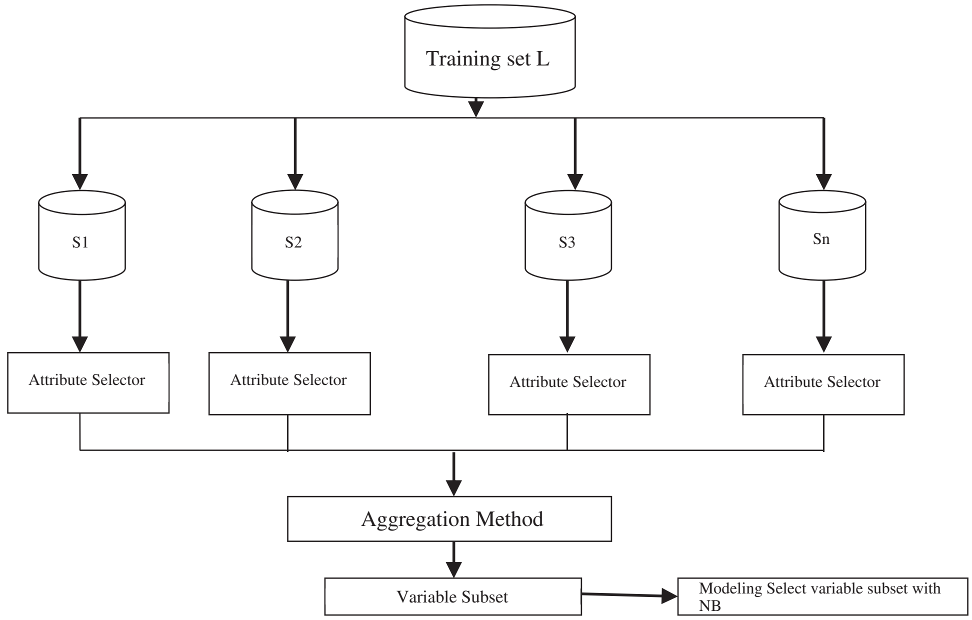

Figure 1: Data perturbation (homogeneous ensemble) approach

3.1.2 Function Perturbation (Heterogeneous Ensemble)

In function perturbation, a dissimilar N attribute selector is utilized on the learning set L and results in different V_n attribute selector outputs. In this approach, a dissimilar attribute selector is utilized on the same learning set L, and the output captured from the different V_n attribute selectors is aggregated into a single model. Furthermore, using different thresholds, the finest attribute subset is captured by the final single model. This approach makes to analyze the weakness and strengths of each method. The procedure for data perturbation is given in Algorithm 2 and in Fig. 2.

Algorithm 2:

```
Input: L-Learning(Training) set, N-number of ranker variable approach, $\tau$-Threshold value
Output: P-Prediction results
Step 1. for each n from 1 to N do
Step 2. Generating $V_n$ ranking list by using attribute selector n
Step 3. end
Step 4. VL = aggregation of ranking $V_n$ by using some suitable aggregation methods.
Step 5.$VL_\tau$ = select the finest attribute set using the threshold value $\tau$ from the VL
Step 6. Apply the NB model to the selected attribute set obtained from $VL_\tau$
Step 7. Obtain prediction results P
```

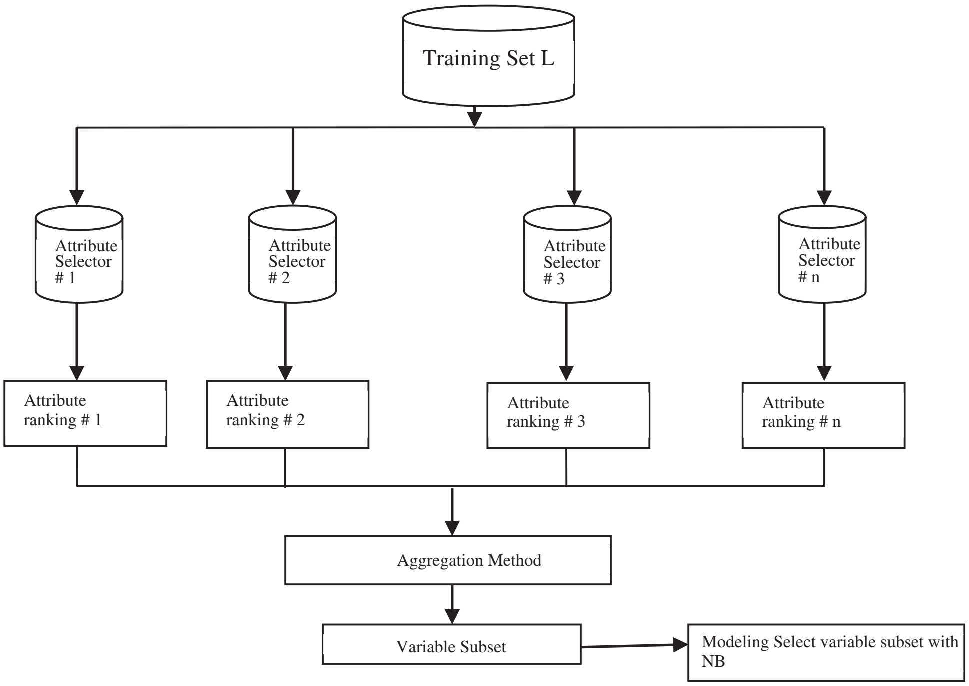

Figure 2: Function perturbation (heterogeneous ensemble)

In both data perturbation and function perturbation approaches, the attribute selectors utilized are based upon ranking the attributes intelligently by correlation with the target label, so there comes in need of using a cuttoff value τ to get the finest steady attribute set from the final ranking list VL_τ. Furthermore, the pertinent variables captured are modeled using the NB classifier to appraise how the performance of the NB is increased using the proposed methodology.

3.2 Variable Selectors Approach

The attribute selection procedure can be arranged into a wrapper, embedded, and filter method depends upon the assessment of the variable set. In this work, five different filter approaches have been considered for this research work.

a) Chi-Square

Chi-Square is one of the most commonly utilized attribute selection approaches. Chi-square is applied to appraise the dependence of the input predictors and the output label, and based upon the correlation measure, the attributes are sorted. Furthermore, with respect to chi-square scores, the attribute subset is considered [18]. The chi-Square formula is represented as

$$x^2 = \frac{(\text{Observed frequency}(O) - \text{Excepted frequency}(E))^2}{\text{Excepted frequency}(E)} \tag{1}$$

b) ReliefF

The relief is developed to overcome the deficiencies associated with relief like multiclass problems, incomplete data, and probability estimation. The quality of the variables is estimated by using the formula given in Eq. (2). Based upon the scores of the variables, the finest attribute set is considered for the assessment using the NB classifier [19].

$$W[V] := W[V] - \sum_{j=1}^{n} \frac{\text{diff}(V, R_i H_j)}{m.n} + \sum_{c=\text{class}(R_i)} \frac{P(C)}{1 - P(\text{class}(R_i)} \sum_{j=1}^{n} \text{diff}(V, R_i, M_j(c))/(m \cdot n) \tag{2}$$

c) Information Gain

Information Gain (IG) is a univariate attribute selection procedure utilized to appraise the dependencies between the output label & the input predictors. IG is based upon an entropy (reduction) measure and assesses how much a variable gives information about the output class. The variables in which partition perfectly will give the maximum information score and the impertinent variables will end up with no information. The variables with high IG values are scored up and the variables with low IG values are scored down. IG of the variable *V* on the target *T* is computed using [20,21]

$$IG(V, T) = H(V) - H\left(\frac{V}{T}\right) \tag{3}$$

where H represents the entropy measure of a random variable and $H(V)$ represents the entropy of *V* and $H(\frac{V}{T})$ represents the entropy of V after observing T.

$$H(V) = -\sum_{i} P(v_i) \log_2(P(v_i)) \tag{4}$$

$$H(V|T) = -\sum_{i} P(t_j) \sum_{j} P(v_i|t_j) \log_2(P(v_i|t_j) \tag{5}$$

d) Gain Ratio

Gain Ratio (GR) is based upon the modification of IG that minimizes the bias value. GR approach considers the size of the branches and the no of branches while picking the variables. GR is computed using

$$GR(V) = \frac{\text{Gain}(V)}{\text{Intrinsic Info}(V)} \tag{6}$$

Intrinsic Information is computed using

$$\text{split info(D)}= -\sum_{j=1}^{v}\left(\frac{|D_j|}{|D|}\right)\log_2\left(\frac{|D_j|}{|D|}\right) \tag{7}$$

Gain value for a variable is measured using

$$Gain(V) = I(D) - E(V) \tag{8}$$

$$I(D) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{9}$$

e) Symmetrical Uncertainty

SU is one of the generally utilized filter-based variable approaches applied to assess the quality of the variables in the target class. Symmetrical Uncertainty assesses the quality of the variables using entropy (information theory). The entropy of the variable *V* is measured by

$$\mathrm{H(V)} = -\sum \mathrm{P(v_i)\log_2(P(v_i))} \tag{10}$$

The entropy of V after observing *T*. measured using

$$\mathrm{H(V|T)} = -\sum_{i} \mathrm{P(t_j)} \sum_{j} \mathrm{P(v_i|t_j)\log_2(P(v_i|t_j)} \tag{11}$$

$\mathrm{P(v_i)}$—prior probabilities(V) and $\mathrm{P(v_i|t_j)}$—posterior probabilities represent *V* given value *T*.

IG of the variable *V* with the target *T* is computed using

$$\mathrm{IG(V,T) = H(V) - H}\left(\frac{\mathrm{V}}{\mathrm{T}}\right) \tag{12}$$

However, IG is biased toward the attributes which have more values. Therefore, SU is indicated as (in which the scores are normalized between the values [0, 1])

$$\mathrm{SU(V,T)} = 2\frac{\mathrm{IG(V/T)}}{\mathrm{H(V) + H(T)}} \tag{13}$$

If SU value is $\mathrm{IG(V|T) = H(V) = H(T)\&SU(V,T) = 1}$, then $V\&T$ are dependent and if SUvalue is $\mathrm{SU(V,T) = 0}$ means not dependent. The feature selectors applied in the research are based upon the ranking of output, which means the feature selector will only rank the features based upon the dependence with respect to output predictors. The attributes are ranked high if hold higher dependency and ranked down if hold no dependency on output predictors. The reason to choose these particular variable selectors is because of 1. Since these five methods are different in metrics and 2. Mostly used by many researchers.

3.3 Aggregation Methodology

The main intention of ensemble methods is to integrate the several outputs captured from the attribute selector into a single output. Depends upon the outcome of the attribute selectors' aggregation methods vary: variable ranking, variable subset, and variable weighting selection. In this research, ranking aggregation is carried out, since the output captured from both ensemble methods is based on ranking. Different types of ranking aggregation are found in practice and detailed information regarding aggregation methods is given below [22].

1. **Mean**: $\mathrm{VL} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{VL}_i$ $\{v_1, v_2, v_3, \ldots., v_n\}$ by total n. Mean aggregation is based upon arithmetic operations. In mean aggregation, an average value is selected from the ranking outcome. $\mathrm{VL} = \{v_1, v_2, v_3 \ldots\ldots v_n/n\}$

2. **Geomean**: $\mathrm{VL} = (\prod_{i=1}^{n} (v_i))^{1/n}$ $\sqrt[n]{v_1 v_2 v_3 \ldots v_n}$ GeoMean aggregation is based upon arithmetic operations. In this aggregation, the Geo mean average value is selected from the ranking outcome. In the mean approach, the ranking output is added and the average is taken. However, in the geo mean approach, the ranking output is multiplied and the average is taken.

3. **Min**: $\mathrm{VL} = \mathrm{Min}\ \{v_1, v_2, v_3, \ldots, v_n\}$. Min aggregation is based upon arithmetic operations. In this aggregation, the minimum value is selected from the ranking outcome.

4. **Median**: $\mathrm{VL} = \mathrm{Median}\ \{v_1, v_2, v_3, \ldots, v_n\}$. Median aggregation is based upon arithmetic operations. In this aggregation, the median value is selected from the ranking outcome.

3.4 Threshold Values

Since in this study, the attribute selection approach was used only to sort the features correspondingly in connection with the output label. Furthermore, to make the finest attribute set, we need to use a cut-off value to pick the attribute set from the final ranked list captured after the aggregation procedure. In this study, three different threshold values were utilized to minimize the data and to pick the best appropriate attribute set [23]. The threshold values used are:

1. $\mathbf{log_2(n)}$: In this threshold value $\log_2(n)$ determine the number of attribute sets selected from the final ranking list VL_τ

2. **10 percentage**: From this threshold value, the top 10 percentage attribute set is selected from the final ranking list VL_τ

3. **50 percentage**: Using this threshold value, the top 50 percentage attribute set is selected from the final ranking list VL_τ

4 Experimental Results

The first methodology is based upon data perturbation (homogeneous ensemble) − NB and the second methodology is based upon function perturbation (heterogeneous ensemble) − NB and the third one Naive Bayes without using any method (Standard Naive Bayes) is studied and experimented.

4.1 Dataset and Validity Scores

The customer dataset utilized in the experiment consists of 45211 instances with 16 input predictors and one target class (with two labels). The customer dataset analyzed is affiliated with the banking sector and the purpose of the prognosis is to determine whether the customer will take up the bank service (term deposit) or not. This prognosis helps to understand the customer pattern and further analysis decisions assist to enhance customer satisfaction with enterprises and try to make a profit with these customers. The empirical outcome captured from the three different methodologies is compared and analyzed using validity scores. The validity scores utilized in the study are accuracy, recall, FPR, precision, FNR, and specificity.

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN} \tag{14}$$

$$\text{Sensitivity} = \frac{TP}{P} = \frac{TP}{FN + TP} \tag{15}$$

$$\text{Specificity} = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{16}$$

$$\text{Precision} = \frac{TP}{FP + TP} \tag{17}$$

$$FNR = 1 - TPR \tag{18}$$

$$\text{FPR} = 1 - TNR \tag{19}$$

4.2 Experimental Procedure

The experimental procedure is performed for three various methodologies: a) the first methodology is based upon data perturbation (homogeneous ensemble) − NB, b) the second methodology is based upon function perturbation(heterogeneous ensemble) − NB and c) the third methodology Naive Bayes without using any method (Standard Naive Bayes).

In data perturbation (homogeneous ensemble) − NB, N = 25 subsamples are generated from the learning set L and the identical attribute selector is utilized on different 25 subsamples. This results in different $\mathrm{V_n}$ $\{\mathrm{v_1, v_2, v_3, \ldots, v_n}\}$ identical attribute selector outputs and the outcomes from the homogeneous ensemble are in ranking format. Furthermore, different aggregation methodologies (Mean, Geomean, Min, Median) are executed to integrate several outputs into one strong output VL = $(\mathrm{v_1, v_2, v_3, \ldots, v_n})$. In the data perturbation filter procedure is utilized, it sorts the features correspondingly in connection with the output label and further to pick the finest attribute set needs a threshold value. Here, three different threshold values ($\mathrm{VL}_\tau = \log_2(\mathrm{n})$, 10, 50) percentages are performed to examine how the outcome changes correspondingly to the selection of the cutoff value. Likewise, the same procedure is utilized for five different filter approaches. The filter approach utilized in the study is chi-square, IG, GR, reliefF, and symmetrical uncertainty. Furthermore, the attribute set captured from the data perturbation (homogeneous ensemble) methodology is appraised using Naive Bayes to compute the performance of the proposed approach using the validity scores.

In function perturbation (heterogeneous ensemble) − NB, the learning set L is utilized with different variable selection technique and results in different $\mathrm{V_n}$ $\{\mathrm{v_1, v_2, v_3, \ldots, v_n}\}$ attribute selector outputs and the outcomes from the heterogeneous ensemble are in ranking format. Furthermore, different aggregation methodologies (Mean, Geomean, Min, Median) are carried out to integrate the multiple outputs into a single output $VL = (v_1, v_2, v_3, \ldots, v_n)$. In function perturbation filter procedure is performed, it only sorts the attributes correspondingly in connection with the class label and further to pick the finest attribute set needs a cut-off value. Here, three different threshold values $(VL_\tau = \log_2(n), 10, 50)$percentages are performed to appraise how the outcome changes correspondingly to the selection of the cut-off value. The filter approach applied in the study is chi-square, IG, GR, reliefF, and symmetrical uncertainty. Furthermore, the attribute set obtained from the function perturbation (heterogeneous ensemble) methodology is evaluated using Naive Bayes to compute the performance of the proposed approach using the validity scores.

3. In the third methodology, Naive Bayes is applied directly to the customer dataset without using any ensemble methods or preprocessing techniques. The outcome captured is computed and analyzed using the validity scores.

4. The outcomes obtained from the proposed methodology were compared and analyzed with Standard Naive Bayes using validity merits.

4.3 Results of Data Perturbation − NB and Standard Naive Bayes

Tab. 1. 10% threshold value is utilized to pick the attribute set from the final sorted attributes captured using different combination methodology to model with Naive Bayes, and evaluation of Naive Bayes without using any preprocessing approach with the summary of FPR, FNR, sensitivity, specificity, accuracy, and precision.

Table 1: Using 10% threshold value to pick the attribute set

	Aggregation	Accuracy	FPR	Sensitivity	FNR	Precision	Specificity
Homogeneous ensemble (IG) − NB	Median	89.19	0.0301	30.3	69.7	57.2	96.99
	Mean	89.19	0.0301	30.3	69.7	57.2	96.99
	Geo mean	89.19	0.0301	30.3	69.7	57.2	96.99
	Min	89.14	0.016	1.9	81	61.7	98.4
Homogeneous ensemble (GainRatio) − NB	Median	89.28	0.014	0.185	81.5	64.7	98.6
	Mean	89.19	0.0301	30.3	69.7	57.2	96.99
	Geo mean	89.19	0.0301	30.3	69.7	57.2	96.99
	Min	89.19	0.0301	30.3	69.7	57.2	96.99
Homogeneous ensemble (Symmetrical Uncertainty) − NB	Median	88.66	0.0305	26.1	73.9	53.2	96.95
	Mean	89.19	0.0301	30.3	69.7	57.2	96.99
	Min	89.19	0.0301	30.3	69.7	57.2	96.99
	Geo mean	89.19	0.0301	30.3	69.7	57.2	96.99
Homogeneous ensemble (ReliefF) − NB	Median	88.34	0.006	4.7	95.3	0.52	99.4
	Mean	88.18	0.014	9.4	90.6	47.6	98.6
	Min	88.18	0.014	9.4	90.6	47.6	98.6
	Geo mean	88.18	0.014	9.4	90.6	47.6	98.6
Homogeneous ensemble (chi−Square) − NB	Median	88.66	0.0305	26.1	73.9	53.2	96.95
	Mean	89.19	0.0301	30.3	69.7	57.2	96.99
	Min	89.19	0.0301	30.3	69.7	57.2	96.99
	Geo mean	89.19	0.0301	30.3	69.7	57.2	96.99
Standard NB		88.0073	0.074	52.8	47.2	48.8	92.6

Tab. 2. 50% threshold value is utilized to pick the attribute set from the final sorted attributes captured using different combination methodology to model with Naive Bayes, and evaluation of Naive Bayes without using any preprocessing approach with the summary of FPR, FNR, sensitivity, specificity, accuracy, and precision.

Tab. 3. $\log_2(n)$ threshold value is utilized to pick the attribute set from the final sorted attributes captured using different combination methodology to model with Naive Bayes, and evaluation of Naive Bayes without using any preprocessing approach with the summary of FPR, FNR, sensitivity, specificity, accuracy, and precision.

Table 2: Using 50% threshold value to pick the attribute set

	Aggregation	Accuracy	FPR	Sensitivity	FNR	Precision	Specificity
Homogeneous ensemble (IG) – NB	Median	88.8235	0.055	45.9	54.1	52.6	95.4
	Mean	88.8235	0.055	45.9	54.1	52.6	95.4
	Geo mean	88.8235	0.055	45.9	54.1	52.6	95.4
	Min	89.2681	0.049	95	54.7	55	95.3
Homogeneous ensemble (GainRatio) – NB	Median	88.4	0.064	38.5	61.5	50.6	93.6
	Mean	88.82	0.055	45.9	54.1	52.6	94.5
	Geo mean	89.77	0.052	43.1	56.9	52.5	94.8
	Min	88.82	0.055	45.9	54.1	52.6	94.5
Homogeneous ensemble (Symmetrical Uncertainty) – NB	Median	88.4	0.05	38.5	61.5	50.6	95
	Mean	88.82	0.055	45.9	54.1	52.6	94.5
	Min	88.46	0.056	43.1	56.9	50.8	94.4
	Geo mean	88.82	0.055	45.9	54.1	52.6	94.5
Homogeneous ensemble (ReliefF) – NB	Median	89	0.032	29.6	71.4	55.6	96.8
	Mean	89.11	0.034	32.3	67.7	52.6	96.6
	Min	88.94	0.032	29.6	70.4	55.1	96.8
	Geo mean	89.11	0.034	32.3	67.7	52.6	96.6
Homogeneous ensemble (chi – Square) – NB	Median	89.27	0.044	41.5	58.5	52.6	95.6
	Mean	88.82	0.055	45.9	54.1	52.6	94.5
	Min	89.82	0.05	42.2	57.8	52.8	95
	Geo mean	88.82	0.055	5.9	54.1	52.6	94.5
Standard NB		88.0073	0.074	52.8	47.2	48.8	92.6

4.4 Results of Function Perturbation (Heterogeneous Ensemble) and Standard Naive Bayes

Tab. 4. 10% threshold value is utilized to pick the attribute set from the final sorted attributes captured using different combination methodology to model with Naive Bayes, and evaluation of Naive Bayes without using any preprocessing approach with the summary of FPR, FNR, sensitivity, specificity, accuracy, and precision.

Tab. 5. 50% threshold value is utilized to pick the attribute set from the final sorted attributes captured using different combination methodology to model with Naive Bayes, and evaluation of Naive Bayes without using any preprocessing approach with the summary of FPR, FNR, sensitivity, specificity, accuracy, and precision.

Tab. 6. $\log_2(n)$ threshold value is utilized to pick the attribute set from the final sorted attributes captured using different combination methodology to model with Naive Bayes, and evaluation of Naive Bayes without using any preprocessing approach with the summary of FPR, FNR, sensitivity, specificity, accuracy, and precision.

Table 3: Using $\log_2(n)$ threshold value to pick the attribute set

	Aggregation	Accuracy	FPR	Sensitivity	FNR	Precision	Specificity
Homogeneous ensemble (IG) – NB	Median	89.2703	0.041	39.4	60.6	55.9	95.87
	Mean	89.2703	0.041	39.4	60.6	55.9	95.87
	Min	89.5733	0.034	36.4	63.6	58.8	96.6
	Geo mean	89.2703	0.041	39.4	60.6	55.9	95.87
Homogeneous ensemble (GainRatio) – NB	Median	89.09	0.041	37.4	62.6	55	95.9
	Mean	88.4	0.05	38.5	61.5	50.6	95
	Min	88.4	0.05	38.5	61.5	50.6	95
	Geo mean	88.4	0.05	38.5	61.5	50.6	95
Homogeneous ensemble (Symmetrical Uncertainty) – NB	Median	89.09	0.041	37.4	62.6	55	95.9
	Mean	88.4	0.05	38.5	61.5	50.6	95
	Min	88.4	0.05	38.5	61.5	50.6	95
	Geo mean	88.4	0.05	38.5	61.5	50.6	95
Homogeneous ensemble (ReliefF) – NB	Median	88.14	0.015	09.3	90.7	46.6	98.5
	Mean	88.94	0.031	28.7	71.3	55.3	96.6
	Min	88.94	0.031	28.7	71.3	55.3	96.9
	Geo mean	88.94	0.031	28.7	71.3	55.3	96.6
Homogeneous ensemble (chi – Square) – NB	Median	89.09	0.041	37.4	62.6	55	95.9
	Mean	89.27	0.042	39.4	60.6	55.9	95.8
	Min	89.27	0.042	39.4	60.6	55.9	95.8
	Geo mean	89.27	0.042	39.4	60.6	55.9	95.8
Standard NB		88.0073	0.074	52.8	47.2	48.8	92.6

Table 4: Using 10% threshold value to pick the attribute set

	Aggregation	Accuracy	FPR	Sensitivity	FNR	Precision	Specificity
Heterogeneous Ensemble − NB	Median	88.77	0.029	0.267	0.732	54.1	97
	Mean	89.19	0.0301	0.303	0.697	57.2	96.99
	Min	89.19	0.0301	0.303	0.697	57.2	96.99
	Geo mean	89.19	0.0301	0.303	0.697	57.2	96.99
Standard NB		88.0073	0.074	0.528	0.472	48.8	92.6

Table 5: Using 50% threshold value to pick the attribute set

	Aggregation	Accuracy	FPR	Sensitivity	FNR	Precision	Specificity
Heterogeneous Ensemble − NB	Median	88.8235	0.0549	0.459	0.541	0.526	0.9451
	Mean	88.8235	0.055	0.459	0.541	52.6	94.5
	Min	89.1929	0.0523	0.471	0.529	54.4	94.76
	Geo mean	89.1929	0.0523	0.471	0.529	54.4	94.76
Standard NB		88.0073	0.074	0.528	0.472	48.8	92.6

Table 6: Using $\log_2(n)$ threshold value to pick the attribute set

	Aggregation	Accuracy	FPR	Sensitivity	FNR	Precision	Specificity
Heterogeneous Ensemble − NB	Median	89.2703	0.041	0.394	0.606	0.559	0.9587
	Mean	89.2703	0.042	0.394	0.606	55.9	95.8
	Min	89.79	0.035	0.391	0.609	59.8	96.5
	Geo mean	89.2703	0.042	0.394	0.606	55.9	95.8
Standard NB		88.0073	0.074	0.528	0.472	48.8	92.6

4.5 Result Discussion

The experimental procedures performed for the three various methodologies are illustrated in Tabs. 1–6. The outcome illustrated in Tabs. 1–3 shows the homogeneous ensemble − NB and standard NB methodology and the outcomes presented in Tabs. 4–6 indicates the heterogeneous ensemble − NB and standard NB methodology. Tab. 1 illustrates the results of a homogeneous ensemble NB using a 10% cut-off value utilized to pick the finest attribute set from the final ranking variables captured from the different combination methodology with the model with Naive Bayes. The results illustrate that using a 10% threshold value, the best performance is acquired when compared to standard NB. Homogeneous ensemble (Gain Ratio) − NB acquires 89.28 higher predictions when compared to standard NB which gets only 88.0073 predictions. However, homogeneous ensemble (ReliefF) − NB acquires 88.14 lesser predictions compared to other Homogeneous ensemble NB approaches, but when compared to Standard NB it is a superior prediction. Tab. 2 indicates the outcome of a homogeneous ensemble NB with a 50% cut-off value utilized to pick the finest attribute set from the final ranking variables list captured from the different combination methodology with the model with Naive Bayes. The results illustrate that with a 50% threshold value, the best performance is acquired when compared to standard NB. Homogeneous ensemble (Gain Ratio) − NB acquires 89.77 higher predictions when compared to standard NB which gets only 88.0073 predictions. However, homogeneous ensemble (Symmetrical Uncertainty) − NB acquires 88.40 lesser predictions compared to other Homogeneous ensemble − NB approaches, but when compared with Standard NB it is a superior prediction. Tab. 3 indicates the outcome of a homogeneous ensemble − NB using $\log_2(n)$ cut-off value utilized to pick the finest attribute set from the final ranking variables list captured from the different combination methodology to model with Naive Bayes. The outcome indicates that using a 10% threshold value, the prominent performance is achieved when compared to standard NB. Heterogeneous ensemble (attribute set obtained from mean, Geomean, min, aggregation using 10% cut-off value) − NB acquires 89.19 higher predictions when compared to standard

NB which gets only 88.0073 predictions. However, heterogeneous ensemble (attribute set captured from median aggregation using 10% cut-off value) − NB acquires 88.77 lesser predictions compared to other heterogeneous ensembles (attribute set obtained from the mean, Geomean, min aggregation using 10% cut-off value) − NB approach, but when compared with Standard NB it is superior prediction. Tab. 5 indicates the outcome of a heterogeneous ensemble NB using a 50% cut-off value utilized to pick the finest attribute set from the final ranking variables list captured from the different combination methodologies to model with Naive Bayes. The results illustrate that using a 50% threshold value, the best performance is achieved when compared to standard NB. Heterogeneous ensemble (attribute set captured from Geomean, min aggregation using 10% cut-off value) − NB acquires 89.1929 higher predictions when compared to standard NB which gets only 88.0073 predictions. However, heterogeneous ensembles(attribute set captured from mean and median aggregation using 10% cut-off value) − NB acquires 88.8235 lesser predictions compared to other heterogeneous ensembles (attribute set obtained from the mean, Geomean, min aggregation using 10% cut-off value) − NB approach, but when compared with standard NB it is superior prediction. Tab. 6 indicates the outcome of a heterogeneous ensemble NB using $\log_2(n)$ cut-off value utilized to pick the finest attribute set from the final ranking variables list captured from the different combination methodology to model with Naive Bayes. The results illustrate that using $\log_2(n)$ threshold value the best performance is acquired when compared to standard NB. Heterogeneous ensemble (attribute set captured from min aggregation using 10% cut-off value) − NB acquires 89.79 higher predictions when compared to standard NB which gets only 88.0073 predictions.

5 Conclusion

Customer analysis is one of the important challenges of each enterprise to understand the behavior of consumers and makes them to develop different marketing plans to increase the consumer satisfaction & the profit of the enterprise's concern. However, the customer dataset captured and stored by the CRM system exists in huge dimensional, and the customer data captured is comprised of redundant and uncertain variables. Furthermore, these consumer databases should not be modeled with the Naive Bayes directly. Since the Naive Bayes model holds an important premise, that is, the dataset applied should not be comprised of correlated variables (the variables applied should be conditional independent with other variables) and the variables in the dataset should be equal. Moreover, due to the modeling of huge dimensional variables with NB increases the complexity of the model. To overcome the above drawbacks, an attribute selection procedure is carried out. In the attribute selection, the filter approach is considered since compared to the wrapper approach, the filter method is best suitable for the large dataset and works fast compared to the wrapper. However, the outcome produced by the filter approach is not satisfying and the outcome is unsteady due to changes in the subsample set. The first methodology is based upon data perturbation (homogeneous ensemble) − NB and the second methodology is based upon function perturbation (heterogeneous ensemble) − NB and the third one Naive Bayes without using any method (Standard Naive Bayes). The experimental procedures performed for three different methodologies are described in Tabs. 1–6. The results described in Tabs. 1–3 show the homogeneous ensemble − NB and standard NB methodology and the results described in Tabs. 4–6 show the heterogeneous ensemble−NB and standard NB methodology. The outcome discloses the proposed approaches perform better compared to standard NB. In standard Naive Bayes acquires a maximum prediction of 88.0073, where the homogeneous ensemble gets a maximum prediction of 89.82 and the heterogeneous ensemble gets a maximum prediction of 89.79. Therefore, from the results, we can witness that the proposed ensemble methodology performs superior compared to Naive Bayes. Further future work can be extended by using an automatic selection of the variable subsets without using any threshold value or integration threshold value within the variable selection procedure itself.

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] S. A. Bin Nashwan and H. Hassan, "Impact of customer relationship management on customer satisfaction and loyalty: A systematic review," *Journal of Advanced Research in Business and Management Studies*, vol. 6, no. 1, pp. 86–107, 2017.

[2] M. M. Migdadi, "Knowledge management, customer relationship management, and innovation capabilities," *Journal of Business & Industrial Marketing*, vol. 36, no. 1, pp. 111–124, 2020.

[3] T. L. Tan and D. T. Dai, "Successful factors of implementation electronic customer relationship management on e-commerce company," *American Journal of Software Engineering and Applications*, vol. 6, no. 5, pp. 121–127, 2017.

[4] P. N. Diaz, F. Ortega, E. Cobos and R. L. Cabrera, "A collaborative filtering approach based on naive bayes classifier," *IEEE Access*, vol. 7, pp. 108581–108592, 2019.

[5] F. Degenhardt, S. Seifert and S. Szymczak, "Evaluation of variable selection methods for random forests and omics data sets," *Briefings in Bioinformatics*, vol. 2, pp. 492–503, 2017.

[6] G. Manikandan and S. Abirami, "Feature selection is important: State-of-the-art methods and application domains of feature selection on high-dimensional data," in: *Applications in Ubiquitous Computing*, 1st ed., Cham: Springer, pp. 177–196, 2021.

[7] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta *et al.*, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, no. 6, pp. 47–58, 2021.

[8] T. Pinto, I. Praca, Z. Vale and J. Silva, "Ensemble learning for electricity consumption forecasting in office buildings," *Neurocomputing*, vol. 423, pp. 747–755, 2021.

[9] O. Mahmoud Elish, T. Helmy and M. I. Hussain, "Empirical study of homogeneous and heterogeneous ensemble models for software development effort estimation," *Mathematical Problems in Engineering*, vol. 2013, no. 6, pp. 1–22, 2013.

[10] Z. Fang, Y. Wang, L. Peng and H. Hong, "A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping," *International Journal of Geographical Information Science*, vol. 35, no. 2, pp. 321–347, 2021.

[11] T. T. Nguyen, N. V. Pham, M. T. D. Manh, A. V. Luong, J. Mccall *et al.*, "Multi-layer heterogeneous ensemble with classifier and feature selection," in *Proc. GECCO*, Cancún Mexico, pp. 725–733, 2020.

[12] M. Pakhter, J. Zheng, F. Afzal, H. Lin, S. Riaz *et al.*, "Supervised ensemble learning methods towards automatically filtering Urdu fake news within social media," *PeerJ Computer Science*, vol. 7, no. 5, pp. e425, 2021.

[13] V. Derbentsev, V. Babenko, K. Khrustalev, H. Obruch and S. Khrustalovac, "Comparative performance of machine learning ensemble algorithms for forecasting crypto currency prices," *International Journal of Engineering*, vol. 34, no. 1, pp. 140–148, 2021.

[14] V. Durgadevi and R. Karthikeyan, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset," *Computer Method and Programs in Biomedicine*, vol. 198, pp. 1–13, 2021.

[15] W. C. Lin, Y. H. Lu and C. F. Tsai, "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Systems*, vol. 36, no. 2, pp. 1–8, 2018.

[16] M. Di Mauro, G. Galatro, G. Fortino and A. Liotta, "Supervised feature selection techniques in network intrusion detection: A critical review," *Engineering Applications of Artificial Intelligence*, vol. 101, article. 104216, 2021.

[17] E. Szmidt, J. Kacprzyk and P. Bujnowski, "Attribute selection for sets of data expressed by intuitionistic fuzzy sets," in *Proc. FUZZ-IEEE*, Glasgow, United Kingdom, pp. 1–7, 2020.

[18] S. Bahassine, A. Madani, A. Sarem and M. Kissi, "Feature selection using an improved chi-square for arabic text classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2018.

[19] R. J. P. Mendoza, D. Rodriguez and L. DeMarcos, "Distributed relief-based feature selection in Spark," *Knowledge Information Systems Journal*, vol. 57, no. 1, pp. 1–20, 2018.

[20] S. W. Sihwi, I. P. Jati and R. Anggrainingsih, "Twitter sentiment analysis of movie reviews using information gain and naïve bayes classifier," in *Proc. ISEMANTIC*, Semarang, Indonesia, pp. 190–195, 2018.

[21] S. Venkataraman and S. Rajalakshmi, "Optimal and novel hybrid feature selection framework for effective data classification," in *Advances in Systems, Control, and Automation*, vol. 442. Singapore: Springer, 2018.

[22] R. Salman, A. Alzaatreh, H. Sulieman and S. Faisal, "A bootstrap framework for aggregating within and between feature selection methods," *Entropy*, vol. 23, no. 2, pp. 1–21, 2021.

[23] M. I. Prasetiyowati, N. U. Maulidevi and K. Surendro, "Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest," *Journal of Big Data*, vol. 8, no. 84, pp. 1–22, 2021.