Article

Analysis of Breast Cancer Risk Factors Data: Association Rule Mining based on Ethnic Groups and Classification using Super Learning

Md Faisal Kabir [1,3]*, Simone A. Ludwig [1] and Abu Saleh Abdullah [2]

[1] Department of Computer Science, North Dakota State University, ND, USA;
[2] Boston Medical Center, Boston University, MA, USA;
[3] Department of Computer Science and Engineering,United International University, Dhaka, Bangladesh.
* Correspondence: Md Faisal Kabir, Department of Computer Science, North Dakota State University, ND, USA; mdfaisal.kabir@ndsu.edu

Version September 12, 2024 submitted to Information

Abstract: Breast cancer is the most commonly diagnosed cancer in women worldwide. Prevention strategies are essential to decrease their impact on the population. Efficient techniques for breast cancer detection are the key to reducing the mortality rate. In the first part of this research, we used rule mining and discovered rules of breast cancer patients based on different ethnic groups. The interpretation of these rules is also discussed. This knowledge in the form of rules can be useful for physicians and other healthcare organizations to understand the characteristics of breast cancer patients for particular races. Ultimately, different prevention programs or processes targeting a specific race can be initiated in the early stage of disease progression. In the second part of this research, the machine learning method named super learning or stacked-ensemble that consist of three distinct base learners/algorithms is used. A comparison of the performance of the super learner and the individual base learners is investigated. The results show that the super learning achieves better predictive performance compared to the individual base learners on the breast cancer risk factors data.

Keywords: data mining; class association rule mining; breast cancer; risk factors; machine learning, classification, super learning, stacked ensemble, H2O.

1. Introduction

Cancer is one of the devastating diseases worldwide. According to the World Health Organization (WHO) [1] there are more than 10 million new cases reported every year. Cancer affects nearly every household although different types of cancer are prevalent in different geographical regions. One example is breast cancer, which is the most common type of cancer in women worldwide with 1.7 million new cases being diagnosed in 2012 [1]. Therefore, prevention strategies are needed to address this issue. The identification of risk factors of breast cancer is important since it allows physicians to be able to inform the patients about the risks associated. Furthermore, the physician will be able to suggest preventive measures.

Data mining can be described a process of extracting implicit, unknown, and useful information from a large volume of data [2]. Data mining encompasses several different techniques. Rule mining is one of the techniques used that provides the mining knowledge in form of rules, which are easily understood [3]. Association rule mining [4] is a special category of rule mining that was introduced in 1993. Since then this technique has been applied to different application domains, in particular, in the medical domain [5] [6] [7] [8]. Association Rule Mining (ARM) has also another name used in industry, which is “market-basket analysis".

In the first part of this paper, we look at the discovery of significant rules for breast cancer patients focusing on different ethnic groups. Predicting the risk of the occurrence of breast cancer is an important issue for clinical oncologists. A reliable prediction will not only help oncologists and other clinicians in their decision-making process but will also allow clinicians to choose the most reliable and evidence-based treatment. Moreover, the best prevention strategies for the patients can be identified.

Classification is a supervised learning technique that classifies unseen data into a finite set of classes [9] by learning a target function that maps each feature into one of the pre-defined classes [10] [11]. The target function is also referred to as the classification model. Classification is applied to many different fields with the aim to come up with the best performing model by experimenting with different classification algorithms. The usual procedure done to achieve better performance for a particular data set is to use a single classifier. However, nowadays a single classifier usually does not provide the best performance so research have looked at different techniques to address this. For example, multiple models could be used for the classification task. Researchers have been applying Bagging (Random Forest) and Boosting (Gradient Boosting Machine) ensemble techniques in different areas in order to obtain better performance [12] [13] [14]. Lately, a super learning or stacking method has been introduced that ensembles several base learners to obtain a better predictive model [9] [15] [16].

In the second part of this paper, we apply a super learner or stacked ensemble technique to the data set. The super learner uses three base learners namely Gradient Boosting Machine (GBM), Random Forest (RF), and Deep Neural Network (DNN). As the meta-learner, the Generalized Linear Model (GLM) is used [17] [18]. A comparison of the performance of the super learner and the individual base learners is conducted.

This paper is comprised of the following sections proceeding the introduction 1. The related work is discussed in Section 2. The preliminary background is described in Section 3. The analytical workflow is discussed in Section 4. In Section 5, we show our experiments. In particular, the evaluation criteria of the classification model along with the results of the super learner are shown and discussed. Section 6 discusses the results of the experiments. Section 7 provides the summary of this paper with a conclusion of our findings.

2. Related Work

Researchers have investigated breast cancer risk factors to find the relationship among them; they have also developed various breast cancer risk prediction models [19] [20][21][22]. The authors in paper [19], used statistical methods to investigate the association between Hormone Replacement Therapy (HRT) and breast cancer risk and they found that HRT increases the risk of breast cancer. In paper [20], authors used Gali model that can estimate the number of breast cancer cases for white women who are examined annually. The authors in paper [21], used commonly identified breast cancer risk factors to describe the model. Furthermore, a data mining approach called k-nearest-neighbor (KNN) is applied to determine the breast cancer risk score that ultimately improves the readability for physician and patients [22].

Data mining technique called association rule mining (ARM) has been applied in the medical field to extract knowledge in a form of rules from the data. In paper [6], the authors implemented the ARM-based technique for finding co-occurrences of diseases carried by a patient from a healthcare database. The method collected data from a patients' healthcare repository from which association rules were discovered. The researcher also investigated class association rule mining which is a variation of ARM technique, to discover the characteristic features [23]. By definition, a class association rule set is a subset of association rules with the particular classes as their consequences [24]. In traditional ARM, if we assign very low support value, then the class ARM will generate overfitting rules for a frequent class. On the other hand, if we specify the support value very high, then the insufficient number of rules for an infrequent class will be generated. In class ARM, this is not an issue since mining is done according to the class, and thus the algorithm is not influenced by the unequal proportion between the classes. As an example, the authors in [5], discovered useful rules of breast cancer and non-breast cancer patients from risk factors data. In the first part of this paper, we discovered hidden but significant rules for breast cancer patients based on different ethnic groups. Rules of breast cancer patients from different ethnic groups can be useful for physicians to make a decision and to inform patients about risk factors. Also, physicians can alert patients about the potential risks of developing breast cancer. By this way, a prevention program or process for particular races can be initiated in the early stage of disease progression.

Machine Learning (ML) techniques have applied in the medical field to help the decision-making process, for instance, for the prediction of cancer risk. Authors in paper [25] applied three different classification methods on breast cancer risk factors data. The authors also used several resampling techniques on the training data as risk factors data have an unequal proportion between cancer and non-cancer cases. Ensemble techniques, which is a popular modern machine learning algorithms, have been applied in different fields including the medical domain to obtain better predictive performance [12] [13] [14]. Super learning or stacking method that ensembles a group of base learners are also used by researchers [15] [16]. In [9], the authors used two different forms of super learner (SL); first one consist of two base learners and other consisting of three base learners. The authors showed that the super learner with three base learners provides better performance than the super learner having two base learners and all individual ML algorithms that they applied for their research. The authors used four popular data sets to assess the performance of their techniques.

In the second part of this paper, we present a super learner technique with three base learners namely gradient boosting machine (GBM), random forest (RF), and Deep Neural Network (DNN); and as a meta-learner Generalized Linear Model (GLM) is used [17] [18]. We compare the performance of the super learner (SL) with the individual base learners; it shows that SL outperforms the individual base learners for the breast cancer risk factors data that we considered for this study.

3. Preliminaries

3.1. Data Description

The data set contains information from 6, 318, 638 mammography examinations that was obtained from the Breast Cancer Surveillance Consortium (BCSC) database [26]. The data collection period was between January 2000 and December 2009. More information about BCSC data resource can be found at http://www.bcsc-research.org.

3.2. Data Pre-processing

The BCSC risk factors data set was pre-processed as outline in [5], and there are a total of 11 attributes or columns with 1,015,583 instances. Among these, number of breast cancer patients and non-breast cancer individuals are 60,800 and 954,783 respectively. The distributions of different attributes of risk factors data can be found in paper [5].

3.3. Further Pre-processing for Rule Discovery of Ethnic Groups having Multiple Consequents

Our goal is to extract hidden but useful information in the form of rules for different ethnic groups of breast cancer patients from the risk factors data set. For that, we merged two attributes named breast cancer history (where the value is Yes – meaning we are considering breast cancer patients) and race attribute. We named it race-cancer-history since it considers breast cancer patients of different ethnic background. The distribution of race-cancer history for the cancer group is shown in Table 1. For instance, the Non-Hispanic-White-Yes value of attribute race/ethnicity represents breast cancer patients of the non-Hispanic White group. After converting into a transaction-like database there was total of 44 items and 60,800 instances for a class association rule mining.

Table 1. Distribution of breast cancer patients based on race or ethnicity

Race or ethnicity	Number of individuals
Non-Hispanic-White-Yes	54869
Asian_or_Pacific Islander-Yes	1867
Hispanic-Yes	2028
Other_or_Mixed-Yes	1055
Non-Hispanic-Black-Yes	736
Native-American-Yes	245

3.4. Data Set for Classification

The BCSC risk factors data was divided into train and test portions for the classification model [25]. A total number of training examples were 812,466 with 48,640 breast cancer patients and 763,826 non-breast cancer individuals while the total number of test instances were 203,117 having 12,160 breast cancer patients and 190,957 non-breast cancer individuals. As risk factors data have imbalanced characteristics which indicate the data has an uneven distribution between the cancer patients and non-cancer individuals; for that reason, the training data has been resampled using different techniques [25]. For our analysis, we selected training data that has been modified using SMOTE and ENN [25]. The distribution of the training data that were obtained by applying SMOTE and ENN is shown in Table 2.

Table 2. Training data that were obtained by applying SMOTE and ENN.

Resampling technique	Class = yes	Class = no	Total instances
SMOTE + ENN	437,256	658,167	1,095,423

4. Analytical Workflow

In this section, we first used class association rule mining on modified risk factors data, discussed in Section 3.3 to extract useful rules of individuals having breast cancer from different ethnic/groups. After that, we applied the ensemble technique called super learning on the BCSC risk factors data as discussed in Section 3.4.

4.1. Association Rule Mining

Association Rule Mining (ARM) is one of the important techniques to generate and extract useful information from a large database. Detailed information on generating association rules along with important measures can be found in [2] [5]. Rules can be generated from data sets by specifying particular classes as their consequence which is named as class association rule technique. More information about class association rule techniques and their usage is available in paper [5] [27]. In this research, we applied class ARM and discovered hidden but significant rules for breast cancer patients of different ethnic groups.

It is to be mentioned that here we have extracted rules having two attributes/items in the consequent at the same time. For that we have merged two attributes into one namely race-cancer-history that indicate breast cancer patients of a particular ethnic group; discussed in Section 3.3. As we consider non-Hispanic white, Hispanic, and Asian-or-pacific-islander races in the class association rule mining process, for that we ran the algorithm with consequent value as any of these three races along with specified support and confidence values. For instance, during rules generation for the non-Hispanic white group, we set the consequent as “race-cancer-history = non-Hispanic-White" along with other important measures. Rules of breast cancer patients from different ethnic groups can be useful for physicians not only to make a decision but also to inform individuals about risk factors.

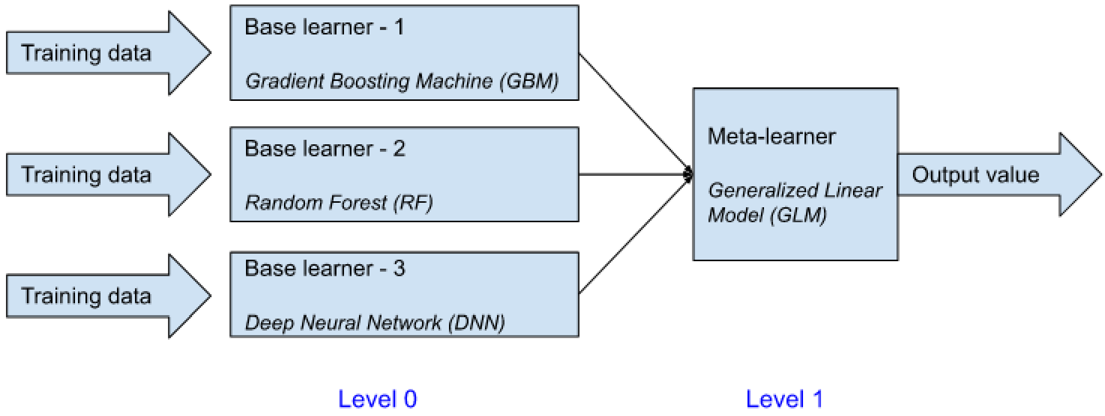

Figure 1. Concept Diagram of Super Learner [9].

4.2. Super Learning

Super learning (SL) or stacked ensemble generally consists of two or more machine learning (ML) algorithms. It is a cross-validation-based technique for combining ML algorithms that generally provide better predictions than those of the base algorithm [15] [16]. Detailed information and applications of super learner can be found in papers [15] [9] [28] [29] and the concept diagram of the super learning method for this research is illustrated in Fig. 1.

4.2.1. Training Super Learner with Base Learners and Meta-Learner

For base learners three machine learning algorithms were specified from H2O namely Gradient Boosting Machine (GBM), Random Forest (RF), and Deep Neural Network (DNN) [17]. The Generalized Linear Model (GLM) was selected as a meta-learner[17] [18].

We trained all individual ML algorithms namely GBM, RF, and DNN. On each of these learning algorithms, default parameters available in H2O were used. Besides, 10-fold cross-validation is performed on each of these algorithms and the cross-validation prediction parameter specified as True. The target column also called as class value for risk factors data is binary; for that reason, the Bernoulli distribution was selected. In Table 3 the important parameters (default values) for each base learners are listed.

Table 3. Default parameter values for corresponding base learners.

Base learner	**hyper-parameter default values**
GBM	learn_rate: [0.1] sample_rate: [1.0] col_sample_rate_per_tree: [1.0] max_depth: [5]
RF	sample_rate: [0.63] col_sample_rate_per_tree: [1.0] max_depth: [20]
DNN	activation: [rectifier] hidden: [200, 200] l1: [0.0] l2: [0.0]

5. Experiments and Results

Results that were obtained using a class association rule are shown in this section. Strong rules for breast cancer patients for different races were generated by selecting an appropriate value of support

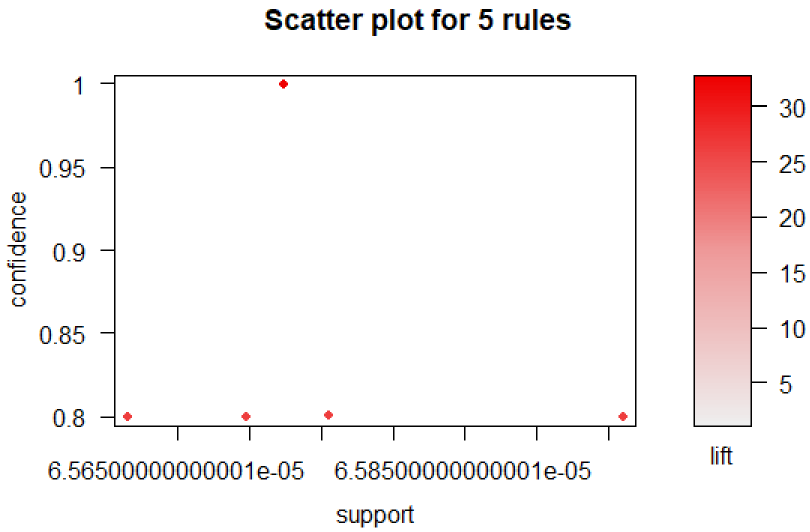

Figure 2. Scatter plot of 5 rules for breast cancer patients of Asian_or_pacific_Islander race with specified support, confidence, and lift values.

and confidence. Interpretations of a few strong rules are also shown in this section. In addition, evaluation measures for the classification model are discussed in this section. The results obtained from SL are also illustrated in this section.

5.1. Rules Discovery

Our objective is to generate characteristics of patients as a form of rules from a particular race/ethnic background with prior breast cancer. For that, we discovered rules using the class association rule technique with the specified support and confidence. We also defined the consequent of a rule (Race_cancer_history) so that we can get our target rules that represent the individual who has breast cancer and a particular race/ethnic background. From the distribution of breast cancer patients of race/ethnic group shown in Table 1, we can see that there are very few numbers of breast cancer patients of native-American, non-Hispanic-Black, and other-or-mixed in the BCSC risk factors data set. For that, we did not consider these ethnic groups having breast cancer in the rule mining process. In the rule mining process, we consider breast cancer patients of the non-Hispanic-White group that is the dominant group compared to other races. We also consider Hispanic and Asian-or-pacific-Islander races in the class association rule mining process.

5.1.1. Rules of Breast Cancer Patients based on Ethnic Groups

After several experiments, the support and confidence values were assigned to 0.005% and 80%, respectively, and we obtained 5 rules. Here, we specified the consequent value "Asian_or_pacific_Islander_Yes", to obtain the rules of breast cancer patients having Asian_or_pacific_Islander ethnic group. These rules are shown in Table 4 while the scatter plot of these rules sort by lift value is shown in Fig. 2.

For breast cancer patients of the Hispanic group, after several experiments, the support and confidence values were specified to 0.005% and 85%, respectively, and we obtained 5 rules. Here, we assign the consequent "Hispanic_Yes", to obtain the rules of breast cancer patients belonging to the Hispanic ethnic group. These rules are shown in Table 5, while the scatter plot of these rules are shown in Fig. 3.

For breast cancer patients of the non-Hispanic white group, the support and confidence values were assigned to 30% and 90%, respectively, and we obtained 23 rules. Here, we set the consequent or

Table 4. Discovered rules using the class association rule method (consequent = Asian-or-pacific-Islander-Yes) with corresponding support and confidence.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Age_group=age_30_34, First_degree_relative=No, Age_menarche=Age_less_12, Age_first_birth=Nulliparous, BIRADS_breast_density=Heterogeneously_dense, BMI_group=10-to-lessThan_25} =>{Race_cancer_history=Asian_or_pacific_Islander_Yes}	0.005	100	32.57
2	{Age_group=age_40_44, First_degree_relative=No, Age_first_birth=Age_greater_equal_30, biopsy=No} =>{Race_cancer_history=Asian_or_pacific_Islander_Yes}	0.005	80	26.05
3	{Age_group=age_30_34, Age_menarche=Age_less_12, Age_first_birth=Nulliparous, BIRADS_breast_density=Heterogeneously_dense, BMI_group=10-to-lessThan_25} =>{Race_cancer_history=Asian_or_pacific_Islander_Yes}	0.005	80	26.05
4	{Age_group=age_30_34, Age_menarche=Age_less_12, First_degree_relative=No, Age_first_birth=Nulliparous, BIRADS_breast_density=Heterogeneously_dense} =>{Race_cancer_history=Asian_or_pacific_Islander_Yes}	0.005	80	26.05
5	{First_degree_relative=No, HRT=No, Age_menarche=Age_greaterEqual_14, Age_first_birth=Age_greater_equal_30, Menopaus=post_menopausal, biopsy=No, BMI_group=10-to-lessThan_25} =>{Race_cancer_history=Asian_or_pacific_Islander_Yes}	0.005	80	26.05

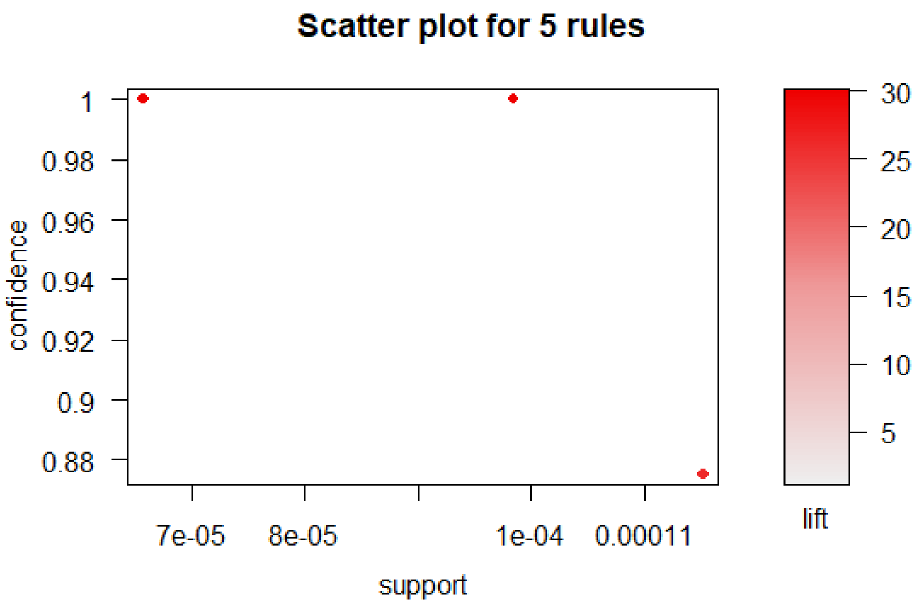

Figure 3. Scatter plot of 5 rules for individuals with breast cancer of Hispanic race with corresponding support, confidence, and lift values.

Table 5. Extracted rules using the class association rule technique (consequent = Hispanic-Yes) with corresponding support, and confidence value.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Age_group=age_18_29, Age_first_birth=Age_20_24, BMI_group=25-to-lessThan_30} =>{Race_cancer_history=Hispanic_Yes}	0.005	100	29.98
2	{Age_group=age_30_34, Age_first_birth=Age_20_24, BIRADS_breast_density=Extremly_dense} =>{Race_cancer_history=Hispanic_Yes}	0.005	100	29.98
3	{Age_group=age_18_29, Age_menarche=Age_less_12, BMI_group=10-to-lessThan_25, BIRADS_breast_density=Heterogeneously_dense} =>{Race_cancer_history=Hispanic_Yes}	0.005	100	29.98
4	{Age_group=age_45_49, First_degree_relative=No, Age_first_birth=Age_less_20, HRT=Yes, BMI_group=10-to-lessThan_25, BIRADS_breast_density=Heterogeneously_dense} =>{Race_cancer_history=Hispanic_Yes}	0.005	87.5	26.23
5	{Age_group=age_65_69, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth=Age_less_20, HRT=No, biopsy=Yes, BMI_group=10-to-lessThan_25, BIRADS_breast_density=scattered_fibroglandular_densities} =>{Race_cancer_history=Hispanic_Yes}	0.005	87.5	26.23

class value was specified as "Non-Hispanic-White-Yes" to obtain the rules of breast cancer patients having the non-Hispanic white ethnic group. The scatter plot of these 23 rules are shown in Fig. 4, while the top 5 rules sorted by the lift value are shown in Table 6.

Table 6. Rules generated using the class association rule technique (consequent set to "Non-Hispanic-White-Yes") with corresponding support,confidence, and lift value. Top 5 rules sort by lift values are shown.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{BIRADS_breast_density=scattered_fibroglandular_densities, HRT=No, Menopaus=post_menopausal, biopsy=Yes } =>{Race_cancer_history=Non-Hispanic-White_Yes}	37	92	1.02
2	{BIRADS_breast_density=scattered_fibroglandular_densities, Menopaus=post_menopausal, biopsy=Yes } =>{Race_cancer_history=Non-Hispanic-White_Yes}	38	92	1.02
3	{BIRADS_breast_density=scattered_fibroglandular_densities, HRT=No, Menopaus=post_menopausal } =>{Race_cancer_history=Non-Hispanic-White_Yes}	38	92	1.02
4	{BIRADS_breast_density=scattered_fibroglandular_densities, Menopaus=post_menopausal } =>{Race_cancer_history=Non-Hispanic-White_Yes}	39	92	1.02
5	{BIRADS_breast_density=scattered_fibroglandular_densities, HRT=No, biopsy=Yes } =>{Race_cancer_history=Non-Hispanic-White_Yes}	40	92	1.01

5.1.2. Interpreting Rules

We can comprehend rule 1 in Table 4 as “If a person’s age range is between 30 and 34, with no breast cancer of first degree relatives, first menstrual cycle is below 12 years, age at first birth is nulliparous, breast density is heterogeneously dense, and body mass index is between 10 and 25 then the individual having race Asian-or-Pacific-Islander can be a breast cancer patient".

Rule 1 in Table 5 can be interpreted as “If a person’s age range is between 18 and 29 having first child birth at age within the range 20 to 24, and body mass index is between 25 and 30 then there is a very high chance that individual of Hispanic ethnicity could have breast cancer".

Similarly, we can interpret rule 1 in Table 6 as “If a individual’s breast density is scattered fibroglandular dense with no records of using hormone replacement therapy having post menopausal status, and no previous breast cancer biopsy then the individual having race non-Hispanic-White is a breast cancer patient".

5.2. Evaluation Criteria of Classification Model

To measure the performance of the super learner, which is a classification model, several evaluation metrics were considered, like Accuracy, Precision, Sensitivity / Recall, and Specificity [30]. These were derived from the confusion matrix, and used to the evaluation of the model, and are shown in Equation (1) through (4).

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{1}$$

$$Sensitivity/recall = TP/(TP + FN) \tag{2}$$

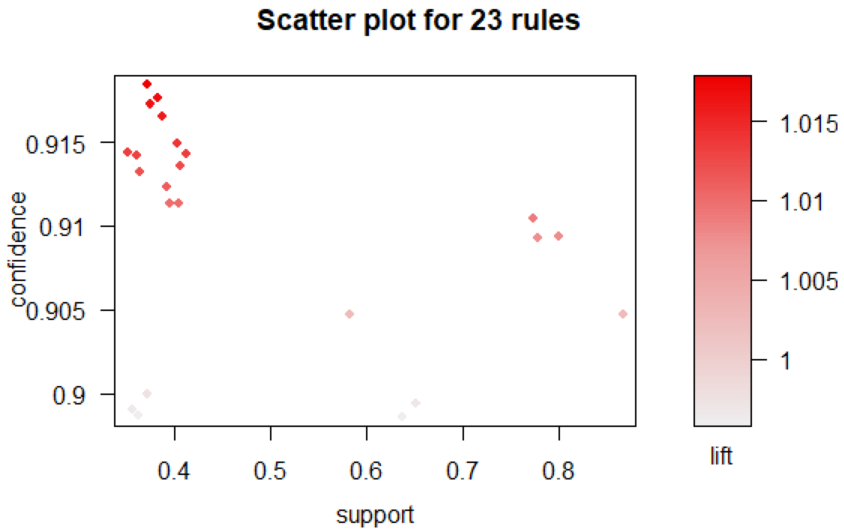

Figure 4. Scatter plot of 23 rules for breast cancer patients of Non_Hispanic_White race with specified support, confidence, and lift values.

$$Specificity = TN/(TN + FP) \tag{3}$$

$$precision = TP/(TP + FP) \tag{4}$$

where:

TP = number of positive instances that classified as positive
TN = number of negative samples accurately classified
FN = number of positive observations that classified incorrectly
FP = number of negative samples does not classified correctly

Also, the Area under the Receiver Operating Characteristic curve (ROC) was considered [30] and a detailed description of this metric can be found in [9].

The F1 measure is another popular performance metric for evaluating the performance of classification techniques, which is defined as given in Equation (5).

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

Also, the G-mean that shows the balance between classification performances on the majority and minority class were also considered. This metric consists of both positive and negative examples. G-mean can be described as the square root of the product of sensitivity and specificity that are shown in Equation (6).

$$G - mean = \sqrt{Sensitivity * Specificity} \tag{6}$$

5.2.1. Results of Super Learner

In this research, a comparison of the performance of the stacked ensemble or super learner (SL) method and the individual machine learning algorithms also named as base learners are conducted. We applied the SL methods on the training data discussed earlier and shown in Table 2. For the evaluation of the model, we used the test data set. Table 7 shows the performance (accuracy, precision, recall/sensitivity, specificity) of SL and three different machine learning (ML) methods on the test data.

Table 7. Performance (accuracy, precision, recall/sensitivity, specificity) of SL and three diverse ML techniques on the test data (**Bold** indicates the best value).

Algorithms	Accuracy (%)	Precision (%)	Recall/Sensitivity (%)	Specificity (%)
GBM	88.92	98.14	89.923	73.20
RF	88.88	98.12	89.12	73.02
DNN	**89.50**	97.67	**91.00**	65.95
SL	88.81	**98.17**	89.77	**73.72**

Table 8 shows the performance (AUC, F1, and G-mean) of SL along with three different ML algorithms on the test data.

Table 8. Performance (AUC, F1, and G-mean) of SL and three individual machine learning algorithms on the test data (**Bold** indicates the best value).

Algorithms	AUC	F1	G-mean
GBM	0.9234	0.9385	0.8113
RF	0.9198	0.9383	0.8102
DNN	0.7877	**0.9422**	0.7747
SL	**0.9247**	0.9378	**0.8135**

Comparing Table 7 and Table 8, if we consider predictive performance based on test data, we can see that for super learning best results were achieved. In Table 7, for accuracy and recall best values were obtained when the individual learner named DNN was applied, however, best values for precision and specificity were obtained using the SL method.

In Table 8, the performance measures AUC, F1, and G-means were listed that are considered important metrics for imbalanced data. In the case of F1, DNN provides the best value which was just slightly greater than SL. However, for AUC, and G-means SL provides the best values that are slightly greater than GBM, and RF methods but considerably greater than the DNN model.

6. Discussion

The data mining approach named rule discovery is very useful since rules can conveniently provide meaningful information. The class association technique was applied in paper [52] that discovered knowledge in the form of rules for both breast cancer and non-breast cancer patients from the BCSC risk factors data. From this information, risk factors associated with breast cancer can be learned. However, it is also important to know similar information based on a particular race. The current paper addresses this issue by identifying significant information in the form of rules for different ethnic groups of breast cancer patients. In machine learning, the classification that is considered as supervised learning is used to classify the unknown or target class as accurately as possible for each instance in the data. Different classification algorithms were applied to the breast cancer risk factors data [25]. In this paper, we tried to improve the performance of the model by applying the super learning method.

The studies that we conducted has a few limitations. First, the BCSC risk factors data that we investigated for this research is robust, however, we did not have information about the overall quality of the data. Second, as there are very few numbers of breast cancer patients of native-American, non-Hispanic black, and other-or-mixed in BCSC risk factors data set, these ethnic groups had to be removed during the rule mining process. During the class association rule mining step, we considered non-Hispanic white, Hispanic, and Asian-or-pacific-islander races. In addition, among these three races the number of breast cancer patients for Hispanic and Asian-or-pacific-islander were very low compared to the non-Hispanic white group. To address this issue, we specified multiple support values for breast cancer patients of both Hispanic and Asian-or-pacific-islander; we specified a very low minimum support as there are few numbers of instances for these two groups. In literature [5] [31], researchers applied the same concept by specifying multiple support values for rare item problems and by applying the same idea such as setting a low support value, we extracted rules for both Hispanic and Asian-or-pacific-islander that are infrequent in the risk factors data. Although minimum support values were very low for breast cancer patients of specified races, however, confidence value that indicates the predictive strength of the rules were assigned high. Third, for a classification model, we used resampled training data that was obtained using SMOTE and ENN techniques. This was done as data used for this study were highly imbalanced - unequal distributions between cancer and non-cancer individuals. By using the super learning approach, we achieved acceptable performance, however, to improve the performance further, more investigation is needed to find or develop appropriate resampling techniques for this particular data set; also different cost-sensitive techniques can be investigated.

7. Conclusion

In this research, a data mining technique named class association rule discovery and a machine learning method called super learning have been investigated for breast cancer risk factors data. In the first part of this study, rule discovery from different ethnic groups of breast cancer patients was done. By using these rules or knowledge, appropriate strategies targeting particular races can be developed. Besides, medical professionals or healthcare organizations can inform vulnerable individuals about their risk. Ultimately, the mortality of breast cancer can be reduced by early detection of cancer cases. Classification is one of the significant tasks of machine learning that correctly classify the target class for each instance in the data. The second part of this paper focused on enhancing the performance of the classification model by using the super learning technique consisting of three diverse algorithms.The results show that for breast cancer risk factors data, super learning provides better predictive performance compared to the individual three machine learning algorithms that were selected as the base learners for this research.

This research can be improved by discovering rules for breast cancer patients of other ethnic groups and extracting knowledge from different ethnic groups for individuals with no breast cancer. Moreover, from the classification perspective this work can be extended by using more diverse techniques with optimal parameters to improve its performance. In addition, as super learning generally provides better performance than the individual learner, the technique can be applied to other research problems.

Author Contributions:
Md Faisal Kabir ran all the experiments for this research and wrote the first draft, Simone A. Ludwig reviewed manuscript after the first draft & made necessary changes, and Abu S Abdullah reviewed and edited discussion part. All authors read and approved the final manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

WHO - World health Organization
ACS - American Cancer Society
ARM - Association Rule Mining
GBM - Gradient Boosting Machine
RF - Random Forest
DNN - Deep Neural Network
GLM - Generalized Linear Model
HRT - Hormone Replacement Therapy
BMI - Body Mass Index
KNN - k-nearest-neighbor
ML - Machine learning
BCSC - Breast Cancer Surveillance Consortium
SL - Super learner / learning

References

1. J. Ferlay, et al. “Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." International journal of cancer 136.5:E359-E386, 2015.
2. J. Han, M. Kamber. “Data mining concept and technology." Publishing House of Mechanism Industry: 70-72, 2001.
3. S. M. Monzurur Rahman, Md. F. Kabir, and F. A. Siddiky. “Rules mining from multi-layered neural networks." International Journal of Computational Systems Engineering1.1: 13-24, 2012.
4. R. Agrawal, T. Imieliński, and A. Swami. “Mining association rules between sets of items in large databases." Acm sigmod record. Vol. 22. No. 2. ACM, 1993.
5. Kabir, Md Faisal, Simone A. Ludwig, and Abu Saleh Abdullah. "Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.
6. K. Majid, and S. T. Tabibi. “Breast mass association rules extraction to detect cancerous masses." Technology, Communication and Knowledge (ICTCK), 2015 International Congress on. IEEE, 2015.
7. C. Ordonez, C. A. Santana, L. De Braal. “Discovering Interesting Association Rules in Medical Data." ACM SIGMOD workshop on research issues in data mining and knowledge discovery, 2000.
8. S. Stilou et al. “Mining association rules from clinical databases: an intelligent diagnostic process in healthcare." Studies in health technology and informatics 2: 1399-1403, 2001.
9. Kabir, Md Faisal, and Simone A. Ludwig. "Enhancing the Performance of Classification Using Super Learning." Data-Enabled Discovery and Applications 3.1 (2019): 5.
10. Agrawal, Rakesh, et al. An interval classi er for database mining applications. Proc. of the VLDB Conference. 1992.
11. Rahman, SM Monzurur, Md Faisal Kabir, and Muhammad Mushfiqur Rahman. Integrated Data Mining and Business Intelligence. Encyclopedia of Business Analytics and Optimization. IGI Global, 2014. 1234-1253.
12. Kaur, Harnoor, and Shalini Batra. HPCC: An ensembled framework for the prediction of the onset of diabetes. Signal Processing, Computing and Control (ISPCC), 2017 4th International Conference on. IEEE, 2017.
13. Gibbons, Chris, et al. Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. Journal of medical Internet research 19.3 (2017).
14. Silwattananusarn, Tipawan, Wanida Kanarkard, and Kulthida Tuamsuk. Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble. Journal of Computer and Communications 4.04 (2016): 20.
15. van der Laan, Mark J., Eric C Polley and Alan E. Hubbard. Super Learner Statistical Applications in Genetics and Molecular Biology, 6.1 (2007): -. Retrieved 19 Mar. 2018, from doi:10.2202/1544-6115.1309.
16. Van der Laan, Mark J., and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media, 2011.

17. Vanerio, Juan, and Pedro Casas. Ensemble-learning approaches for network security and anomaly detection. Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks. ACM, 2017.
18. Aiello, Spencer, et al. Machine Learning with Python and H20. H2O. ai Inc (2016).
19. N. Hou, et al. "Hormone replacement therapy and breast cancer: heterogeneous risks by race, weight, and breast density." Journal of the National Cancer Institute 105.18:1365-1372, 2013.
20. M. H. Gail et al. "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." JNCI: Journal of the National Cancer Institute 81.24: 1879-1886, 1989.
21. W. E. Barlow, et al. "Prospective breast cancer risk prediction model for women undergoing screening mammography." Journal of the National Cancer Institute 98.17: 1204-1214, 2006.
22. E. Gauthier et al. "Breast cancer risk score: a data mining approach to improve readability." The International Conference on Data Mining. CSREA Press, 2011.
23. Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple class-association rules. In: Proceedings of the 2001 international conference on data mining. San Jose, CA, US; 2001. p. 369–76.
24. P. Razan et al. "Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain." Journal of biomedical informatics 48: 73-83, 2014.
25. Kabir, Md Faisal, and Simone Ludwig. "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.
26. Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C).A list of the BCSC investigators and procedures for requesting BCSC data for research purposes, last retrieved July 2018 from http://www.bcsc-research.org.
27. P. Razan et al. "Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain." Journal of biomedical informatics 48: 73-83, 2014.
28. Nykodym, Tomas, et al. Generalized Linear Modeling with H2O. Published by H2O. ai, Inc (2016).
29. LeDell, Erin. Scalable Super Learning. Handbook of Big Data 339 (2016).
30. Fawcett, Tom. An introduction to ROC analysis. Pattern recognition letters 27.8 (2006): 861-874.
31. B. Liu, H. Wynne, and M. Yiming. "Mining association rules with multiple minimum supports." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.