International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

Heterogeneous Ensemble Variable Selection To Improve Customer Prediction Using Naive Bayes Model

[1]Siva Subramanian.R, [2]Girija.P, [3]Dr.Anuradha.M,[4] Dr.Dinesh M G,[5] Dr.Aswini.J,[6] Divya.P
[1]Associate Professor, Dept of CSE, RMK College of Engineering and Technology, Puduvoyal, India
Puduvoyal, Tamilnadu, India
sivasubramanian12@yahoo.com
[2]Assistant Professor, Vel Tech Rangarajan Dr.Sagunthala R&D Insitute of Engineering and Technology
Avadi, Tamilnadu, India
pygirijads@gmail.com
[3]Professor, S.A.Engineering College
Poonamallee, Tamilnadu, India
anuparini@gmail.com
[4]Associate Professor, Dept of CSE, Easa College of Engineering and Technology
Coimbatore, Tamilnadu, India
dineshbabu.mg@gmail.com
[5]Professor, Dept of AI & ML, Saveetha Engineering College(Autonomous), Affiliated to Anna University
Chennai, Tamilnadu, India
aswini.jayaraman@gmail.com
[6]Assistant Professor, Dept of CSE, IFET College of Engineering
Villupuram, Tamilnadu, India
Divyait2009@gmail.com

Abstract—The analysis of customer patterns and behaviours is essential for all businesses, as the customer is the sole source of revenue. Understanding customer patterns and behavior enables businesses to enhance their business processes and customer happiness. The availability of voluminous client datasets within organizations facilitates efficient customer analysis. Yet, the inclusion of interrelated, irrelevant, as well as missing factors leads to a poor forecast of the dataset. Feature selection techniques are investigated in order to handle the problem. Objective of feature selection is to pick the pertinent variables from a complete set of associated, irrelevant, and missing variables. In general, FS is classified into 3 types: filter, wrapper, & hybrid method. The filter method is a quick one, but the variables used are ineffective. Similarly, a wrapper method is effective yet computationally inefficient. In this study, an ensemble feature selection strategy is presented and tested to circumvent the issue with these feature selections. There are two techniques to ensemble FS: one is homogenous and the other is heterogeneous. This study employs a heterogeneous ensemble feature selection method. In the suggested method, the learning dataset is applied to five distinct filter FS approaches, and the ranked attributes that result are aggregated using two distinct methods: the mean method and the min method. Relevant variables are chosen to further build the final sorted qualities using the cut off value as a guide. As the HEVS technique's filter approach simply ranks the variables, it is necessary to select the variable subset cut off value. The experimental technique is conducted from two distinct vantage points: Heterogeneous ensemble variable selection with Naive Bayes and Naive Bayes without variable selection. In the end, the outcomes that were obtained via the use of the two different approaches are compared using different factors. The experimental results demonstrate that the suggested HEVS method outperforms the usual Naive Bayes model. As relevant variables are included when modeling using NB, the computational complexity of this proposed methodology is also minimized.

Keywords- Customer Pattern; Naive Bayes; Ensemble; Feature Selection; Filter; Wrapper

I. INTRODUCTION

For a firm to thrive in today's competitive corporate environment, it must deliver superior service that efficiently satisfies its customers. Businesses can't achieve financial success in today's market if they fail to satisfy their customers. It is possible for businesses to better understand their customers by conducting analyses of the patterns and behaviours of their customers. Some of the advantages of doing a customer analysis include the following: 1. Specific customer value. 2. Customer experience. 3. Customer retention. 4. Increased quality. 5. Consumer acquisition [1]. In today's highly connected and technologically evolved society, the amount of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

data being collected is staggering. These datasets make it possible to unearth previously concealed information about customers, which, once processed, can be put to productive use in corporate settings. Yet, the customer datasets that have been collected are of an enormous number, and they might include correlated, irrelevant, or missing variables. This issue arises because of the many methods used to obtain the data, as well as the high-dimensional nature of the information that was gathered. It is not possible to do an effective analysis of the consumer with these data. The utilization of FS is being contemplated as a potential solution to the problem. One of the methods that fall under the category of pre-processing is known as feature selection. The aim of this method is to pick the relevant variables from the entire dataset by omitting correlated, irrelevant, and missing variables. The selection of the variables is done by a certain statistical or evaluative technique. In general, feature selection may be broken down into three different types: the first kind is a filter, the second type is a wrapper, and the third type is embedded. The findings that are obtained using a straightforward FS method like a filter, notwithstanding their speed and effectiveness, are ineffective. Filter FS has several benefits, the most important of which are that it is 1) independent of the model being used, 2) reliant on the features being used, 3) computationally efficient, and 4) uses statistical approaches [2]. The filter FS does not take into account the feature dependencies, which is one of its disadvantages. Similarly, the wrapper FS computes the model performance by making use of every feature subset possible. Wrapper feature sets have several benefits, the most important of which are reduced over fitting, consideration of feature dependencies, interaction with the classifier, and improved classification results. Wrapper file systems have the disadvantage of being computationally demanding. In this research, ensemble feature selection is presented and experimented with as a means of overcoming the drawbacks associated with the feature selection processes for filters and wrappers. One of the two types of ensemble feature selection is referred to be homogenous (data perturbation), while the other type is referred to as heterogeneous (function perturbation). When we talk about homogenous ensemble feature selection, we mean that the same FS strategy is used on various set of the training data, and then we integrate the multiple outputs into a single one by employing an aggregation method. In a similar vein, heterogeneous ensemble FS refers to a situation in which numerous feature selection strategies are utilized on the same training dataset, and the results of these strategies are combined into a single output by the application of an aggregation technique. In this research, a new method known as the heterogeneous ensemble variable selection (HEVS) technique is developed and tested. This method is based on the heterogeneous ensemble approach. Within the framework of the HEVS technique, the training data is put through five distinct feature selection procedures. In this section, filter feature selection is carried out. In addition, the outputs that were taken from the various feature selections were aggregated into a single output utilizing a variety of different techniques. In conclusion, a relevant subset of variables is obtained by utilizing some appropriate threshold value. Then, a Naive Bayes model is constructed using the captured variables subset. The experiment is done out using three distinct customer datasets provided from the UCI library. The experiment is carried out from two perspectives: one employs the suggested HEVS approach, while the other employs the Standard Naive Bayes methodology. These two methodologies are compared to one another. The results that were produced from the two separate methods are projected and compared utilizing a variety of validity metrics. The rest of the work is presented: 2. A Review of the Relevant Literature, 3. Methods, 4. Experiment Results, and 5. Conclusion.

II. LITERATURE SURVEY

[3], the significance of feature selection in data mining is emphasized. The author discusses the limitations of single feature selection approaches and proposes an ensemble methodology as a solution to this problem. The ensemble approach combines multiple methods and procedures to obtain more reliable and stable results compared to other feature selection techniques. The experiment was conducted using two different datasets: a small-scale and a large-scale dataset. The SVM model was utilized to model a subset of the extracted features. The analysis of the results indicated that the ensemble learning method outperformed the single-feature selection method. Overall, the study highlights the effectiveness of ensemble learning in feature selection for data mining tasks. **[4],** in this research, the author talks about how important the mechanism for selecting features is. Based on the FS, the author proposes and tests a hybrid ensemble FS. In the proposed method, the first step is to use the CDF-g algorithm to select the primary subset of variables. The second step is to use data perturbation to capture the second subset of variables. In the second step, relevant variables are pulled out of the second subset of variables using function perturbation. UCI's phishing dataset is used for the experimental procedure. Experimental results show that the hybrid ensemble FS selects the stable relevant subset and improves the prediction of the classifier in a system for detecting phishing. **[5],** conducted a study to investigate the significance of feature selection in high dimensional domain datasets such as image, biomedical, and document analysis. It is known that the utilization of all these data with Machine Learning (ML) leads to poor prediction and high computational time. To address this issue, the author examined the ensemble FS approach. In this approach, seven

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

different ranking methods are considered to rank the features. Then, the aggregation technique is utilized to combine the results obtained from these ranking methods. By setting a suitable threshold value, the final variable subset is derived. The experimental procedure involved various datasets, and a stability analysis was conducted to evaluate the stability of the captured feature subsets. The results of the experiments demonstrate that the ensemble FS approach selects a stable and relevant subset of features, which in turn improves the prediction accuracy of the classifier. This research emphasizes the significance of feature selection in high dimensional domain datasets and highlights the effectiveness of ensemble FS in obtaining stable and relevant feature subsets**. [6],** the author talks about how important credit scores are for each customer in the financial sector. The author implies that credit scores are based on information from the past. But the fact that these datasets contain correlated, irrelevant, and missing data makes it hard to use them to make decisions. In this research, the author suggests a hybrid model that combines FS with ensemble learning as a way to solve the problem. The first step is to do some preliminary work and assign ranks. Then, an ensemble FS is modeled with data that has already been processed, and in the last step, a variable subset is modeled with a multilayer ensemble framework. Different customer credit datasets are used for the experimental procedure, and the results show that the proposed approach works better. **[7],** the author addresses the importance of breast cancer classification in this study, and based on the research the author proposes an ensemble filter FS approach (EnSNR). The proposed method uses a combination of SNR and entropy evaluation, which are two different ways to measure the value of a variable. The genetic algorithm is used to model the subset of variables that the EnSNR method produces. The procedure for the experiment is done with a microarray dataset that has 50,739 attributes. The results of the experiments show that the proposed method is a better way to choose the important variables, get rid of the unimportant ones, and increase the efficiency of cancer classification. **[8]**, the author performs an in-depth analysis of the ensemble learning technique and suggest how it might be used to improve feature selection and classification accuracy. The ensemble technique is founded on the premise that combining many model results is always superior to a single model result. The ensemble technique in earlier days is only used to improve categorization. Nonetheless, due to its efficacy, the ensemble method is utilized alongside the feature selection method. The objective of this work is to present readers with an in-depth understanding of ensemble learning, its fundamental concepts, and its future tendencies. **[9],** the significance of customer analysis in enterprises is emphasized. The authors suggest that efficient analysis of customers can aid enterprises in improving customer satisfaction and their products or services. The presence of correlated and irrelevant variables can result in poor prediction and a violation of the Naive Bayes (NB) assumption. To address this issue, the authors propose an ensemble learning approach and perform experiments to evaluate its effectiveness. The ensemble learning approach includes both homogenous and heterogeneous methodologies. In homogenous ensemble FS, the same feature selection approach is applied to different subsets of training data. In contrast, heterogeneous ensemble FS involves applying different feature selection approaches to the same training dataset and integrating the multiple outputs using an aggregation method. The experimental procedure involved using a customer dataset captured from UCI. The results obtained indicate that the proposed ensemble learning approach performs better in addressing the issue of correlated and irrelevant variables compared to the traditional methods. This research highlights the significance of ensemble learning in improving the prediction accuracy of customer analysis in enterprises. **[10],** the author discusses the significance of IDS in the network security area and argues that IDS efficiency should be improved. The inclusion of irrelevant and duplicated datasets makes classifier efficiency analysis challenging. In this research, ensemble, and Feature selection are implemented to overcome the issue. In the first step, CFS-BA is used to reduce the dimensionality of the dataset, and then an efficient analysis of IDS is undertaken using an ensemble approach. The experimental technique is conducted utilizing three distinct datasets, and the findings demonstrate that the suggested CFS-BA-Ensemble methodology outperforms previous alternatives. **[11],** the author discusses the significance of network cyber-attacks and suggests the use of IDS to combat the aforementioned problems. In addition, the author suggests that the backup IDS is significantly affected by high dimensions, irrelevant data, and redundancy. In this research, CFS-FPA is presented and tested to provide a solution to the problem. Ada boosting and bagging ensemble learning algorithms are incorporated into enhanced intrusion detection. The experimental approach is conducted using the CICIDS2017 dataset, and the acquired results demonstrate that the suggested method achieves 99.7 percent accuracy, 0.053 percent FNR, and 0.001 percent FAR. **[12],** the author states that depression is one of the most prevalent mental illnesses that profoundly affect human lives. A sad individual can be detected based on user-shared social media material and machine learning. Due to the presence of redundant and high-dimensional elements, the human state cannot be accurately identified. This research proposes and experiments with a Hybrid strategy that combines the FS and Ensemble approaches to address the challenge. The experimental findings indicate that the proposed method has an accuracy of 90.27 percent.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

III. METHODOLOGY

Customer is one of the most essential people in a business, as the enterprise's business is entirely dependent on the client. Assessments of these prospective customers facilitate the development of lucrative businesses and customer happiness. In the modern, technologically advanced world, data regarding customers are collected in vast quantities and with high dimensions. It is conceivable for correlated, irrelevant, noisy, and missing variables to exist as a result of the heterogeneous manner of data collection, as the consumer data is obtained in a variety of ways. Due to the presence of associated and irrelevant characteristics, it is impossible to conduct an efficient analysis of the consumer using these data. Using these datasets with machine learning also increases the temporal complexity. To solve the issue of correlated and irrelevant variables, approaches to feature selection are addressed. FS is an essential pre-processing approach that permits the selection of relevant factors and the elimination of correlated and irrelevant variables, hence enhancing the performance of ML. In general, FS is classified into 3 types: filter, wrapper, & hybrid method [13]. The filter method is a quick one, but the variables used are ineffective. Similarly, a wrapper method is effective yet computationally inefficient. This research proposes and experiments with ensemble feature selection as a solution to the problem with FS. Ensemble learning that enables the combination of numerous outputs into a single output and provides a stable subset of features. In this manner, a better subset of variables is extracted and modelled using ML. There are two approaches in ensemble learning: the homogeneous approach and the heterogeneous approach. Homogeneous approach denotes the application of the same feature selection strategy to distinct subsets of training data and the aggregate of various outputs into a single output using some aggregation technique. Heterogeneous approach denotes the application of different feature selection techniques to the same training dataset and the combining of several outputs into a single result using some aggregation technique. Based on the heterogeneous ensemble method, this study proposes and tests the heterogeneous ensemble variable selection (HEVS) method. This HEVS method employs ensemble with feature selection to determine the subset of important variables to be modelled using ML classifier.

This proposed HEVS approach works by combining feature selection with the ensemble learning model. The overall architecture HEVS described in figure 1.

A. Data Collection

The research is conducted using three different dataset captured from UCI repository. First dataset is Bank dataset (BM), second dataset is German credit dataset (GC) and third one is Australian credit dataset (AC). Bank dataset consists of 45211 instances with 17 variables. Likewise German credit dataset consists of 1000 instances with 20 variables. Likewise Australian credit dataset consists of 690 instances with 15 variables.

B. Data Pre-Processing

Data pre-processing is one of the important technique in Machine Learning which helps to remove the noisy and missing data. Since the customer data are collected in real time there is high possible of holding noisy and missing data and using the data pre-processing the above problem can be eliminated.

C. Feature Selection

Feature selection, also known as variable selection or attribute selection, is an operation applied to a dataset in order to pick relevant variables from the entire dataset and exclude irrelevant and correlated variables. The objective of FS is to select a subset of relevant variables without minimising the knowledge underlying the dataset. In general, FS is classified into 3 types:

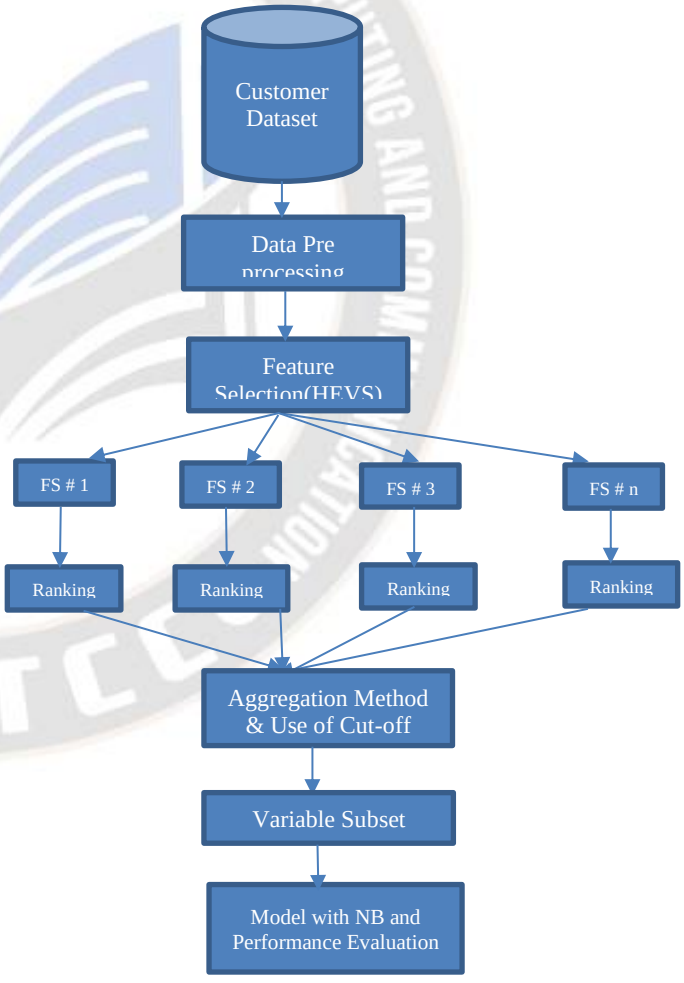

Figure 1: Overall approach of proposed HEVS methodology

Filter, wrapper, & hybrid method. The filter method is a quick one, but the variables used are ineffective. Similarly, a wrapper method is effective yet computationally inefficient. To solve the disadvantages of feature selection, ensemble learning with FS

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
Article Received: 22 February 2023 Revised: 28 March 2023 Accepted: 12 April 2023

is proposed and tested in this study. There are two techniques to ensemble learning: the homogeneous approach and the heterogeneous approach. In this study, a mixed methodology is employed. Heterogeneous approach denotes the application of different feature selection techniques to the same training dataset and the combining of several outputs into a single result using some aggregation technique. In this study, heterogeneous ensemble variable selection is presented and tested based on a heterogeneous approach.

Procedure:

Input: Training data D, Variable selector vs , Aggregation-v_a, Threshold t

Output: Final Variable Subset- FVS

1. Let $D = \{v_1, v_2, v_3, \ldots . v_n | \, C_n \}$, where v_n represents the variables and C_n represents the class label.
2. Using pre-processing the customer data is processed.
3. Dataset D is applied to different vs
4. Generating v_r ranking list by using vs
5. v_a combining using suitable aggregation procedure obtained from v_r
6. Select relevant variable subset from v_a using suitable Threshold- t
7. Model the selected variable subset with NB Model.
8. Compare performance evaluation using different validity scores.

The processed customer dataset is applied to five filter FS approach and each FS approach rank the variables accordingly to class label. Chi-square, ReliefF, IG, GR and SU, FS are applied to customer dataset. Further the using FS the variables in the customer dataset are ranked.

D. Aggregation Method & Cut-off Value

The paper provided a strategy for capturing a meaningful variable subset by integrating data from five filter FS techniques using Mean and Min aggregation methods. To capture the relevant variable subset, an appropriate threshold (t) was applied to the final aggregated sorted list. The variable subset in this investigation was chosen using a 50% threshold value [14]. Lastly, the collected variable subset was modelled with the Nave Bayes classifier, and the resulting results were projected with various validity scores.

E. Performance Evaluation

The results obtained from the proposed methodology is projected using different validity scores. The validity Scores applied are Accuracy, Sensitivity, Precision and Specificity.

IV. EXPERIMENTAL PROCEDURE

The suggested method is tested using three distinct datasets obtained from the UCI repository. The experimental technique is carried out from two perspectives: Heterogeneous ensemble variable selection with Nave Bayes and Naive Bayes without variable selection. Finally, the results of the two methodologies are compared using different parameter scores.

A. Parameter Score:

The results obtained from the above two methodologies are projected and compared using 4 different parameter score: Accuracy, Sensitivity, Precision and Specificity [15].

$$Accuracy = \frac{TN+TP}{TP+TN+FP+FN} \quad (1)$$

$$Sensitivity = \frac{TP}{FN+TP} \quad (2)$$

$$Precision = \frac{TP}{FP+TP} \quad (3)$$

$$Specificity = \frac{TN}{FP+TN} \quad (4)$$

B. Experimenatl Procedure

1. The dataset required for the analysis is captured from the UCI library.

2. In the first methodology (HEVS-NB), the learning set is applied to 5 different filter FS approaches. After evaluation, the variables are sorted accordingly to relevance to the target class. Then using some aggregation methods the output obtained from the five filters FS approaches is combined into a single output. Further picking the efficient variable subset threshold value is applied, since the filter approach used in the HEVS approach only sorts the variables accordingly to the relevance with the class label. Lastly, the subset of the variables captured from HEVS is modeled with the Naive Bayes model, and the results obtained are projected using different parameter scores.

3. In the second methodology, without using any feature selection or pre-processing step the complete customer dataset is modeled with the Naive Bayes model.

4. Lastly results obtained from two methodologies are projected and compared using different parameter scores: Accuracy, Sensitivity, Precision, and Specificity

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

C. Results of Hetergeneous Ensemble Variable Selection-NB

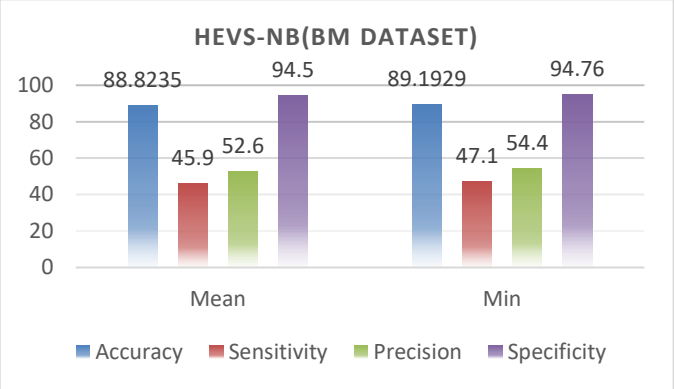

Figure 2: Results of Heterogeneous Ensemble variable selection-NB using 50% Threshold Value

Figure 2 represents the results of HEVS-NB using BM dataset. Here the proposed approaches achieves 88.82 accuracy, 45.9 Sensitivity, 52.6 Precision and 94.5 Specificity in Mean aggregation. Likewise the proposed approaches achieves 89.19 accuracy, 47.9 Sensitivity, 54.4 Precision and 94.76 Specificity in Min aggregation

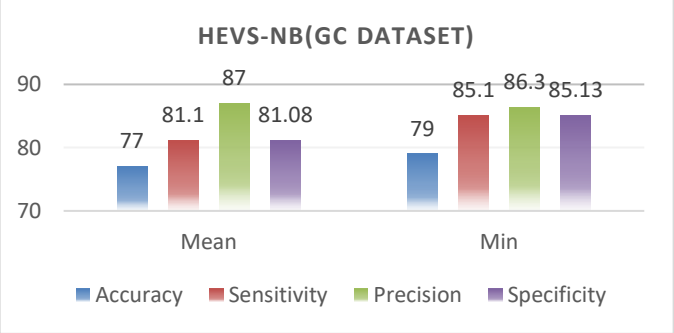

Figure 3 represents the results of HEVS-NB using GC dataset. Here the proposed approaches achieves 77 accuracy, 81.1 Sensitivity, 87 Precision and 81.08 Specificity in Mean aggregation. Likewise the proposed approaches achieves 79 accuracy, 85.1 Sensitivity, 86.3 Precision and 85.13 Specificity in Min aggregation

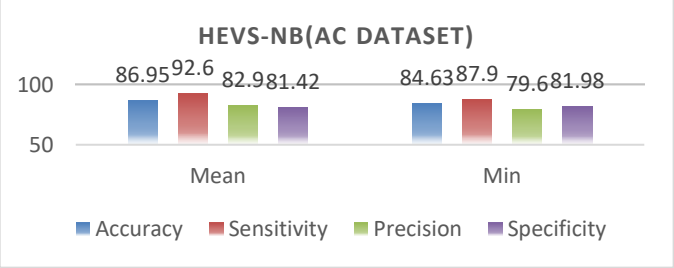

Figure 4: Results of Heterogeneous Ensemble variable selection-NB using 50% Threshold Value

Figure 4 represents the results of HEVS-NB using AC dataset. Here the proposed approaches achieves 86.95 accuracy, 92.6 Sensitivity, 82.9 Precision and 81.42 Specificity in Mean aggregation. Likewise the proposed approaches achieves 84.63 accuracy, 87.9 Sensitivity, 79.6 Precision and 81.98 Specificity in Min aggregation.

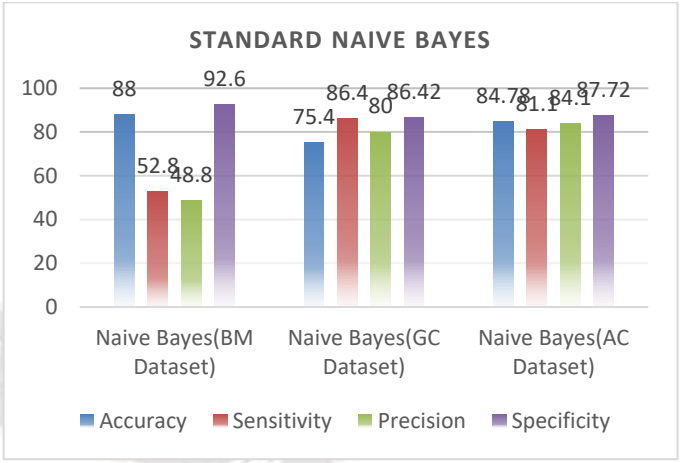

Figure 5: Results of Standard Naïve Bayes without using Feature Selection

Figure 5 represents the results of Naive Bayes without any feature selection with three different datasets. Naive Bayes using Bank Dataset archives 88 archives, 52.8 Sensitivity, 48.8 Precision, and 92.6 Specificity. Likewise, Naive Bayes using German Credit Dataset archives 75.4 archives, 86.4 Sensitivity, 80 Precision, and 86.42 Specificity. Likewise, Naive Bayes using Australian Dataset archives 84.78 archives, 81.8 Sensitivity, 84.1 Precision, and 87.72 Specificity

D. Result Discusssion

The experimental results that were acquired from the two distinct viewpoints are projected from figure 2 through figure 5. Figure 2 illustrates the findings obtained by HEVS-NB when applied to the bank marketing dataset. In this context, the proposed methods obtain an accuracy of 88.82, a sensitivity of 45.9, a precision of 52.6, and a specificity of 94.5 in the mean aggregation. Similarly, the proposed methods obtain an accuracy of 89.19, a sensitivity of 47.9, a precision of 54.4, and a specificity of 94.76 in the Min aggregation. Figure 3 illustrates the findings obtained by HEVS-NB when applied to the German Credit Dataset. In this context, the offered methods obtain an accuracy of 77, a sensitivity of 81.1, a precision of 87, and a specificity of 81.08 in the mean aggregation. In the same vein, the proposed methods reach a level of accuracy of 79, a sensitivity of 85.1, a precision of 86.3, and a specificity of 85.13 in the Min aggregate. The findings of HEVS-NB when run using the Australian Credit Dataset are displayed in figure 4. In this context, the offered methods obtain an accuracy of 86.95, a sensitivity of 92.6, a precision of 82.9, and a specificity of 81.42 in the mean aggregate. In the same vein, the proposed methods obtain an accuracy of 84.63, a sensitivity of 87.9, a precision of 79.6, and a specificity of 81.98 in the Min aggregation. The outcomes of using Naive Bayes with three different datasets are depicted in figure 5, which does not include any feature selection. The Naive Bayes algorithm, when applied to the Bank

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

Dataset archives, achieved a sensitivity of 52.8, a precision of 48.8, and a specificity of 92.6. Similarly, Naive Bayes, when applied to the German Credit Dataset archives, achieved a sensitivity of 86.4, precision of 80, and specificity of 86.42. Similarly, the Naive Bayes algorithm, when applied to the Australian Dataset archives, obtained 84.78 archives, 81.8 Sensitivity, 84.1 Precision, and 87.72 Specificity. When compared to the findings presented in figures 2 through 5, it can be seen that the suggested HEVS-NB achieves significantly better results than the traditional NB does. By employing FS approaches to exclude associated and inappropriate variables from the dataset and selecting just relevant variables to model with NB, one can achieve far better results than would otherwise be possible. Because essential variables are taken into consideration when modeling with NB, this proposed methodology additionally reduces the time complexity of the process.

E. Result Findings:

1. The presence of correlated, irrelevant, and inappropriate variables are unavoidable in the real-time dataset.

2. The use of this dataset without any pre-processing with machine learning results in poor prediction results.

3. To overcome the issue use of feature selection helps to improve the prediction results.

4. In this research, Heterogeneous Ensemble based variable selection is proposed to choose the relevant variable subset to model with ML.

5. The proposed approach helps to choose the stable and relevant variable subset from the complete dataset.

6. Finally experimental results reveal the proposed methodology works better and archives better prediction results compare to the standard Naïve Bayes model.

7. Further time complexity is also reduced with this proposed methodology, since relevant variables are considered to model with NB

V. CONCLUSION

The growth of the enterprise is primarily dependent on how well the firms understand the behavior patterns of customers within the enterprise and how well they improve the level of pleasure felt by those customers. The study of the behaviours or patterns of customers not only helps businesses determine how long customers will remain loyal to them but also gives them a clear image of how they may increase their customers' level of satisfaction. The expanded usage of the internet and technology in today's world helps to acquire large amounts of data about the client in a variety of various ways and perspectives. When these customer datasets are used effectively, it helps to identify the underlying information about the client in a way that is both efficient and successful. Nevertheless, because the data were collected in a variety of different ways, the results included correlated, irrelevant, missing, and unsuitable factors. With these datasets, it is impossible to conduct an efficient analysis of customer information. To solve the issue that was described earlier, the technique of FS is utilized. The goal of the FS process is to choose the pertinent variables from the entire dataset without reducing the amount of information known about the dataset. The first type of FS is a filter, the second is a wrapper, and the third is a hybrid approach. In general, these three categories are categorized as follows: Although it is a quick method, the filter strategy does not choose effective variables. In a similar vein, the wrapper solution is an efficient one, but it is inefficient computationally. In this research, a heterogeneous ensemble feature selection is presented and experimented with as a potential solution to the difficulty given by these feature selections. When compared to other FS approaches, utilizing an ensemble approach makes it much easier to select a stable variable subset. The learning dataset is put through five distinct filter FS procedures in the suggested method, after which the resulting ranking attributes are aggregated utilizing two distinct methods: the mean method and the min method. After that, further, build the final ranking qualities by utilizing the cut-off value, and then choose the essential variables. Because the filter method that was employed in the HEVS approach could only rank the variables, the next step, which was to select the variable subset cut-off value, was necessary. Then, using the NB model, the variables that were collected by the suggested method were explored. The experimental technique is carried out utilizing two distinct points of view: the first makes use of heterogeneous ensemble variable selection in conjunction with naive Bayes, while the second makes use of naive Bayes alone and does not make use of variable selection. In the final step of this process, the results generated from the two methods are compared using different parameter scores. The experimental results that were acquired from the two distinct viewpoints are projected from figure 2 through figure 5. When compared to the findings presented in figures 1 through 4, it can be seen that the suggested HEVS-NB achieves significantly better results than the traditional NB does. By employing FS approaches to exclude associated and inappropriate variables from the dataset and selecting just relevant variables to model with NB, one can achieve far better results than would otherwise be possible. Because essential variables are taken into consideration when modeling with NB, this proposed methodology additionally reduces the time complexity of the process.

REFERENCES

[1] R.Siva Subramanian & Dr.D.Prabha, 'A survey on customer relationship management', 4th International Conference on Advanced Computing and Communication Systems

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6599
***Article Received**: 22 February 2023 **Revised**: 28 March 2023 **Accepted**: 12 April 2023*

(ICACCS), Coimbatore, Electronic ISBN: 978-1-5090-4559-4 pp. 1-5, 2017

[2] Dr.S.Balakrishnan & Dr.M.Karpagam, ,'Performance Evaluation of Naive Bayes Classifier with and without Filter Based Feature Selection', International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol-8 Issue-10, 2019, pp.2154-2158.

[3] Chih-Wen Chen, Yi-Hong Tsai, Fang-Rong Chang,Wei-chao Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results", Expert Systems, Vol.37, Issue 5, 2020, e12553.

[4] Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin S.C.Yong, Wei King Tiong," A New hybrid ensemble feature selection framework for machine learning-based phishing detection system", Information Sciences, Vol.484, 2019, pp.153-166.

[5] Barbara Pes," Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains", Neural Computing and Applications, Vol.32, 2020,pp.5951-5973.

[6] Diwakar Tripathi, Damodar Reddy Edla, Ramalingaswamy Cheruku, venkatanareshbabu Kuppilli," A novel hybrid credit scoring model based on ensemble feature selection and multiplayer ensemble classification", computational Intelligence, Vol.35, issue.2,2019, pp.371-394.

[7] Supoj Hengpraprohm, Thaksin Jungjit," Ensemble Feature Selection for Breast Cancer Classification using Microarray Data", Intelligencia Artificial, vol.23, 2020, No.65.

[8] Veronica Bolon-Canedo and Amparo Alonso-Betanzos," Ensembles for feature selection: A review and future trends", Information fusion, Vol.52, 2019,pp.1-12.

[9] Siva Subramanian.R and Prabha.D," Ensemble Variable Selection for Naïve Bayes to Improve Customer Behaviour Analysis", Computer Systems Science & Engineering, Vol.41, No.1, 2022, pp.339-355.

[10] Yuyang Zhou, Guang Cheng, Shanqing Jiang, Mian Dai, " Building an efficient intrusion detection system based on feature selection and ensemble classifier", Computer Networks,Vol.174, 2020, 107247.

[11] Doaa N.Mhawi, Ammar Aldallal, Soukeana Hassan, Advanced Feature-selection-Based Hybrid Ensemble Learning Algorithms for Networks Intrusion Detection Systems", Symmetry, vol.14, 2022, No.7, 1461.

[12] Jingfang Liu and Mengshi Shi," A Hybrid Feature Selection and Ensemble Approach to Identify Depressed Users in Online Social Media", Frontiers in Psychology, 2021,Vol.12.

[13] Dr.J.Aswini, Mrs. B.Maheswari & Mrs.M.Anita,. Alleviating NB conditional independence using Multi-stage variable selection (MSVS): Banking customer dataset application. Journal of Physics: Conference Series. 1767. 012002, 2021.

[14] R.Siva Subramanian, R & Dr.D.Prabha, "Ensemble variable selection for Naive Bayes to improve customer behaviour analysis", Computer Systems Science & Engineering, Tech Science Press, Vol.41, No.1, 2022,pp.339-355.

[15] S. N. Bushra, Dr.J.Aswini, G.Nirmala, B.Maheswari,2022, Customer Analysis using Machine Learning with Feature Selection Approaches: A Comparative Study", 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, pp. 196-202,2022