Optimizing Data Mining Techniques to Improve Breast Cancer Predictions

Authors: Erin Berg, Dylan Nadeau, Kyle Tan

Abstract

The purpose of this paper is to discuss and analyze how accurately we can predict whether a patient has breast cancer based on attributes that any woman can obtain from a routine checkup. Our aim is to identify the optimal conditions for several classification tasks and enhance the one that produced the best results. Classification tasks that were implemented included Artificial Neural Networks, Support Vector Machines, and Logistic Regression. We faced two major challenges: incorporating with a relatively high number of attributes (10) and working with a small data set (116 instances). These problems give rise to high complexity and possible overfitting. To combat these obstacles, we implemented techniques such as feature selection and k-fold cross-validation. In the end, we discovered the ideal attributes and parameters to perform logistic regression and obtain a prediction accuracy of 77.59%. Doing so, however, we stripped our data to the bare minimum and were not able to improve our prediction accuracy with other adjustments.

Introduction

Cancer is the second leading cause of death in the United States, taking the lives of almost 600,000 people each year and impacting countless others, according to the Centers of Disease Control and Prevention.[1] Although cancer manifests itself in many ways, breast cancer is among the deadliest. About one in eight women will be diagnosed with the disease at some point during their lives, meaning it is the most common malignancy among that demographic.[2]

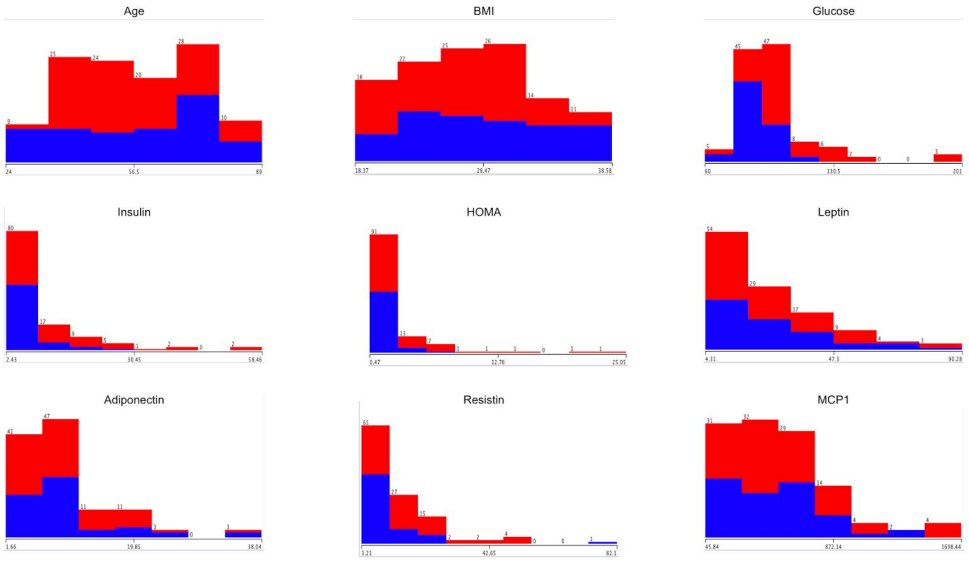
Finding and treating cancer at an early stage can save lives. Studies show that more than 90% of women diagnosed with breast cancer at the earliest stage live with their disease for at least five years compared to 15% for women diagnosed with the most advanced stage of the disease. Early detection through regular screenings with mammography have resulted in 30% fewer deaths.[3]

If the presence of breast cancer can be predicted based chemical levels and characteristics of patients taken during routine blood work by using models over a dataset, then presence of breast cancer in new patients could be determined earlier, thus increasing the likelihood of a successful treatment and longer life. The Breast Cancer Coimbra Dataset taken from the UCI

Learning Repository was used for this study because it includes nine attributes, all which are characteristics and chemical levels that would be taken during routine blood work of any patient in a doctor's office. The instances in this set are Portuguese women, of whom 64 were newly diagnosed with breast cancer and 52 are healthy controls. This dataset only included women with a BMI below 40 kg/m^2 because a woman with a BMI higher than this threshold is considered obese.

It has been proven that obesity is linked to breast cancer, so removing those instances will give a more accurate prediction for healthy individuals. The nine attributes of this dataset that were used to predict the presence of breast cancer were age, body mass index (BMI), glucose level, insulin level, HOMA level, leptin level, adiponectin level, resistin level, and MCP.1 level. These attributes are intertwined with each other and many can be related to breast cancer. For example, glucose is a simple sugar that is an important energy source in living organisms and is a component of many carbohydrates. High glucose levels promote cell growth, including the growth of cancer cells.[4] Insulin is a protein that regulates the amount of glucose in the bloodstream and promotes cell growth.

Obesity is correlated to higher levels of insulin because people diagnosed with obesity are likely to be insulin-resistant so their bodies must produce extraordinary amounts of insulin. Insulin resistance puts patients at an increased risk for breast cancer.[5][6] Resistin is a cysteine-rich adipose-derived peptide hormone that modulates steps in the insulin-signaling pathway and induces insulin resistance.[7] Thus, it is also correlated with obesity and an increased risk for breast cancer. Leptin is a protein produced by fatty tissue and believed to regulate fat storage in the body. Leptin levels increase as the result of higher fat concentrations (high BMI). BMI is a measure of a person's weight relative to their height.[8] Adiponectin is a protein hormone involved in regulating glucose levels as well as fatty acid breakdown. Lower levels of adiponectin indicate a higher likelihood of breast cancer and are associated with obesity (high BMI).[9] Homeostatic Model Assessment (HOMA) is a measurement of insulin sensitivity. MCP.1 is a cytokine that reduces the intake of insulin-stimulated glucose in muscle cells.[8] Below are graphs demonstrating the breakdown of each of the nine attributes by class. The healthy controls are represented in blue and the patients with breast cancer are represented in red.



The graphs demonstrate that glucose, insulin, HOMA, and resistin have the strongest correlation with breast cancer. Higher levels of each of these seem to indicate a higher susceptibility to breast cancer. On the other hand, age and MCP.1 can be expected to be less helpful predictors because the classes are split more evenly across the different attribute values. This data gives an idea of the features that could be removed during feature extraction. A challenge that must be confronted in this study is overfitting that is difficult to avoid due to the small sample size of the Breast Cancer Coimbra Data Set as it may lead to artificially high accuracy results.

Methods

Given the characteristics of the Breast Cancer Coimbra Dataset, we hypothesized that Artificial Neural Networks, Support Vector Machines, and Logistic Regression would generate the highest performance accuracies. In order to observe how well these classification tasks functioned, we then conducted a 10-fold cross-validation test on the data using various classifiers with default parameters. WEKA and Matlab were implemented throughout our entire project. Below is a table of our results:

Classification Algorithm	Accuracy
IBK	66.38%
Logistic Regression	**73.27%**
Multilayer Perceptron	**65.5%**
One R	65.6%
SMO	**66.4%**

Testing on an untouched data set using default parameters, we anticipated low accuracies. We decided on a baseline accuracy of 50% to show that these classifiers perform better than simple chance. As observed from the table above, all of the chosen classifiers yielded accuracy results ranging from about 65% to 75%, providing relatively similar results when compared against one another. These outcomes allowed us to have a general idea of the performance of the classifiers, while also providing a goal to improve upon.

The first classifier we enhanced to obtain a higher accuracy rate was Artificial Neural Networks. Artificial Neural Networks (ANN) mimics the functionality of the human brain in order to complex and nonlinear decision boundaries to classify the data. A simple model, known as a perceptron, comprised of an input node, representing the input attributes, and an output node, representing the output of a model. Connecting each node is a link that assigns a weight learned from the data. A multilayer network, used for more sophisticated classification, builds off of a perceptron and additionally contains hidden layers that allow for complex decision boundaries to be improved. Within each hidden layer, a predetermined number of hidden nodes process the data and transmits the results to the following layer. Each weight, along with a bias, is iteratively updated using gradient descent in order to produce a better output. The simple algorithm for updating the weights and biases follows:

$$w_{ij}^{(k+1)} = w_{ij}^{k} + \lambda \frac{\partial E}{\partial w_{ij}}$$

$$b_{i}^{(k+1)} = b_{i}^{k} + \lambda \frac{\partial E}{\partial b_{i}}$$

Where w represents the weight, b represents the bias, and λ represents the learning rate.

This classifier was chosen due to its unique characteristics for classifying data. ANNs can easily learn, even if presented with less relevant or redundant attributes, as weights are automatically learned. Previously mentioned, our data set contains main attributes that are highly correlated, such as resistin level, glucose level, and HOMA value, to name a few. ANNs assign similar weights to these redundant attributes so that one does not influence the quality of the classifier more than the others. Additionally, ANN was chosen due to its ability to classify nonlinear data. The networks are highly expressive and we presumed its capability to create nonlinear decision boundaries would help in producing a high classification accuracy for our complex data set.

To optimize ANNs, we needed to determine the ideal parameters for the number of hidden layers and hidden nodes. Because every data set contains different attributes, instances, and values, there is no specific formula in determining how many layers and nodes to incorporate. One may assume that a high number of layers and nodes will produce the best results, but that leads to a strong chance of overfitting the data. On the other hand, a small number of layers and nodes leads to underfitting. As a general rule of thumb implemented by data scientists, only a small number of layers are necessary for performing just about any task as long as it is not overly complex.[10] Additionally, the number of hidden nodes should not exceed the maximum number of attributes.[11] Using this logic, we decided to construct the ideal ANN for our data set by testing every possible combination of hidden layers, ranging from 1 to 3, and hidden nodes, ranging from 1 to 9.

Another classifier we optimized was Support Vector Machines. Support Vector Machines, SVMs, can create nonlinear decision boundaries that maximize the distance between the two classes. Doing so, the chance of misclassification is reduced.To determine the decision boundary, SVMs only use a subset of instances that are difficult to classify, known as support vectors. These support vectors are located close to the determined boundary.

SVMs proved to be promising due to its unique method in computing nonlinear decision boundaries. Nonlinear decision boundaries are computed by transform the data into high dimensional space but computes results in terms of the coordinates in the original space. This

method, known as the Kernel Trick, allows for simpler calculations and visuals when dealing with higher dimensionality data. Dealing with the curse of dimensionality, this feature of SVMs proved to be favorable. The formula for computing the Kernel Trick follows:

$$K(u, v)=\langle\varphi(u),\varphi(v)\rangle=f(u, v)$$

Where u and v are the instances and φ is the nonlinear transformation of the data. It is advantageous because the mapping of φ does not need to be known and the curse of dimensionality is avoided computing the dot product of φ(u) and φ(v) in the original attribute space.

In order to optimize SVMs, we calculated the polynomial Kernel function for various dimensions:

$$K(u,v)=(u^T v+1)^p$$

Varying the polynomial p transforms and computes the data in different dimensions, each giving a different accuracy.

The last classifier improved was Logistic Regression. This classification approach is based on solving a regression subtask that estimates the following odds ratio:

$$P(y = \text{class 1} \mid \text{attributes x}) / P(y = \text{class 2} \mid \text{attributes x})$$

The odds ratio can also be estimated as:

$$\delta((w^T \bullet x) + b)$$

Logistic regression solves this classification task by casting it as an optimization task by minimizing a loss function to find ideal weights for w and b. This loss function also contains a ridge penalty that fights multicollinearity. Multicollinearity is the existence of near-linear relationships among the independent variables. It is a problem because it can create inaccurate estimates of the regression coefficients, which are susceptible to very high variance. In short, the ridge penalty prevents effects of multicollinearity by controlling coefficients of each attribute.

Like ANNs and SVms, logistic regression can handle redundant attributes because weights are automatically learned. Additionally, this classification task works well with high-dimensional data sets. In order to optimize this classifier, we sought the ideal ridge penalty that yielded the best accuracy performance without heavily influencing the data.

To further improve our results, other techniques must be implemented. One of the most common methods to enhance machine learning models is feature selection. Also known as attribute selection, this technique is used to eliminate features used in constructing the model. This is different from feature extraction, despite their similarities. Feature extraction seeks to find the best combination of the original attributes. A common type of feature extraction that we considered using was principal components analysis, or PCA. In PCA, the original features are transformed into a new equally sized or smaller set of linearly uncorrelated attributes. The algorithm seeks to find attributes with the highest amount of variance between the classes while maintaining the ability to reconstruct the attributes.[12] A downside of PCA is that the attributes need to be scaled in a particular way for the best results, so feature selection seemed to work more naturally with this dataset.

Although reducing the number of attributes may seem counterintuitive, it must be noted that the removed features are either redundant or irrelevant. This will leave the model with only the important features to use, translating to minimal information loss. Feature selection serves several important roles. Not only does it simplify the model, decreasing training time and making the model easier to interpret, but it also increases generalization by reducing overfitting.[12] By using fewer features, the model is overall simpler and easier to use.

The breast cancer dataset is small in size, so few feature selection methods would actually improve performance. The first technique tested was the correlation based feature selection. As the name implies, features that are highly correlated to the class are given a higher priority than features with little to no correlation. Given the interconnected nature of our dataset, this type of selection fits very nicely with our model. Attributes such as glucose, insulin and resistin are both incredibly related to each other and have a direct impact on whether a person has cancer or not. It follows that giving higher preference to these and ignoring other, less impactful data points may provide the model with an easier set of data to predict from, leading to a higher performance accuracy.

The correlation between the attributes and the class can be seen from the previous graphs that has each attribute plotted against the class. Attributes that are cleanly split between positive and negative cancer results would be more strongly correlated than those that are scattered

among the graph. Initial indications showed that glucose, insulin and resistin would be very highly correlated with breast cancer, while BMI and age would not be. Using WEKA, we are able to use a statistical significance test to determine the ranked attribute list detailing how connected each feature is to the class.

In addition to correlation based feature selection, learner based feature selection proved to be a promising method for the dataset. In much of the current machine learning literature, learner based feature selection is praised as one of the better attribute selection techniques to use. [16] The learner based feature selection takes many subsets of the attributes and tests each one against one base classifier. The classifier does not have to be the one used in our model; rather, it should be quick and easy to train.[16] When the data was feed through the algorithm, a J48 decision tree was used to test each subset of features. J48 decision trees are among the simplest and quickest classifiers to use, so it was appropriate to use for testing.

After each subset is put through the J48 classifier, the algorithm compares each classification accuracy against the others. Whichever subset has the best performance is declared the most relevant subset. The intuition behind why this method should be successful lies in the fact that the features in the most successful subset were tested against every single other subset, so there can't possibly be a better way to rearrange the attributes.

While optimizing our model was a primary concern, seeking simpler, more established models that consistently perform well was a move that could produce better results than what our model can do. To improve the quality of results of one classifier, it follows that using many classifiers may increase the accuracy of the model. This basis forms the foundation of meta-learners, or ensemble classifiers. These classifiers use multiple algorithms to come to different results. Those results are usually interpreted via a weighted average, and that decision is the prediction for that set of data. We used a specific type of ensemble learner called boosting. Boosting is known for its ability to minimize the effects of bias, variability and noise.[17] In order to do that, it uses an iterative approach to test the data. Using a base classifier, the algorithm runs the first iteration of testing data through. It identifies misclassified records, and puts emphasis on correcting those mistakes. By weighting misclassified errors more heavily, it focuses on fixing the issues and spends less time training for something it already knows.

Classifiers we decided that would not help us achieve our goal as well were clustering, decision trees, and K-nearest neighbor. Clustering was not promising because are attributes are greatly connected, so the model would not clearly define accurate subsets. This same rationale goes for decision trees as well, as they do not account for interactions between attributes. K-nearest neighbor was not used either due to our high number of attributes and difficulty in mapping.

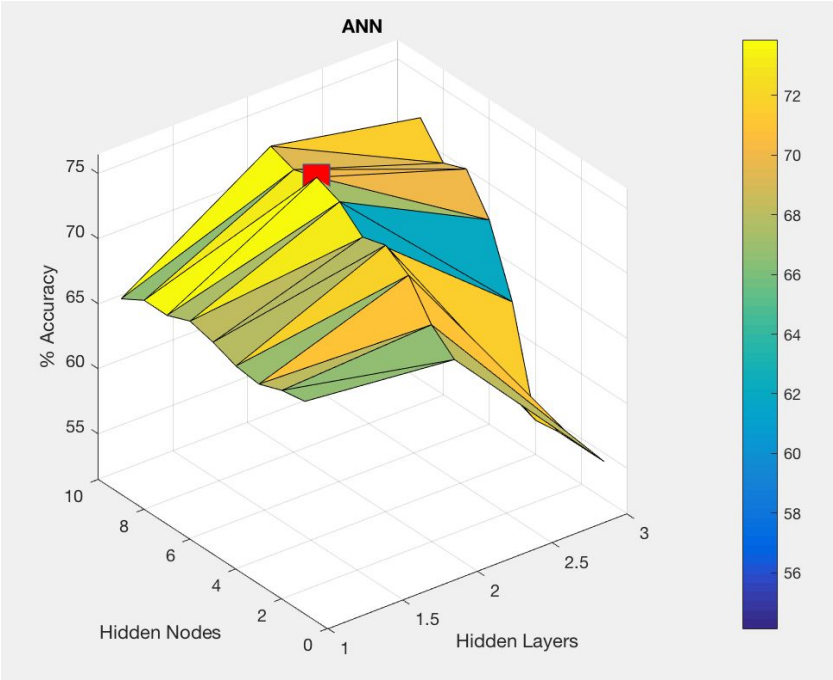
Our data set was relatively small, with only 116 instances. For this reason, overfitting was notion we hoped to avoid through a 10-fold cross-validation for every experimental run. Unfortunately our data set was not large enough to be effectively divided into training, test, and validation sets. Due it its size, however, it made our runtime fairly quick. It is important to note, however, that our observations were not statistically significant at the 0.05 level. Increasing it may have made our results significant, but more prone to error.

Results

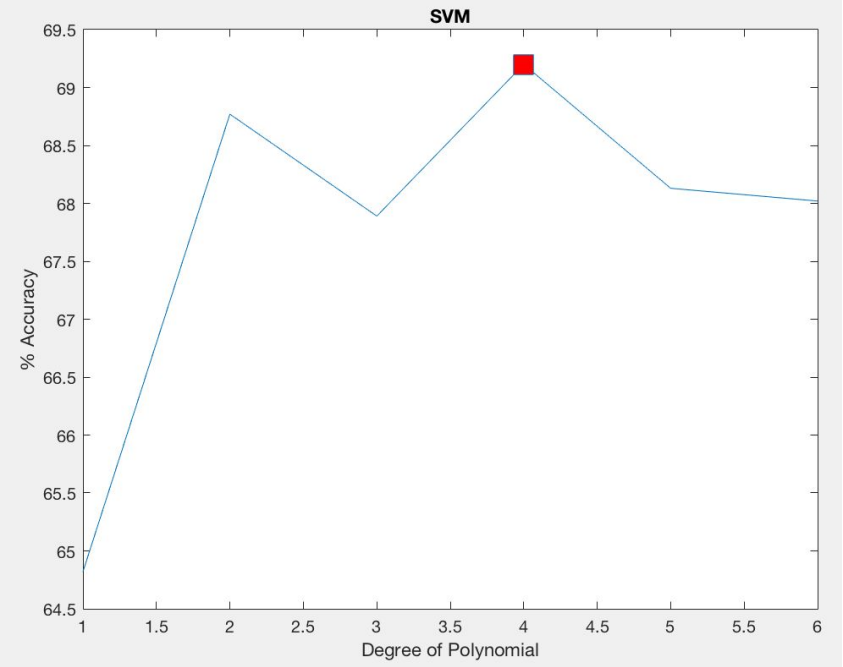
Optimizing the dataset using the techniques described in our methods, we determined the ideal parameters for each classification task and found that we improved our results.

Classifier	Optimization	Previous Accuracy	New Accuracy
ANN	2 hidden layers, 7 hidden nodes	65.50%	71.97%
SVM	Kernel polynomial of degree 4	66.40%	69.20%
Logistic Regression	Ridge penalty of 0.4	73.27%	74.84%

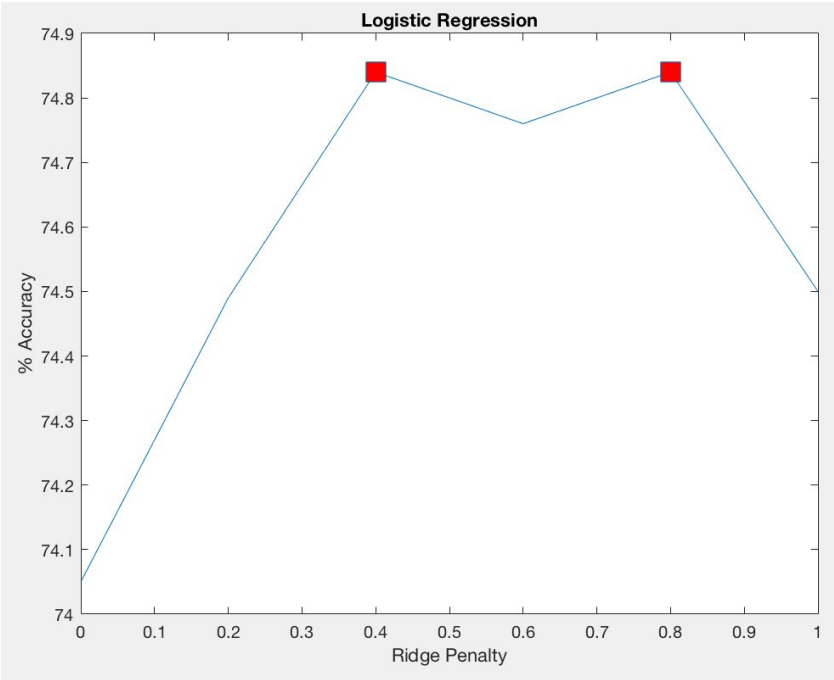
Below are visualizations of our results:


ANN
% Accuracy
75
70
65
60
55
10
8
6
4
2
0
Hidden Nodes
1
1.5
2
2.5
3
Hidden Layers
72
70
68
66
64
62
60
58
56

ANN


SVM
% Accuracy
69.5
69
68.5
68
67.5
67
66.5
66
65.5
65
64.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
Degree of Polynomial

SVM



Logistic Regression

For logistic regression, it is important to note that a ridge penalty of 0.4 and 0.8 produced the same accuracy of 74.84%. We chose the smaller penalty of 0.4 in order to avoid influencing the data as much as possible.

As demonstrated, each of our optimizations increased the performance accuracy by an average of 3.5 percentage points. Our most successful model was the logistic regression, with an accuracy of 74.84%

Moving forward, enhancing our best model was the best opportunity to predict breast cancer occurrences at the most successful rate. Since the base logistic regression model had the highest classification accuracy, incorporating feature selection was the next natural move.

Starting with the correlation based subset of attributes, we processed the dataset to determine correlation strength. We found that the following ranked attribute list was created.

```
Attribute Evaluator (supervised, Class (nominal): 10 Classification):
        Correlation Ranking Filter
Ranked attributes:
 0.38432   3 Glucose
 0.28401   5 HOMA
 0.2768    4 Insulin
 0.22731   8 Resistin
 0.13259   2 BMI
 0.09138   9 MCP.1
 0.04355   1 Age
 0.01949   7 Adiponectin
 0.00108   6 Leptin

Selected attributes: 3,5,4,8,2,9,1,7,6 : 9
```

Although no individual attribute was extremely correlated with breast cancer, some attributes were statistically more significant than other attributes. Glucose, HOMA and insulin were the three most associated features, while leptin, adiponectin and age were the three least related to whether or not a patient has breast cancer. While insulin and HOMA are directly tied to glucose, and glucose is closely correlated with cancer rates, this does not imply that high glucose levels directly cause cancer; that claim is outside the scope of our research, though we do find a clear association between them.

To determine the appropriate threshold to use as a cut-off point, we decided to test every subset of features. Each test would have one less attribute, namely the lowest correlated feature at that time. After testing each of the eight subsets, our team found that the best threshold was a correlation that was between 0.04 and 0.02. This meant that eliminating exactly two attributes, leptin and adiponectin, yielded the highest accuracy of 77.58%. This increase in accuracy was significantly higher than the previous improvements we've made to the classifier. It also meant that the correlation based feature selection was successful in enhancing our model, as the previous best model produced a classification accuracy of 74.84%.

Knowing feature selection can make a positive impact on our model, we then tried to analyze the effects of learner based feature selection. Using WEKA to calculate the best subset of attributes, the following list was produced: BMI, glucose leptin and resistin.

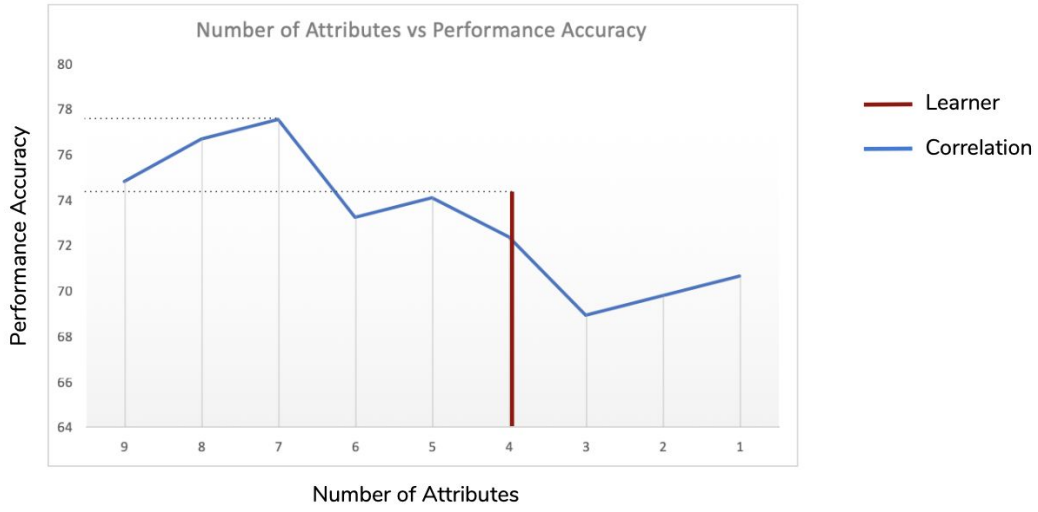
```
Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.752

Attribute Subset Evaluator (supervised, Class (nominal): 10 Classification):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 2,3,6,8 : 4
                     BMI
                     Glucose
                     Leptin
                     Resistin
```

The two feature selection methods has both similarities and differences. They both chose glucose as an important attribute. They difference most drastically on the significant of leptin; while the correlation based model did not see the value in the feature, the learner method found the best results using it for training. The models also differed slightly on how determining the value of BMI and resistin, but to less of an extent.

The results aligned with our initial expectations. Using the new list of selected attributes, the learner based model yielded an accuracy of 74.14%, 0.7% worse than the base model. Although we were correct in our prediction that the correlated nature of the attributes would lead to correlation based feature selection being better, we did not anticipate the learner based model to perform worse than our optimized logistic regression model. We hypothesize that the learner model stripped the dataset of too much valuable information. The list of four features was likely overfitting the J48 classifier, which is why it was able to perform better than the other subsets of attributes. When it was applied to our logistic regression model, it was not equip to handle the data since it was under new circumstances. It likely did not have enough data to accurately predict whether the patient had breast cancer.



Finally, we looked to compare the results of our best classifier to date, the correlation based feature selection, to the boosting model. Since ensemble techniques are known to work better than single classifiers, we were expecting this to perform similarly, and likely more accurately, than our current model. After implementing the boosting model with the J48 decision tree as the base classifier, we found that the accuracy was 76.72%. This was 0.86% lower than our finely tuned logistic regression model, which is not statistically significant, but a difference nonetheless.

Conclusion

After continuously improving upon our findings each iteration, we found that the best model to predict whether or not a patient has breast cancer is a logistic regression algorithm that implements correlation based feature selection. While it is not statistically more improved than the boosting method, it does perform the best with an accuracy of 77.58%. We were able to bring the accuracy from its base performance of 73.27% to 74.84% by improving the parameters that the model used to train before incorporating the new feature set and arriving to our best model. A notable observation is that while we were able to increase the accuracy by 4.31 percentage points, or 5.9%, our largest improvement was with the artificial neural network. Just by tuning the number of hidden nodes and layers, the classification accuracy jumped from 65.5% to

71.97%. This is an increase of 6.47 percentage points, or 9.9% of the original accuracy. In the future, implementing feature selection for this model may prove to yield even more successful results.

Our initial hypotheses were overall correct. The attributes in our dataset was overwhelming connected with each other, so we used that fact to our advantage. We assumed that classifiers that account for these correlations would perform better than those that did not. Seeing that the artificial neural network and logistic regression successfully predicted the results, we were correct in that regard. We also noted that correlation based attribute selection would be one of the best methods at improving our performance accuracy for the same reason. This turned out to be true as well, providing us with our best model to date. One way we were incorrect was our assumption that the ensemble learning technique would be the best. It did perform better than other base models, but it did not beat our fully optimized logistic regression model. This may be because of the same dataset; given so few instances to work with, it was not able to produce a good enough model to accurately predict cancer incidences.

If we were to continue our work, the natural next move would be to rework our dataset. It was incredibly limiting to only have 116 instances. In accordance with good data mining practice, we would've had enough samples to set aside a test set. This would've been unseen by the model so it could better test how accurate it performs. Since we had such little data to work with, we felt it was better to use it all in training the model. Given more time, we would expand our dataset with more examples so we could make better predictions. In addition, we would expand the types of models tested and see if any other types of techniques work well with our dataset. There are many other meta-learning methods, such as bagging and stacking, and variations of logistic regression that could possibly have better results than the correlation based logistic regression model.

References

[1] National Vital Statistics System. Vital statistics of the United States, 1980. Volume II —Mortality, part A. 1985; public-use 2016 Mortality File; Xu JQ, Murphy SL, Kochanek KD, Bastian B, Arias E. Deaths: Final data for 2016. National Vital Statistics Reports; vol 67 . Hyattsville, MD: National Center for Health Statistics. 2018. Available from: https://www.cdc.gov/nchs/products/nvsr.htm.

[2] Ripperger, T., Gadzicki, D., Meindl, A., & Schlegelberger, B. (2008). Breast cancer susceptibility: current knowledge and implications for genetic counselling. *European journal of human genetics : EJHG*, *17*(6), 722-31.

[3] Why is early diagnosis important? (2018, August 23). Retrieved from https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important

[4] Hou, Y., Zhou, M., Xie, J., Chao, P., Feng, Q., & Wu, J. (2017). High glucose levels promote the proliferation of breast cancer cells through GTPases. *Breast cancer (Dove Medical Press)*, *9*, 429-436. doi:10.2147/BCTT.S135665

[5] Higher Insulin Levels Linked to Worse Prognosis in Metastatic Breast Cancer. (2015, November 19). Retrieved from https://www.breastcancer.org/research-news/insulin-levels-linked-to-mets-prognosis

[6] Sentinelli, F., Romeo, S., Arca, M., Filippi, E., Leonetti, F., Banchieri, M., . . . Baroni, M. G. (2002). Human Resistin Gene, Obesity, and Type 2 Diabetes. *Diabetes, 51*(3), 860-862. doi:https://doi.org/10.2337/diabetes.51.3.860

[7] Jamaluddin, M. S., Weakley, S. M., Yao, Q., & Chen, C. (2012). Resistin: functional roles and therapeutic considerations for cardiovascular disease. *British journal of pharmacology*, *165*(3), 622-32.

[8] General (current) breast cancer knowledge: Ripperger, T., Gadzicki, D., Meindl, A., & Schlegelberger, B. (2008). Breast cancer susceptibility: current knowledge and implications for genetic counselling. *European journal of human genetics : EJHG*, *17*(6), 722-31.

[9] Macis, D., Guerrieri-Gonzaga, A., & Gandini, S. (2014). Circulating adiponectin and breast cancer risk: a systematic review and meta-analysis. *International journal of epidemiology*, *43*(4), 1226-36.

[10] Juhola, M. (2018, June 11). *Neural Networks Basics*. Retrieved from http://www.uta.fi/sis/tie/neuro/index/Neurocomputing2.pdf

[11] Panchal, F. S., & Panchal, M. (2014). Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. *International Journal of Computer Science and Mobile Computing, 3*(11), 455-464. Retrieved from https://www.ijcsmc.com/docs/papers/November2014/V3I11201499a19.pdf

[12] Esbensen, K. (1987). Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems, 2(37), 52nd ser. Retrieved from http://files.isec.pt/DOCUMENTOS/SERVICOS/BIBLIO/Documentos de acesso remoto/Principal components analysis.pdf

[13] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR (2013) Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Health Med Inform 4:124. doi: 10.4172/2157-7420.1000124

[14] Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*. https://doi.org/10.1177/117693510600200030

[15] Hall, M. A. (2000). *Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning*(Working paper No. ISSN 1170-487X). Hamilton, NZ: University of Waikato.

[16] Hall, M. A., & Smith, L. A. (n.d.). *Feature Subset Selection: A Correlation Based Filter Approach* (Working paper). Hamilton, NZ: University of Waikato.

[17] Schapire, Robert E. (2001). *The Boosting Approach to Machine Learning An Overview.* Pages 1-16. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.5565&rep=rep1&type=pdf