International Journal of Management Science and Information Technology
IJMSIT

E-ISSN: 2774-5694 P-ISSN: 2776-7388 OPEN ACCESS

Volume 4 (2), July-December 2024, 312-320
DOI: https://doi.org/10.35870/ijmsit.v4i2.3061

Application of C4.5 algorithm with PSO Feature Selection and Bagging Technique on Breast Cancer Classification

Fika Ulfa Widowati
Department of Digital Business, Faculty of Economics and Business, Universitas 17 Agustus 1945 Semarang, Indonesia

Email: fikaulfa-widowati@untagsmg.ac.id

Article history:
Received August 5, 2024
Revised August 7, 2024
Accepted August 8, 2024

Abstract

The second greatest cause of death for women worldwide is breast cancer. When abnormal cells in the body proliferate out of control, cancer is the result. The diagnosis of breast cancer was established using anthropometric data obtained from standard blood tests. From the UCI Machine Learning Repository, the Breast Cancer Coimbra Data Set was obtained and put to use. One popular classification decision tree method is the C4.5 approach. By selecting the appropriate features and applying the appropriate strategy to address class imbalance throughout the classification process, the performance of the C4.5 algorithm may be enhanced. To determine the accuracy of the classification, tests are conducted using a confusion matrix. Accuracy in this study is anticipated to increase with the application of the C4.5 Algorithm, the Bagging approach to address class imbalance, and the PSO feature selection method. The C4.5 Algorithm, PSO, Bagging Technique produce the best accuracy results, with an average of 86.36 percent. The C4.5 classification method has the second highest accuracy, with a PSO accuracy of 79.39 percent. Utilizing the Bagging Technique in conjunction with the C4.5 Algorithm, the accuracy of 75.0% is the third highest. Furthermore, it has a 65.71 percent accuracy with the C4.5 categorization. As a result, the increase in accuracy from before adding PSO and Bagging Technique was 20.65%, indicating that the inclusion of PSO and Bagging Technique had a substantial impact on the calculation process.

Keywords:
Classification; C4.5 Algorithm; Decision Tree; PSO; Bagging; Breast Cancer.

1. INTRODUCTION

Cancer is an illness brought on by the body's unchecked, aberrant cell growth. This abnormal cell growth can damage normal cells around it and in other parts of the body (E. j. Corwin, 2009). The second leading cause of death worldwide is cancer, as it generally does not cause symptoms at the beginning of its development, so it is only detected and treated when it reaches an advanced stage (Breast cancer, 2018). One of the main causes of cancer is genetic mutation in cells. One type of cancer is breast cancer. To receive immediate treatment and have an 80–90% chance of recovery, breast cancer must be detected early. Detection of breast cancer using several methods, including a medical interview, a physical examination in person to detect changes in the breast and lymph glands, as well as a supporting examination using a medical device. Further examinations can be performed with mammograms, ultrasound (USG) of the breast to determine a solid mass or cyst containing fluid, biopsy with sampling of tissue to be examined in the laboratory and to determine whether the cell examined is benign or malignant. A ten-year study demonstrates a positive correlation between breast cancer mortality rates and non-functioning glucose absorption (S. M. Samuel et al., 2018). A study in Brazil found that a high body mass index (BMI) accounts for 3.8% of newly diagnosed cancers in the country. And from observations, a higher BMI is greater in women, especially in cases of breast cancer (L. F. M. de Rezende et al., 2018).

In order to identify correlations, data mining combines statistical and mathematical methods with pattern recognition technologies. Thus, it is possible to diagnose breast cancer using data mining techniques (O. Maimon and L. Rokach, 2005). In data mining, classification is a type of data analysis procedure used to extract models that characterize data classes. The acquired data will go through two phases of training and testing in the categorization process. The data will be categorized and shown in pattern form throughout the training phase. Then at the testing stage, the data is analyzed with algorithms or models for classification. This phase serves to determine the accuracy derived from the data based on the classification rules.

Feature selection is a part of preprocessing that aims to reduce the complexity of attributes at the time of data analysis. Feature selection aims to reduce the number of dimensions in data sets by identifying relevant features while maintaining prediction accuracy (M. H. Aghdam, S. Heidari, 2015). In certain situations, Particle Swarm Optimization (PSO), a metaheuristic optimization technique for feature selection, has been shown to outperform genetic algorithms, particularly in the optimization domain (B. Xue et al., 2012). In addition to having too many features available, data imbalances on datasets are also common. Algorithm performance has been demonstrated to be negatively impacted by data imbalance, one classification issue (T. W. Cenggoro, 2018). Bagging and Boosting are the two most used ensemble techniques (D. Opitz and R. Maclin, 1999). Bagging techniques are better than boosting when dealing with noise-containing data (T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, 2010). Besides being easy to develop, the Bagging technique is also strong in dealing with class imbalances if applied properly (W. Feng et al., 2018). The bagging technique can be applied to the decision tree method to increase the accuracy value to be produced later (C. D. Sutton, 2005). Based on the above problem description, then the accuracy of breast cancer classification can still be improved by using the C4.5 algorithm that involves the PSO method for feature selection and Bagging technique to overcome the imbalance class, so that the results obtained are better than previous studies with a higher degree of precision.

Based on the background described earlier, there is a need for optimization in the treatment of breast cancer by selecting the right method of classification. The classification method used previously to increase accuracy belongs still low. For that, the problem is how to improve the accuracy of breast cancer classification using the C4.5 algorithm with feature selection on Particle Swarm Optimization (PSO), as well as Bagging Techniques for the balance class.

Based on the existing problem formula, the aim of this study is to improve the performance of the C4.5 algorithm with feature selection, as well as finding the right method to deal with class imbalances on breast cancer diagnosis datasets in order to get better accuracy results. This research is expected to contribute to the society as well as the current scientific and technological developments, namely: to provide an overview to the public about the importance of breast cancer detection with existing parameters as a means for early-stage prevention, to produce the best accuracy with the C4.5 classification model using PSO method as feature selection and to be able to handle data that imbalance so that it can be used for breast cancer diagnosis with maximum results.

2. RESEARCH METHOD

Some studies discussed the identification of breast cancer using several methods, including the C4.5 classification method. Research related to the accuracy rate of breast cancer detection is like the research conducted by Arnab Kumar Mishra, et al (Arnab Kumar Mishra, Pinki Roy, and Sivaji Bandyopadhyay, 2020) pointing out that the best prevention efforts depend on an early diagnosis of breast cancer. Discrete optimization performance is enhanced by using Binary Particle Swarm Optimization (BPSO) feature-based selection for breast cancer prediction. The prediction system's effectiveness is assessed using a number of metrics, including the average prediction accuracy, sensitivity, specificity, and Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve. The Coimbra Dataset (BCCD) and the Wisconsin Diagnostic Dataset (BCWDD) have average prediction accuracy of 80.83% and 98.24% respectively, results obtained after feature selection. Pre-function selection and feature selection have increased by 4-8% and 1-4%.

Research done by Farah Sardouk, et al (2019) it aims to prevent breast cancer with early-stage screening using a number of parameters including BMI, age and regular blood sugar testing. Breast cancer datasets tested with AdaBoost, Rregression, Random Forest, RBFNN, Jrip, J48 algorithms. The highest accuracy of the classification of both cases (precision) using the Classification Via Rregression method in the study was 76.2%.

On the research Maria Inês Cruz, Jorge Bernardino (2019) mentioned that there is a need for invasive treatment of breast cancer because this type of cancer has spread around normal tissue. The highest accuracy in this study is using the logistic regression method of 74%, then using the ensemble technique with several algorithms obtaining a 95% AUC value and a precision of 87%.

Research conducted by Raka Hendra Saputra dan Budi Prasetyo (2020) explains the difficulty of identifying breast cancer that is done manually, and requires a health support tool (biomarker). The methods used in this study are PSO and bagging as feature selection to address class imbalances, therefore in the classification process, the C4.5 algorithm's performance is more ideal. A confusion matrix is used to test classification and determine the accuracy outcome. Using C4.5 algorithms, PSO, and the bagging technique, accuracy is increased to 5.11% from 93.43% to 98.54% at first.

M A Muslim et al (2018) investigating the most prevalent kind of breast cancer in females. In order to improve therapy, the study attempts to ascertain the percentage of breast cancer cases that are diagnosed early. A Wisconsin Breast Cancer (WBC) dataset from UCI machine learning was employed in the study. Particle Swarm Optimization (PSO) and C4.5 algorithms were utilized in this work to optimize the C4.5 methodology and select features. 10-Fold-Cross validation is employed as a confusion matrix and validation technique. PSO with the C4.5 algorithm can increase the result's accuracy by 0.88%.

Research done Muhammad Usman Ghani, et al (2019) explains about breast cancer caused by abnormally growing breast cells. Anthropometric information and characteristics obtained from standard blood tests are used to forecast breast cancer. The Recursive Feature Elimination approach may be used to identify the most significant characteristic in the dataset as a biomarker. (RFE). The most effective biomarkers for breast cancer are age, BMI, glucose, HOMA, and resistin. They employed KNN, ANN, decision trees, and naive Bayesian classification approaches. The Artificial Neural Networks approach provided the best accuracy of 80.00% in this investigation.

On the research Yolanda D. Austria, et al (2019) Explaining that the breast cancer detection device (X-ray mammography) used in the early stages was considered less optimal compared to routine blood sampling for breast cancer. Eleven classification algorithms, including Random Forest (RF), Decision Tree (DT), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Gradient Boosting Method (GBM), and Naive Bayes, were used by researchers to diagnose breast cancer. (NB). Breast cancer prediction is identified by hyper-parameter modification. With an accuracy rate of 74.14%, the machine learning technique Gradient Boosting (GB) is the best classifier for predicting breast cancer using the Coimbra Breast Cancer Dataset. The Nonlinear Support Vector Machine (SVM) classifier generates the quickest testing time of zero seconds, while the k-Nearest Neighbor (kNN) classifier gives the fastest training time of 0,000598 seconds. With a 50% accuracy rate, body mass index (BMI) values are the best predictor, followed by glucose.

2.1. Data Mining

In order to extract and identify relevant information and related knowledge from a variety of large databases, data mining is a process that makes use of statistical techniques, mathematics, artificial intelligence, and machine learning (Joko Purnomo, 2014). Data mining is a method used for a large database. The primary purpose of data mining is to predict. In data mining processing there are many aspects, one of which is data sets that have a big influence. The characteristics of the data set include data records, graph-based data, and sequential data. The main components of data mining are Databases, data warehouses, or database servers.

2.2. Decision Tree

Tree of Decisions A tree is a type of data structure made up of an edge, or rib, and a node. The decision tree node is composed of three components: the node root, the intersection knot, and the end knot (F.A.Himawati, 2013). Decision tree is a hierarchical or tree-structured prediction model. One style of flowchart that has a tree-like form is the Decision Tree. The leaf node depicts the distribution of classes inside or across classes, each internal node represents a test on an attribute, and each branch shows the result of the test. A hierarchical or tree-structured prediction model is called a decision tree. Unclassified records can be categorized using the decision tree by selecting the branch that matches the test findings. This will place the internal node—a node having one intake and two or more outlet branches—where another attribute or leaf test is necessary. The record whose class is not known is subsequently assigned a class that matches the leaf node's class. A class label is indicated on the decision tree by every leaf node. Transforming the data from a table into a tree and subsequently the tree's model into a rule is the procedure involved in creating a decision tree.

2.3. C4.5 Algorithm

The C4.5 technique is used to generate the decision tree from the data (Rahmayuni, 2014). The ID3 algorithm, which is also a decision tree construction method, provides the foundation for the C4.5 algorithm. Until there are no more branches, every decision node is examined repeatedly by the C4.5 algorithm, which then chooses the optimal intersection. To choose the root property by considering all possible characteristics with the highest gain value. The C4.5 algorithm chooses the best interference by utilizing the ideas of entropy reduction or information gain. When used properly, data mining is not a standalone technique but rather a component of the Knowledge Discovery in Database (KDD) process. One of the crucial phases in the KDD process is data mining, which is particularly relevant to the process of calculating and extracting patterns from the survey data. The decision tree's rules will combine to create an if-then condition. These rules are obtained by tracing the decision tree from root to leaf. Each node and condition of the intersection forms a condition or a if, whereas for the values present on the sheet will form a result or a then.

2.4. Particle Swarm Optimization

Because Particle Swarm Optimization (PSO) has characteristics with Genetic Algorithms (GA), PSO is frequently employed in research. The PSO's benefits are simple to use, and there are several adjustable factors. Starting with a population of random solutions, the PSO system updates the results of each generation in pursuit of the optimal point. The method found answers by applying more methodical mathematics. Kennedy and Edward developed the concept of particle swarm optimization (PSO) in 1995. This algorithm's thought process was influenced by the social behavior of animals, such a school of fish or a flock of birds (Muhammad Ali Ridla, 2018). Unlike GA, PSO doesn't have evolutionary operators like crossovers and mutations. The line in the matrix is called a particle (sama dengan kromosom GA). They contain coded variable and non-binary values. Each particle moves around the surface of the particle at a speed.

2.5. Bagging Technique

Combining bootstrapping and aggregating is known as bagging. By altering the number of elements or resampling the same number of elements as in the original dataset, the bootstrap sample is produced (E. Alfaro et al., 2013). This method is part of supervised learning that aims to improve the accuracy of predictions (O. Somantri et al., 2014). The process of improving the accuracy of the classification algorithm on the Ensemble method is by building several classifiers from the data training then at the time of classification using the voting/aggregating of those classifier-classifiers (K. Tan et al., 2006).

In this study, the datasets are taken from the UCI Machine Learning Repository's public data. Based on the results of several previous studies, the C4.5 method showed good performance especially in some studies related to classification. On the study “Classification of Breast Cancer Using Data Mining” (Farah Sardouk et al., 2019) It explains about breast cancer prevention by examining the early stages using a number of parameters including BMI, age, and blood sugar testing on a regular basis. The parameters used have a major influence on the classification process. Next M a Muslim et al (2018) using Particle Swarm Optimization (PSO) as a feature selection method to optimize the C4.5 algorithm and doing research on breast cancer. Confusion matrix and validation technique both use 10-Fold-Cross Validation. Using the C4.5 algorithm and PSO can improve accuracy. The same research was done by Raka Hendra Saputra and Budi Prasetyo (2020) but in this study, the researchers added the Bagging method as a feature selection to give more optimal results. The Bagging function is to overcome class imbalance, consequently the C4.5 algorithm operates in the classification process more effectively. From some of these studies, it shows some weaknesses associated with the use of data that imbalance and selection of features that allow irrelevant features may reduce the performance of classification. That could cause the accuracy of C4.5 to be less than maximum. The study used the Bootstrap Aggregating method to address class imbalances in datasets. It is hoped that using such there are techniques to increase the classification accuracy for breast cancer.

3. RESULTS AND DISCUSSION

3.1. Data Collection

The Machine Learning Repository (UCI) website provides a public dataset that is used in the research with the following link https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra that contains breast cancer data obtained from the University Hospital Centre of Coimbra. Clinical characteristics were observed from 64 breast cancer patients and 52 healthy controls. The predictors are anthropometric data and the parameters used from the analysis of blood tests, as well as the data taken before surgery and treatment. The category variables in the data set show values 1 and 2, each functioning to divide into two categories: healthy patients and breast cancer patients. The data set includes 116 patient records with ten quantitative predictors, including: Age (years), BMI (kg/m2), Glucose (mg/dL), Insulin (µU/mL), HOMA (Homeostasis Assessment Model), Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), MCP-1(pg/dL), Classification (1=Healthy controls,2=Patients).

Table 1. Breast Cancer Dataset

Quantitative Attributes	Data Type	Unit	Value Range
Age	Integer	Years	24-89
BMI	Real	kg/m2	18,37-38,579
Glucose	Integer	mg/dL	60-201
Insulin	Real	µU/mL	2,432-58,46
HOMA (Homeostasis Assessment Model)	Real		0,467-25,05
Leptin	Real	ng/mL	4,311-90,28
Adiponectin	Real	µg/mL	1,656-38,04
Resistin	Real	ng/mL	3,21-82,1
MCP-1	Real	pg/dL	45,843-1698,44
Classification (1=Healthy controls,2=Patients)	Integer		1 or 2

The characteristics of the datasets described in Table 1 are multivariate with integer and real attributes that are classified in breast cancer diseases. In this study there are two classes: Healthy controls and Patients.

Table 2. Description Dataset

Dataset Characteristics	Multivariate	Number of Instances	116
Atribute Characteristics	Integer, Real	Number Of Atributes	10
Associated Tasks	Classification	Missing Values	N/A

From Table 2 above it describes a data set of breast cancer that has the attribute characteristics of Integer and Real with data as many as 116 taken from healthy control patients and breast cancer patients. There are 10 attributes in that data set.

3.2. Test Results Feature Selection PSO

Currently, a preprocessing step is completed before testing is conducted using the C4.5 classification algorithm. The outcomes of the C4.5 categorization will be contrasted with the C4.5-algorithm optimization results using PSO and Bagging approaches. The C4.5 technique will be used to process datasets used in research after they have successfully completed the feature selection preprocessing step. The classification C4.5 is used in the test step of the C4.5 method. The C4.5 classification process using 10-Fold-Cross Validation will be carried out with the process of separating training data and data testing. Tests are carried out to obtain maximum accuracy values using rapidminer tools. This new particle is then inserted again into the upper circuit and compares the highest corresponding accuracy. The particles with the highest fitness will be pbest and gbest. As soon as the x values are close to each other and converge with each other then iteration will be stopped. The processing of data will be done in the discussion with the help of tools Rapidminer.

3.3. Test Results Technical Bagging

In order to compare the test results of the Bagging methodology with the prior C4.5 method, the C4.5 classification method was used to the breast cancer data set obtained from the UCI Machine Learning Repository, which has an accuracy value of 65.71%. The experiment was done using the same tools as the rapidminer. Here are the calculation steps using the bagging technique.

Step 1 : Execute the bootstrap process on the training data by dividing the data according to the number specified by the bag. In this study used 100 bags.
Step 2 : Classify each bag using the C4.5 classifier to get the model.
Step 3 : The model was then tested using testing data.
Step 4 : Each bag produces accuracy, then selects all the results.
Step 5 : The final result is the result based on the most votes.

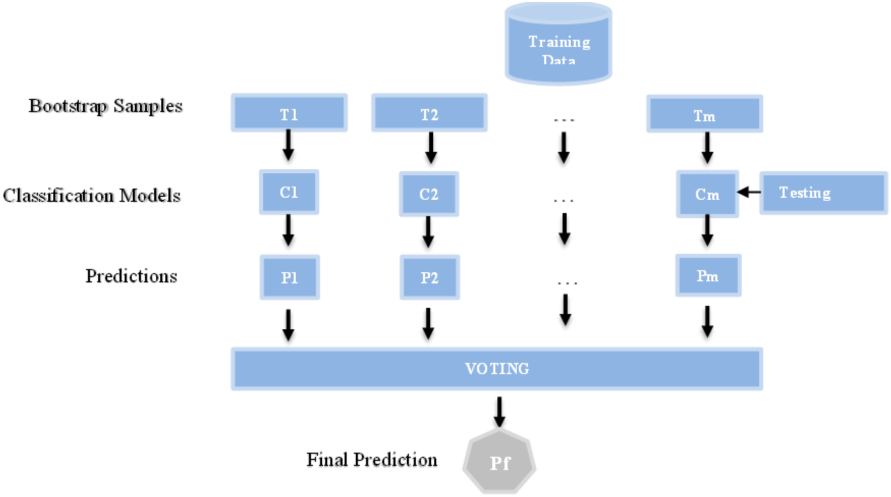

Figure 1. Scheme of Bagging Techniques (Faisal, 2016)

Scheme The above bagging technique describes about the process or steps of each stage starting from the data training. Training data using data samples (bootstrap). Then the data is classified by defining the model first (classification models). Once the classification process is done using the appropriate model, then all the results of the precision can be selected, based on the outcome of the vote. Voting is done to determine the final prediction based on the largest number of votes.

3.4. C4.5 Algorithm Test Results

In order to get the same classification results as the previous approach, or to compare with it, the test results using the breast cancer data set in this experiment were performed using the C4.5 algorithm classification method.

3.5. Evaluation Results with Confusion Matrix

This phase is carried out in all processes, from pre-processing to classification. Procedure for assessing and verifying categorization performance outcomes. This study used a confusion matrix for evaluation in order to determine how well the classification findings performed, and 10-fold cross-validation was used for validation. At the time of validation, the sequence of the existing set of documents will be checked. This is intended to avoid grouping documents from specific categories. From the experiments that have been carried out using several classification methods namely C4.5, PSO and Bagging Technique, the highest values were obtained on the classification C4.5 and PSO, with an accuracy of 86.36%. With previously performed calculations before preprocessing with data of 116 data with true negative values of 57 data, false positive data of 9, false negative of 7 and true positive of 43 data.

3.6. Results of Model Test Evaluation

The assessment and verification procedure for classifying performance outcomes. In this work, the confusion matrix was used to evaluate the performance of the classification findings, and 10 fold cross-validation was used for validation.

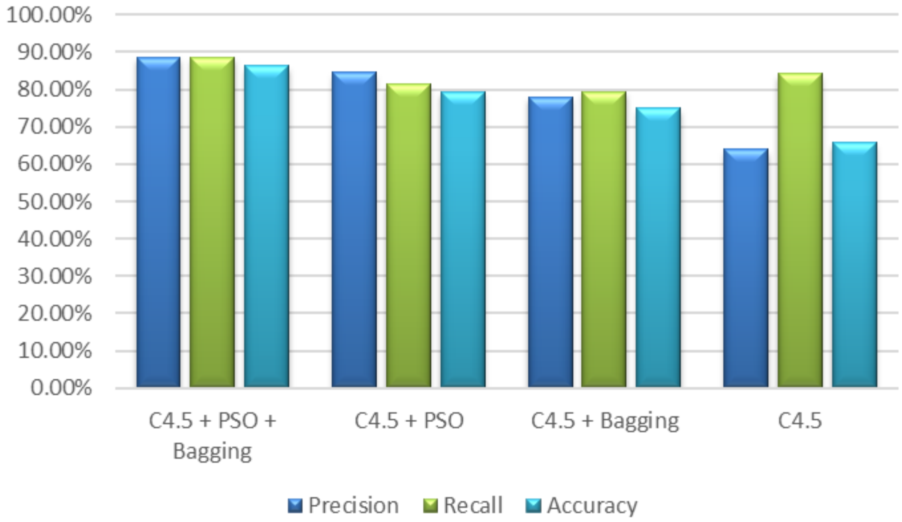

Figure 2. Comparison Diagram of Precision, Recall and Accuracy Values on C4.5 Algorithms, PSO and Bagging Techniques

From the graph in Figure 2 above it can be concluded that breast cancer data sets using feature selection on PSO and algorithm C4.5, as well as Bagging Technique have the highest value or accuracy of 86.36% with 116 datasets. The C4.5 algorithm with PSO technique earned the second-best accuracy, at 79.39%. The third achieved a 75.00% accuracy rate by using the Bagging technique in conjunction with the C4.5 algorithm. The data set was then tested using the C4.5 method, which had an accuracy of 65.71%.

4. CONCLUSION

Using the Breast Cancer Coimbra Data Set, the C4.5 algorithm was utilized in conjunction with Particle Swarm Optimization (PSO) for feature selection and bagging techniques for balancing class in the study conducted to detect breast cancer. the procedure for dividing a PSO's original nine qualities into five during the data set feature selection phase. It is used to resolve class imbalance on data sets, in contrast to the bagging approach used in research, and it yields the best bag of the vote result, making the C4.5 algorithm's performance more optimum in terms of enhancing accuracy. With the C4.5 method, 65.71% accuracy was achieved without the use of PSO or the bagging approach. In contrast, the accuracy result is 86.36% when PSO and Bagging methods are used to algorithm C4.5. The accuracy comparison that was achieved indicates a 20.65% improvement. Thus, it can be said that PSO and bagging strategies are crucial for enhancing the accuracy of the C4.5 algorithm by maximizing its performance.

ACKNOWLEDGEMENTS

Fika Ulfa Widowati from the University of 17 Augustus 1945 Semarang wrote this journal paper. based on studies showing how the C4.5 algorithm, PSO Feature Selection, and Bagging Technique may be applied to the classification of breast cancer. The UCI Machine Learning Repository was acknowledged by the authors for supplying the datasets utilized in this investigation. This repository is a very valuable resource for the machine learning community and has been very helpful in testing and validating the proposed methods. We are grateful to the RapidMiner development team for offering a robust and intuitive data analysis environment that is so useful for experimenting with and putting the C4.5 algorithm, PSO Feature Selection, and bagging approaches into practice. The author alone is responsible for all of the journal's content.

REFERENCES

Arnab Kumar Mishra, Pinki Roy, and Sivaji Bandyopadhyay. (2020). Binary Particle Swarm Optimization Based Feature Selection (BPSO-FS) for Improving Breast Cancer Prediction. Proceedings of International Conference on Artificial Intelligence and Applications ICAIA. India: Springer. doi: https://doi.org/10.1007/978-981-15-4992-2_35

B. Xue, M. Zhang, and W. N Browne. (2012). Particle swarm optimization for feature selection in classification: A multi-objective approach. transactions on cybernetics, 43(6), 1656-1671.

Breast cancer. (2018). Acesso em 24 de 9 de 2018, disponível em http://www.who.int/cancer/prevention/diagnosis screening/breast-cancer/en/

C. D. Sutton. (2005). Classification and regression trees, bagging, and boosting. Handbook of statistics, 24, 303-329.

D. Laudisio et al. (2018). Obesity and Breast Cancer in premenopausal women: Current evidence and future perspectives. European Journal of Obstetrics & Gynecology and Reproductive Biology.

D. Opitz and R. Maclin. (1999). Popular ensemble methods: an empirical study. Journal of Artificial Intelligence, 11, 169-198.

E. Alfaro, M. Gámez, and N. Garcia. (2013). Adabag: and package for classification with boosting and bagging. Journal of Statistical Software, 54(2), 1- 35.

E. j. Corwin. (2009). Saku Patofisiologi Corwin. Jakarta: Aditya Media.

F.A. Himawati. (2013). Data Mining. Yogyakarta: ANDI.

Farah Sardouk, Dr. Adil Deniz Duru, Dr. Oğuz Bayat. (4 de 1 de 2019). Classification of Breast Cancer Using Data Mining. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 51(1), 38-46. Fonte: https://asrjetsjournal.org/

I. A. Gheyas, and L. S. Smith. (2010). Feature subset selection in large dimensionality domains. Pattern Recognition, 43(1), 5-13.

J. Li and X. Han. (2018). Adipocytokines and breast cancer. Curr. Probl. Cancer, 42(2), 208–214.

Joko Purnomo, W. L. (2014). Implementasi Algoritma C4.5 Dalam Pembuatan Aplikasi Penunjang Keputusan Penerimaan Pegawai CV. DINAMIKA ILMU. TIKomSiN, 24-32.

K. Tan, Pang, N., Michael, S. & Vipin. (2006). Introduction To Datamining. Boston: Pearson Addison Wesley.

L. F. M. de Rezende et al. (February de 2018). The increasing burden of cancer attributable to high body mass index in Brazil. Cancer Epidemiol, 54, 63–70.

M A Muslim et al. (2018). Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis. International Conference on Mathematics, Science and Education 2017 (ICMSE2017), 1-7. doi: https://doi.org/10.1088/1742-6596/983/1/012063

M. H. Aghdam, S. Heidari. (2015). Feature selection using particle swarm optimization in text categorization. Journal of Artificial Intelligence and Soft Computing Research, 5(4), 231-238.

Maria Ines Cruz, Jorge Bernardino. (2019). Data Mining Techniques for Early Detection of Breast Cancer. In Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019), (pp. 434-441). Coimbra, Portugal. doi: https://doi.org/10.5220/0008346504340441

Much. Rifqi Maulana, M. Adib Al Karomi. (2015). Information Gain Untuk Mengetahui Pengaruh Atribut Terhadap Klasifikasi Persetujuan Kredit. J. Litbang Kota Pekalongan, 9, 113–123.

Muhammad Ali Ridla. (Juni de 2018). Particle Swarm Optimization Sebagai Penentu Nilai Bobot Pada Artificial Neural Network Berbasis Backpropagation Untuk Prediksi Tingkat Penjualan Minyak Pelumas. Jurnal Ilmiah Informatika, 3(1). Fonte: ejournal.amili.ac.id

Muhammad Usman Ghani, et al. (2019). Comparison of Classification Models for Early Prediction of Breast Cancer. 2019 International Conference on Innovative Computing (ICIC) , 1-6. Fonte: IEEE 978-1-7281-4682-9/19

N. Rout, D. M. (2018). Handling imbalanced data: a survey. International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications, 431-443.

Nur Fitriyah, Budi Warsito, Di Asih I Maruddani. (2020). Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (SVM). JURNAL GAUSSIAN, 9(3), 376 -390. Fonte: https://ejournal3.undip.ac.id/index.php/gaussian

O. Maimon dan L. Rokach. (2005). Data mining and knowledge discovery handbook.

O. Somantri, G. W. Sasmito, M. S. Sungkar, And Erwadi. (2014). Optimalisasi Neural Network Dengan Bootstrap Aggregating (Bagging) Untuk Penentuan Prediksi Harga Listrik. Scientific Journal of Informatics, 1(2), 185-192.

Pat Langley, Stephanie Sage. (1994). Induction Of Selective Bayesian Classifier. Proceeding Of the Tenth Conference on Uncertainty In Artificial Intelligence, 399-406. doi: https://doi.org/10.1016/B978-1-55860-332-5.50055-9

Rahmayuni, I. (1 de April de 2014). Perbandingan Performansi Algoritma C4.5 Dan Cart Dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang. TEKNOIF, 2(1), 40-46.

Raka Hendra Saputra, Budi Prasetyo. (3 de September de 2020). Improve the Accuracy of C4.5 Algorithm Using Particle Swarm Optimization (PSO) Feature Selection and Bagging Technique in Breast Cancer Diagnosis. Journal of Soft Computing Exploration (JOSCEX), 1(1), 47-55. Fonte: https://shmpublisher.com/index.php/joscex

S. A. Alasadi And W. S. Bhaya. (2017). Review Of Data Preprocessing Techniques In Data Mining. Journal of Engineering and Applied Sciences 12, 12(16), 4102–4107. doi:Doi: 10.3923/Jeasci.2017.4102.4107

S. M. Samuel, E. Varghese, S. Varghese, and D. Büsselberg. (8 de 2018). Challenges and perspectives in the treatment of diabetes associated breast cancer. Cancer Treat, 70, 98–111.

Santosa, Budi, Paul Willy. (2011). Metoda Metaheuristik: Konsep dan Implementasi. Surabaya: Guna Wydia.

Selvia Lorena Br Ginting, W. Z. (15 de November de 2014). Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014, 263-272.

Siburian, B. R. (Agustus de 2014). Aplikasi Data Mining Untuk Menampilkan Tingkat Kelulusan Mahasiswa Dengan Algoritma Apriori. Pelita Informatika Budi Darma, 7(2), 56-61.

Suyanto. (2014). Algoritma Optimasi Deterministrik Atau Probabilistik. Yogyakarta: Graha Ilmu. Fonte: http://ruangbaca-if.unesa.ac.id/

T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 41(3), 552-568.

T. W. Cenggoro. (2018). Deep learning for imbalance data classification using class expert generative adversarial network. Procedia Computer Science, 135, 60- 67.

UCI Machine Learning Repository: Breast Cancer Coimbra Data Set. (20 de 9 de 2018). (Archive.ics.uci.edu) Acesso em 2021, disponível em https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

W. Feng, W. Huang, W, and J. Ren. (2018). Class imbalance ensemble learning based on the margin theory. Applied Sciences, 8(5), 815.

Wibowo, M. N. (16 de August de 2012). Implementasi dan Demo Pohon Keputusan ID3 dan C4.5 menggunakan PHP.

Yolanda D. Austria, et al. (2019). Comparison of Machine Learning Algorithms in Breast Cancer Prediction using the Coimbra Dataset. 23.1-23.8. doi: https://doi.org/10.5013/IJSSST.a.20.S2.23