Machine learning evaluations using WEKA

by

Thomas Johnson III

University Honors Thesis

Elizabeth City State University

University Honors Program

Elizabeth City, North Carolina

Spring 2020

ECSU
ELIZABETH CITY STATE UNIVERSITY
FOUNDED 1891

Machine learning evaluations using WEKA

By

Thomas Johnson III

Defense Date:
May 7, 2020

Approved:

Malcolm

Malcolm D'Costa PhD, Advisor
Assistant Professor of Computer Science
Dept. of Mathematics Computer Science, &
Engineering Technology
Elizabeth City State University

Krishna H. Kulkarni

Krishna H. Kulkarni, PhD Professor
Dept. of Mathematics Computer Science,&
Engineering Technology
Elizabeth City State University

Andre P. Stevenson

Andre P. Stevenson, PhD
Professor of Social Work
Director, University Honors Program
Elizabeth City State University

Table of Contents

Abstract

Computer science is a growing field and machine learning is a growing area within computer science. The development of various machine learning algorithms that have been created has been diverse. Using WEKA, the study used the mammography dataset to examine machine learning algorithms to explain what components of the machine learning algorithms may affect performance. The logistic regression model classified the most instances of the provided partitioned mammogram dataset. Results indicated an expansion in the assortment of machine learning algorithms would be employed generating a larger collection of models.

Keywords: machine learning, algorithms, WEKA, logistic regression

Introduction

Machine learning is a growing field in computer science. One which has had growing implications across the various domains of STEM. Machine learning utilizes various algorithms derived from mathematical and statistical functions and concepts for the purpose of allowing computers to process data for various purposes while using the data that is inputted to improve the algorithm. The more data that is available, the more effective the machine learning algorithm becomes. There is a considerable number of machine learning algorithms that have been developed for such goals as classification and prediction. This diverse pool of options for machine learning algorithms has evolved because of the necessity of various algorithms being implemented to contribute to various goals. Simply because a portion of machine learning algorithms was developed for the same purpose does not mean that the algorithms will be equally effective in the same scenario. For an overview of machine learning algorithms, the benefits of a specific algorithm, and the vulnerabilities of a specific algorithm, sources such as Types of machine learning algorithms by Ayodele written in 2010 can provide some information. For the purpose of this thesis, the focus will be exploring the most mammography on a given set of features.

Using the mammography dataset, we can examine different machine learning algorithms for the purpose of examining how effective a given machine learning algorithm will perform as opposed to another. Utilizing said machine learning algorithms will offer a chance to examine what components of each machine learning algorithms may affect the performance of the algorithm. The machine learning algorithms will be implemented through WEKA, a library in Java that provides access to machine learning algorithms within the Java programming language.

To do so, various machine learning algorithms will be used and evaluated using a select portion of variables to classify the level of distraction imposed by various scenarios. The results

will reveal which machine learning algorithm yields the best classification results for the mammography dataset and permit discussion as to why each machine learning algorithm performed as it did on the mammography dataset.

Proposed Hypothesis

The null hypothesis is that the machine learning algorithms will all perform at the same level of accuracy on the same dataset. The alternative hypothesis is there will be machine learning technique that performs better than the others on the same dataset.

Machine Learning Theory

Machine learning theory will be critical to the implementation of work revolving around machine learning. Machine learning theory is a composition of scenarios where a program must be able to become more efficient or effective based on data provided to said program (Mooney, pp. 1). Some of the critical chunks of machine learning theory can be observed to refer to long-term phenomena that can be seen across many instances of machine learning being employed (Mooney, pp. 1). One such observation concerns the quantity of computing assets that will be required to tackle a suggested scenario (Mooney, pp. 1). If a model has virtually unrestricted access to an inexhaustible pool of assets, then the model is likely to reach a state of high efficiency or correctness (Mooney, pp. 1).

Another basic idea of learning theory are the asset investment and configuration of the training phase that will provide for the best model (Mooney, pp. 1). Other than inexhaustible computing assets, pouring data into the training of the model to an infinite amount or significantly enormous amount should allow the model to become top tier in correctness or efficiency (Mooney, pp. 1). With the aim of attaining the highest level of correctness or efficiency that is plausible, a model can achieve such through development from mistakes made

during the training stage (Mooney, pp. 1). Training stage should be seen as the process of getting best model coming forth as a fusion of the previous model's successes and failures, although no state of pure perfection is certain (Mooney, pp. 1). Machine learning theory also encapsulates scenarios where no feasible program with given resources and time will be able to provide a model that is best for tackling said scenario (Mooney, pp. 1).

Machine learning is primarily associated in more recent times with Mr. Arthur Samuel, whose contributions are still revered today (McCarthy and Feigenbaum, 2020). Arthur Samuel's lifespan ranged from 1901 and concluded in 1990 (McCarthy and Feigenbaum, 2020). Machine learning theory is essentially all work that involves or encapsulates machine learning. The ideas of machine learning theory are being applied, juxtaposing practical implementations versus the boundless assumptions that are attached to the abstract ideas of machine learning theory. An ideal case would be that a project possesses a virtually inexhaustible quantity of data to train machine learning models on. The reality of the matter is that work has deadlines, and the datasets are limited by the time used to collect information as well as the information that is collected. Furthermore, without at least some comprehension of machine learning theory little productivity could be achieved with validity when machine learning is employed.

Related Literature

Comparison the various clustering programs of WEKA tools was a research endeavor performed by Narendra Sharma, Aman Bajpai, and Mr. Ratnesh Litoriya on using WEKA's clustering programs on two extensive collections of data to affirm a particular machine learning program that the most expansive selection of persons with access to WEKA could apply (2012). The sources of the data are the "ISBSG and PROMISE data repositories" (Sharma, Bajpai, and Litoriya, 2012) for the usage of evaluating the clustering programs that are available on WEKA.

The usage of WEKA in Comparison the various clustering programs of WEKA tools project is explained by referencing the "graphical user interface" as well as the relative ease of using WEKA without possessing expert experience in "data mining" (Sharma, Bajpai, and Litoriya, 2012). The clustering programs introduced are discussed in both how the clustering programs work along with the pros and cons of said clustering algorithms (Sharma, Bajpai, and Litoriya, 2012). In affirming the proposal of measuring the most rudimentary machine learning resource in WEKA, k-means clustering was found to be the most rudimentary resource (Sharma, Bajpai, and Litoriya, 2012). This project is similar to my own in the evaluation of multiple machine learning algorithms that are available through WEKA. In committing to this research Sharma, Bajpai, and Litoriya are addressing the obstacle of getting persons of diverse backgrounds to consider machine learning as a valuable resource due to the computer science specific knowledge that is typically required (2012). The machine learning field has accumulated a multitude of various algorithms that can be applied to complete tasks or comprehend the connections that lie within data. There was no standard for the most rudimentary machine learning program for the persons who would be able to utilize WEKA, so the discussion of the paper was to determine such (Sharma, Bajpai, and Litoriya, 2012). Whereas Sharma, Bajpai, and Litoriya's discussion is on a machine learning program that could be used as a starting point (2012), the discussion of this thesis is a machine learning program is most efficient in classification of a specific dataset. Both objectives aim to reduce the mystery that is involved in machine learning through WEKA.

AL-Rawashdeh and Bin Mamat worked with WEKA for classification of spam and non-spam email in Comparison of four email classification algorithms using WEKA (2019). Naïve Bayes Classification Algorithm, Bayes Net Classification Algorithm, J48 Classification Algorithm, LAZY-IBK Classification Algorithm are trained and tested with explanations as to how each

machine learning algorithm is intended to be utilized for the classification of spam email (AL-Rawashdeh and Bin Mamat, 2019). The dataset provided for the endeavor is the SPAM E-mail Database that can be located in the UCI Machine Learning Repository (AL-Rawashdeh and Bin Mamat, 2019). The dataset is defined by 57 features as well as 4601 total emails (AL-Rawashdeh and Bin Mamat, 2019). The effectiveness of the Naïve Bayes Classification Algorithm, Bayes Net Classification Algorithm, J48 Classification Algorithm, and LAZY-IBK Classification Algorithm are determined by taking into account the instances of true positive, false positive, false negative, and true negative possibilities that can be observed in the Confusion Matrix of Table 1 (AL-Rawashdeh and Bin Mamat, 2019). The J48 algorithm is observed in the context of the study to provide the prime overall capabilities for classification of spam and non-spam email of the SPAM E-mail Database (AL-Rawashdeh and Bin Mamat, 2019). Future considerations are made to observe testing the other algorithms that were used in this paper so that their effectiveness can delineated in greater detail (AL-Rawashdeh and Bin Mamat, 2019). AL-Rawashdeh and Bin Mamat's contributions are similar in concept as to what will be done within this Honors thesis. This is to state that there will be machine learning algorithms that will be picked to trained and tested on the multimodal dataset for distracted driving to reveal which machine learning algorithm will yield the superlative model. AL-Rawashdeh and Bin Mamat's provoke thought on concerns of cyber security (2019), which will not be an aim of this thesis.

Educational data mining for student placement prediction using machine learning algorithms was an endeavor using WEKA and R studio to run algorithms in the aim of analyzing educational data on students in the aim of whetting student placement services (Rao, Swapna, and Kumar, 2018). The dataset was the integral training with Random Tree, Random Forest, Naïve Bayes, and J48 from WEKA resources on one portion of the research while binomial logistic

regression, regression tree, neural networks, recursive partitioning, conditional inference tree, and multiple regression was utilized from R studio trained on the dataset (Rao, Swapna, and Kumar, 2018). The models that were the result of the training and testing on the dataset of educational data are observed, and the most effective from WEKA and R Studio are given significant attention (Rao, Swapna, and Kumar, 2018). WEKA will be the only machine learning software package utilized within this thesis. While Rao, Swapna, and Kumar demonstrated that there are various software packages or instruments to access machine learning instruments, the focus of this thesis will be evaluating the machine learning algorithms available in WEKA against the multimodal dataset for distracted driving. The question of which algorithm obtains superb correctness with the provided data further supports the necessity of increased study as to the application of machine learning.

The research endeavor entitled The prediction of Breast Cancer Biopsy Outcomes Using two CAD approaches that Both Emphasize an Intelligible Decision Process is where the dataset used for this research was developed from (Elter, Schulz-Wendtland, & Wittenberg, 2007). The endeavor was based upon 2100 items that were obtained from DDSM (Elter, Schulz-Wendtland, & Wittenberg, 2007). There are 961 instances that are available in the dataset which can be acquired by accessing the data repository that is managed and maintained by the University of California, Irvine (Elter, Schulz-Wendtland, & Wittenberg, 2007). There are six features available, with the class being the severity, severity in the context of this data discerning whether that instance was classified as a “malignant” or “benign” (Elter, Schulz-Wendtland, & Wittenberg, 2007) instance. The aim is that the dataset, when employed in a proper environment for medicine, can allow healthcare professionals to evaluate the capabilities of current hardware as well as software (Elter, Schulz-Wendtland, & Wittenberg, 2007). The time span in which the dataset of

mammogram information was accumulated starts in the year 2003 and ends in the year 2006 (Elter, Schulz-Wendtland, & Wittenberg, 2007). The partition of the 961 instances that benign possesses 515 entries while the partition of instances that are malignant possess 446 entries for future usage (Elter, Schulz-Wendtland, & Wittenberg, 2007).

Comparative Analysis of Classification Algorithms on Different Datasets using WEKA explores the usage of two machine learning programs being applied across multiple datasets to verify which program will be more effective in this scenario (Arora and Suman, 2012). The two machine learning programs in question are the multilayer perceptron and the J48, both of which are accessible through WEKA's toolset (Arora and Suman, 2012). The datasets: vehicle, glass, lymphography, diabetes and balance-scale; present in the research were all obtained courtesy of the University California Irvine Machine Learning Repository (Arora and Suman, 2012). Through the training and testing phases carried out, the multilayer perceptron was more effective in the grand scheme of the project (Arora and Suman, 2012). This research endeavor is similar to the endeavor of this thesis in that the machine learning algorithm that exhibited the most correctness was to be determined from the final results. The question of what machine learning algorithm to employ for a given scenario is once again raised to be examined.

Evaluating the better option from a selection of more than one machine learning program requires a consideration (Bouckaert, 2003). A significant hindrance towards obtaining the most correct or most efficient machine learning program comes from the quantity of data accessible (Bouckaert, 2003). Having any quantity of data or datasets then considering the full breath of machine learning options available to a user in any given situation generates a number of uncertainties (Bouckaert, 2003). Attempting to extract a duo of machine learning options from the

broader expanse for evaluating which is superior only reintroduces those uncertainties on a reduced scale (Bouckaert, 2003).

Choosing between two learning algorithms based on calibrated tests was a research endeavor conducted on three machine learning methods (Alam and Pachauri, 2017). Utilizing the OneR, Naïve Bayes, and J48 machine learning methods for molding a tool for the indications of possible instances of fraud is the topic of Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA (Alam and Pachauri, 2017). The information being used for the machine learning programs was sourced from the Institute for Statistics Hamburg with 1000 entries of German Credit information (Alam and Pachauri, 2017). The entries have been modified to ensure machine learning programs which can only execute with non-categorical features can be utilized information (Alam and Pachauri, 2017). The article proceeds with a section solely focused on the explanation of the machine learning methods used in the research (Alam and Pachauri, 2017). At the conclusion of the research, the J48 machine learning algorithm was evaluated to possess the shortest duration of time to prepare a model as well as possessed the greatest accuracy amongst the suggested machine learning methods to be evaluated (Alam and Pachauri, 2017).

Within supervised machine learning algorithms: classification and comparison, an evaluation of a multitude of machine learning programs are implemented on one dataset are observed and recorded (Osisanwo et al., 2017). The machine learning methods that are the subject of this endeavor is concentrated on supervised machine learning methods, in which there are labels fed to the machine learning program in training to steer the model said program generates to be able to classify new instances that the model was not previously trained on (Osisanwo et al., 2017). The machine learning programs picked to generate models were J48, JRip, the perceptron variation

of neural networks, decision table, support vector machine, Naïve Bayes, and random forest (Osisanwo et al., 2017). A brief delineation of each of the machine learning programs is provided before the statistics of the performance of each machine learning program is given (Osisanwo et al., 2017). The dataset employed was sourced from the online data repository of the University of California, Irvine, and was originally compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (Osisanwo et al., 2017). Multiple tables provide information on various measurements taken on the machine learning models, Table 1 detailing the measurements of machine learning programs when contrasted with one another, and Tables 3 and 4 listing out in-depth statistical measurements of the models on implemented (Osisanwo et al., 2017). The seven machine learning models generated are the deconstructed using the statistical measurements obtained from the implementation of each model (Osisanwo et al., 2017). From those measurements, further generalizations are made by way of the delineation of the machine learning programs that generate the models earlier, and the measurements that have been retrieved for the evaluation of the machine learning models that were generated (Osisanwo et al., 2017). The article presents a generalization of the usage of the traits of each machine learning method for determining afterwards the best scenarios to employ said machine learning methods.

The Comprehensive Analysis of Data Mining Classifiers Using WEKA is an article detailing the progress of Hemlata to delineate the possibilities of advancing data analysis using WEKA as well as attempt to make a go-to reference in regard to how to use the machine learning algorithms accessible within WEKA (2018). Initially, an overview of data analysis is produced which is summarized in a flowchart beneath the initial overview (Hemlata, 2018). Machine learning algorithms are then partitioned into six groups that are each given their respective text, with classification granted the largest portion of text for thorough explanation (Hemlata, 2018).

The machine learning algorithms that are employed upon the Pima_Diabetes, possessing nine features as well as two binary possibilities for the class, dataset that is acquirable through the UCI Machine Learning Repository for the research are: decision table, J48, random tree, Naïve Bayes, Rules ZeroR, Attribute Selected Classifier, Random Tree, SGD Function, Input Mapped Classifier, IBK Lazy, Rules ZeroR (Hemlata, 2018). Each model is subject to 10-fold cross validation as well as applying the model to be tested upon the complete span of the training data (Hemlata, 2018). The SGD Function surpasses the other models in regard to accuracy across most of the features of the Pima_Diabetes dataset (Hemlata, 2018). Therefore, in this scenario, the SGD Function was shown to be the foremost selection to be employed for this research (Hemlata, 2018). Hemlata notes in the conclusion a mention of possible succeeding work in that there are opportunities to continue defining the research Hemlata conducted due to the large quantity of factors to be thoroughly explored (2018). Of these factors, evaluation settings, machine learning algorithms that were not initially utilized within this form of the research as well as other data collections for further delineation of the work of the machine learning field (Hemlata, 2018).

Methodology

Data Construction

The data on mammography exams is constructed with six different features: BI-RADS assessment, age, shape, margin, density and severity (Elter, Schulz-Wendtland, & Wittenberg, 2007). The BI-RADS assessment, and density features are both ordinal in nature, the shape and margin features are nominal, the age feature is an integer, and the severity feature has binary values (Elter et al., 2007). The severity feature is the class for the data, 0 being a representation of a non-malicious growth, 1 being a representation of a malicious growth (Elter et al., 2007). There is a total of 961 entries in the original uncleaned dataset (Elter et al., 2007), but after cleaning the

dataset, a new one was saved with 830 entries. The dataset was obtained secondhand from UCI Machine Learning Repository.

The dataset has dispersion of various data types throughout the entries that encompass the dataset. To enable a wider collection of features that can provide more data per entry for the purpose of providing further data for the machine learning model to use for classification, the data must be cleaned and processed. There were some values that were missing for a few entries, so those 131 entries were removed to allow for a larger breadth of options for which machine learning programs to employ for processing the newly acquired dataset. The remaining entries were preserved in a .arff file to be read and processed by WEKA for future usage. Before running any machine learning features, the class is set as Severity in WEKA so that WEKA will be able to perform the correct calculations with the other features to predict the outcome of said class. Furthermore, there were a few typos that were discovered within the dataset. The way in which said typos were handled was by guessing the correct data input based on the parameters for each feature.

WEKA

WEKA is a program that possesses a multitude of resources for machine learning tasks that can be accessed a multitude of ways, including through the use of a "[graphic] user interface" (Machine Learning Group at the University of Waikato, 2020). WEKA was built to allow users to quickly grasp the necessary procedures so that a broader base could partake in using machine learning. This is partially done by allowing usage of machine learning resources without the prerequisite of the user being familiar with constructing code to implement machine learning algorithms and other constructs (Machine Learning Group at the University of Waikato, 2020). As stated on the website for WEKA, WEKA has served in "teaching, research, and industrial

applications" (Machine Learning Group at the University of Waikato, 2020). WEKA is compatible with other machine learning software including Deepleaning4j, scikit-learn, and R (Machine Learning Group at the University of Waikato, 2020). Additional details on additional, compatible software along with valuable tutorial supplements are available on the WEKA website as well (Machine Learning Group at the University of Waikato, 2020).

In using WEKA, the data is opened using the WEKA GUI which loads the dataset with options that can be selected to modify, organize, set the target feature for classification and other possibilities. The data must first be transferred to a .arff file for WEKA to be able to fully utilize the data. This is simple using file features that come out of the box with WEKA's software.

Machine Learning

Constructing the machine learning models will require two phases. The first phase is the training phase that will essentially center on providing data for the machine learning algorithms to train the models on. After the training phase, a model is generated that can perform classification to some degree of accuracy. The next phase is to test the model on data that the model has not been exposed to before for the purpose of ensuring the model is classifying data correctly. In the event that a model is seen to perform horribly, the model can be retrained again under the same parameters then retested to observe if there is an improvement in the results. During the training phase, a number of attributes can be altered to enhance the machine learning model if necessary, to allow said model to learn effectively. This does not change the overall concept of the model, but can lead to variations of the implemented structure and augmentation of parts to heighten or hamper the correctness of the model.

Cross validation will partition the dataset over iterations, with each iteration specifying the portion of data to be partitioned for the purpose of testing with remainder used to train the model

for that iteration. This allows the model to train and test on the entire dataset. The full extent of the results of said training as well as testing will be demonstrated in the correct and incorrect results yielded in the WEKA GUI. There are a number of mathematical measures that can provide deeper insights as to the errors that are made in the machine learning model's classification of the test data.

Building the Models

Once a dataset is properly cleaned, the dataset can be uploaded to the WEKA program. The user may then click on *classify* to gain access to the classification algorithms that are available. Once a classification algorithm is selected, the algorithm will begin to design a model based on the inputted data with the aim of making the model as accurate as possible when classifying the data. Before constructing model, the user has access to various parameters in regard to how the model will be trained, tested, and what information will be available to assist in the evaluation of the model that will be constructed. After that the user waits until the model is constructed, the model processes the data, and then various, customizable statistics are portrayed in the WEKA window in regard to the model's performance. Further options such as visualizations in regard to the model may be accessed as well.

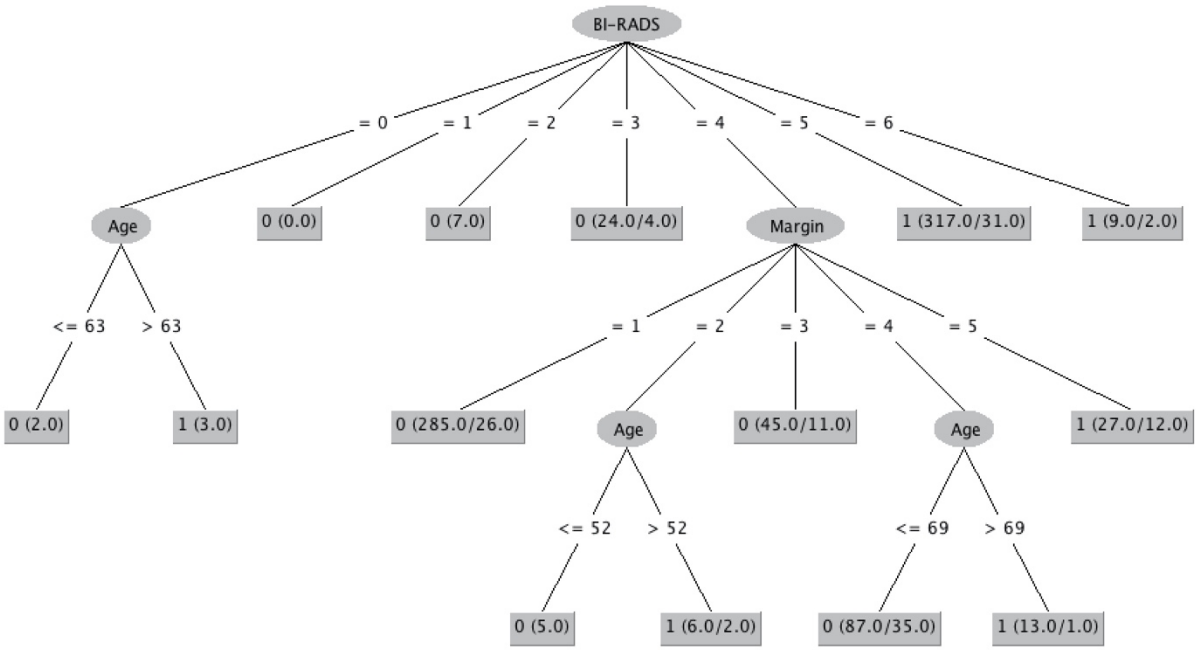
The J48 decision tree was ran using 10-fold cross validation and achieved an accuracy of 81.57% when rounded to the nearest hundredth. The mean absolute error is 0.2444 and root mean squared error is 0.364. The mean absolute error and root mean squared error are both small which is great in being indicative of error being minimized in the model. The J48 decision tree is a constructed from the concept of the Iterative Dichotomiser 3 where information gain is critical to the effectiveness of the J48 model (Girones, 2020). Information gain refers to the value of details that are present within information, which in turn allows the decision trees created from the J48

algorithm to place a higher emphasis on specific features (Girones, 2020) for the purpose of maximizing the accuracy achieved. A better visual of the accuracy of the J48's produced model can be observed in the confusion matrix constructed display the errors made visually:

Confusion Matrix for the J48 Model		
benign	malignant	Classified As
353	74	benign
79	324	malignant

There were 353 instances of a benign mass were correctly classified by the J48 decision tree as well as 324 instances of a malignant mass were correctly classified by the J48 decision tree. 74 instances were misclassified as benign masses but were actually malignant and 79 instances were misclassified as benign but were actually malignant.

Below is a visualization of the J48 decision tree model that was acquired via the visualization features that accessible through the WEKA GUI. In it, the breakdown of the analysis that the J48 model devised is apparent. This can be used as further reference as to what processes were occurring within the J48 decision tree, and simpler to follow for most persons versus the code outputs that are associated with a decision tree.



J48 Decision Tree Visualization 1:This is a graphical representation of the J48 decision tree model that was generated from the data. Within it, you can see the calculation process that was made for each input for determining the classification within the severity class.

The Naïve Bayes algorithm was utilized to construct a model using the mammogram dataset as well. There were 82.89%, when rounded to the nearest hundredth, correctly classified instances. The mean absolute error is 0.1839 and the root mean squared error is 0.3654. 688 instances from the dataset were classified correctly and 142 instances from the dataset were incorrectly classified by the Naïve Bayes algorithm. The Naïve Bayes algorithm uses graphs that maintain a parent to child connection for the purpose of constructing models (Osisanwo et al., 2017). A visual detailing of the performance of the Naïve Bayes model can be observed within the confusion matrix for the Naïve Bayes model:

Confusion Matrix for the Naïve Bayes Model		
benign	malignant	Classified As
343	84	benign
58	345	malignant

It can be observed that 343 instances were correctly classified as being benign, but 84 instances were misclassified as being malignant when said 84 instances were actually benign. There were 345 instances correctly classified as being malignant, but 58 instances were misclassified as being benign when said 58 instances were actually malignant.

Multilayer perceptron is WEKA's variation of the algorithm for spawning artificial neural networks. An artificial neural network consists of an input layer which has a constitution of input nodes where data is initially received, stored and processed. From the input layer, the inputted information is transported to the hidden land calculated with weights. There can be multiple hidden layers to add further transfers and computation in hopes of reducing error within the final output. The output layer is when the final calculations are made, and the final output is retrieved to determine the classification of each input. The multilayer perceptron has correctly classified 80.60%, rounded to the nearest hundredth, of the 830 provided instances. The mean absolute error is 0.2268 and the root mean squared error is 0.374. A visual of the details of the correct and incorrect classification can be observed in the confusion matrix below:

Confusion Matrix for the Naïve Bayes Model		
benign	malignant	Classified As
348	79	benign
82	321	malignant

There are 348 instances that were correctly classified as being benign, and 79 instances that misclassified as being malignant despite actually being instances of benign masses. There are 321 that were correctly classified as being malignant, yet there are 82 instances that were misclassified as being benign when said 82 instances were actually malignant.

The fourth algorithm that is utilized is WEKA's adaptation of logistic regression. Logistic regression basically works with a binary classification, however WEKA's adaptation uses a variation of the logistic regression that can be taken beyond binary outputs. Due to the severity class only having two ends, either malignant or benign and nothing else, logistic regression will be working towards either one of those two outputs. The percentage of classified instances achieved by the logistic regression model is 83.494%, while the mean absolute error is 0.2319 and the root mean squared error is 0.3483. Below, a confusion matrix will visually describe where the logistic regression model's errors lie:

Confusion Matrix for the Naïve Bayes Model		
benign	malignant	Classified As
366	61	benign
76	327	malignant

There were 366 instances that were correctly classified as being benign, but 61 instances were malignant masses that were misclassified as being benign. There were 327 instances that were classified as being malignant, yet 76 instances that were benign masses were misclassified as being malignant.

Metrics for Models			
Model	Accuracy	Mean Absolute Error	Root Mean Squared Error
J48 Decision Tree	81.5663%	0.2444	0.364
Naïve Bayes	82.8916%	0.1839	0.3654
Multilayer Perceptron	80.6024%	0.2268	0.374
Logistic Regression	83.494%	0.2319	0.3483

The mean absolute error of the J48 decision tree is lower than that of the multilayer perceptron with $0.2444 < 0.2268$, plus the accuracy of the J48 decision tree is higher than that of the multiplayer perceptron with $81.57\% > 80.60\%$. The mean absolute error for the Naïve Bayes models is noticeably lower than that of the J48 decision tree. Furthermore, the accuracy of the Naïve Bayes Model was slightly greater than that of the J48 model, as seen in the comparison $82.89\% > 81.57\%$. From looking at the accuracy alone, we can see that the Naïve Bayes model surpassed the J48 decision tree in this scenario. Such evaluation immediately reveals that the Naïve Bayes model has also surpassed the multilayer perceptron within the confines of this scenario as well. The logistic regression model surpassed all the previously mention models in correctly classified instances from the mammogram dataset.

Possible reasons for the results of this scenario is that logistic regression model, was in its optimum environment with only two outcomes to be concerned. This could have played an advantage over the other models that were constructed. Furthermore, there were only five features

to use for the purpose of attaining the correct outcome for the class. This would leave more sophisticated algorithms at a disadvantage when attempting to build accurate models due to the lack of details that are available for each entry of the dataset.

Discussion

The results acquired from the testing evaluation are can be observed in the Metrics for Models. The question that emerges is as to why the logistic regression model outperformed the other models that were generated by the machine learning algorithms present. First, examining the surface details of the data itself, it is immediately apparent that the partitioned dataset is small, containing 830 instances. Generally, machine learning algorithms will become better with the increased amount of data that is available. The partitioned dataset used for training the machine learning models then testing said models is tiny compared to more expansive datasets of thousands or millions of instances. This would impair the models that would be more effective when the dataset has an enormous quantity of instances to process. Also, there were five features and one class. If more features were available there would be a possibility that more sophisticated machine learning algorithms could take advantage of more information due to the increased number of features. Such, however, was not the case with only five features to determine the value of the class for each instance. The fact that there were only 830 instances employed for training and testing further compounded the capabilities of the models from the constraints of the quantity of available data. If the data had information completed for each entry, then the full dataset might have been applicable. Furthermore, errors in the entries of the data, if fixed by the original managers of the data, would have saved time as well as errors from initially building models with errors in the data.

The mean absolute error and the root mean squared error are both mathematical evaluations of the dispersal of errors in classifying instances of the dataset. These mathematical values grant a more defined observation of the errors that have occurred within classification of the partitioned dataset. Mean absolute error is typically applied for the examination of errors in data that possess characteristics of a uniform distribution (Chai and Draxler, 2014). The root mean squared error makes calculations for errors that maintain characteristics of a normal distribution lacking any misevaluations in predicted and actual values (Chai and Draxler, 2014). For both metrics, the aim is that the value will be as tiny as possible. The Naïve Bayes model has the lowest mean absolute error value of 0.1839 and has the second highest accuracy of the models generated and tested. The logistic regression model can be observed to hold the lowest root mean squared error value of 0.3483. Although the metrics are not perfect at predicting which models will be correctly classify the most instances from the provided partitioned mammogram dataset, they can be good indicators of the top performing machine learning algorithms. Furthermore, both metrics act as superb delineations of the four errors within the models.

Given that there were only four machine learning algorithms that were utilized to generate models, this is not conclusive in saying that the most effective machine learning algorithm in regard to classification for the partitioned dataset is the logistic regression algorithm. There are far more machine learning algorithms designed for classification that can be applied to the partitioned dataset beyond the four that were used in this endeavor. Only upon testing each applicable algorithm under the same parameters for training and testing can an evaluation be made for which machine learning algorithm performs best under those given parameters, not overall. Testing for the overall best machine learning algorithm would involve altering the parameters of training and

testing and configurations for the models to be generated to ensure that the model chosen is one that was able to provide the best results amongst numerous shifting factors.

Utilization of the WEKA software package and the WEKA GUI has placed a granted the machine learning resources without stressing the necessities of expertise in coding to conceptualize then implement. The WEKA GUI has removed most of the usage of coding, except for possibly cleaning the data, although there are features in WEKA that provide such functionality without exiting the WEKA GUI. Generation of .arff files from file formats such as .csv can be performed using built-in functionalities of WEKA. The main limitation to coding in this respect is that most of WEKA's applications are confined to as well as buttressed by .arff files. Once a .arff file is acquired, widgets can be used to select the machine learning algorithms to generate the models the parameters for the training, the parameters for testing, then the initiating the building of the model. After the model runs, said model can be saved for future usage through mouse actions versus properly preparing lines of code to specify a file and location to save the model beforehand.

The assortment of machine learning algorithms available in WEKA is extensive. A multitude of machine learning algorithms for classification are sorted amongst groupings that are constructed from similar characteristics. There are also machine learning algorithms for clustering, as well as association. Then there are algorithms for evaluating the significance of features of a dataset that can determine which features have the most influence in predicting the class. Further options can be added through add-ons or potentially coding new functionality to the WEKA software package. Concerning the built-in machine learning algorithms, there is an observed restriction of which machine learning algorithms will be implemented through the characteristics of the data that is loaded into WEKA. This is helpful in preventing misuse of machine learning

algorithms on data that would not be applicable for the implementation of said machine learning algorithm.

Concerning the validity of the results of the machine learning models that are observed within the context of this endeavor, similar results may be reproducible. There is no guarantee that the exact same model will be generated with the same data nor is there any guarantee of the same outputs. Once similar or differing results are obtained, those results can be used to validate or examine the context that allowed for the generation of differing models. Furthermore, the training and testing process of machine learning allow for the machine learning models to be validated within the process. This validation ensures that the model is producing reasonable results. Not validating the model generated is notoriously bad practice as not testing the validity of the model leaves room for improperly generated models or outputs to be made.

Future Work

In the future there would be an expansion in the assortment of machine learning algorithms that would be employed generating a larger a collection of models. This would allow for further testing to determine the classification accuracy of various models to determine which possesses the best results for the provided partitioned mammogram dataset. There could be additional metrics added to verify the machine learning model that minimizes classification errors as well. On another level, there can be tweaks to the configurations of the machine learning models to enable examination as to what configurations allow for better models to be outputted for the partitioned mammogram dataset. Increased selection of models and variations of those models' configurations will yield more results in regard to which model will provide the best set of results overall.

Considerations can be made for the data in future research endeavors. Although the mammogram dataset did allow the generation of some successful models, locating a dataset with more instances, more features, or restricting the instances or features can enable study of machine learning model development with more data. This should lead to each machine learning model benefitting from the greater or smaller breadth of data available. The machine learning algorithm that outperforms the remainder could vary as well. More data or minimized data in the case of instances or features for each respective instance could backfire as well. There could be cases where one feature gains far too much significance within the model or that the increased amount of data hinders the model's performance such as underfitting or overfitting. Encountering such possibilities will contribute more to the continued research of machine learning phenomena.

Conclusion

There are considerable capacities for machine learning. In the applications of four machine learning algorithms in this endeavor, four models were generated. The logistic regression model classified the most instances of the provided partitioned mammogram dataset in this scenario. Said event is not to be taken as logistic regression in the superb machine learning algorithm in this scenario, but a reaffirmation as to how machine learning algorithms do not have a superior choice present initially without extensive amounts of testing to determine such for a specific scenario. Further evaluations will have to be completed to determine the superb choice for the partitioned mammogram dataset.

References

Arora, R., & Suman, S. (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications*, *54*(13), 21-25. doi: 10.5120/8626-2492

Alam, F., & Pachauri, S. (2017). Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA. *Advances in Computational Science and Technology, 10(6),* 1731-1743.

AL-Rawashdeh, G. H., & Mamat, R. B. (2019). Comparison of four email classification algorithms using WEKA. *International Journal of Computer Science and Information Security (IJCSIS)*, *17*(2).

Ayodele, T. O. (2010). Types of machine learning algorithms. In *New advances in machine learning*. IntechOpen.

Bouckaert, R. R. (2003, August). Choosing between two learning algorithms based on calibrated tests. In *ICML* (Vol. 3, pp. 51-58).

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Dcosta, M. (2017). Simulator Study I: A Multimodal Dataset for Various Forms of Distracted Driving.

Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical physics*, *34*(11), 4164-4172.

Girones, J. (2020). *J48 decision tree - Mining at UOC*. Data-mining.business-intelligence.uoc.edu. Retrieved 28 March 2020, from http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree.

Hemlata. (2018). COMPREHENSIVE ANALYSIS OF DATA MINING CLASSIFIERS USING WEKA. *International Journal of Advanced Research in Computer Science*, *9*(2), 718-723. https://doi.org/10.26483/ijarcs.v9i2.5900.

Machine Learning Group at the University of Waikato. (2020). WEKA 3 - Data Mining with Open Source Machine Learning Software in Java. Retrieved 19 January 2020, from https://www.cs.waikato.ac.nz/ml/WEKA/

McCarthy, J., & Feigenbaum, E. (2020). Arthur Samuel. Retrieved 30 January 2020, from http://infolab.stanford.edu/pub/voy/museum/samuel.html

Mooney, R. *CS 391L: Machine Learning: Computational Learning Theory* [Ebook] (1st ed., pp. 1-7). Austin: University of Texas at Austin. Retrieved from https://www.cs.utexas.edu/~mooney/cs391L/slides/colt.pdf

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.

Pavlidis, I., Dcosta, M., Taamneh, S., Manser, M., Ferris, T., Wunderlich, R., ... & Tsiamyrtzis, P. (2016). Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors. *Scientific reports*, *6*, 25651.

Rao, K. S., Swapna, N., & Kumar, P. P. (2018). Educational data mining for student placement prediction using machine learning algorithms. Int. J. Eng. Technol. Sci., 7(1.2), 43-46.

Sharma, N., Bajpai, A., & Litoriya, M. R. (2012). Comparison the various clustering algorithms of WEKA tools. *facilities*, *4*(7), 78-80.