Breast Cancer Data Analysis Project Paper

By

Abdullah Ayesh, Kortney Curry, Ryan Dagnino, Brian Marmolejo

Final Project Paper

Document

Submitted towards fulfillment of the requirements for
Mathematical Methods in Data Science – Fall 2023

Arizona State University
Tempe, AZ 85281

Introduction

There are a multitude of mathematical methods across concepts such as linear algebra and other fields which can be used for various applications such as data analysis and data science. The utilization of these various concepts can be incredibly useful for deriving essential information from given datasets. The dataset which will be analyzed will be that of the breast cancer diagnostic dataset obtained by the University of Wisconsin. Some of the data which will be derived and analyzed from this dataset includes whether the cancer is malignant or benign, a linear regression of the data, and various probabilities pertaining to the cancer. In order to formulate evaluations and predictions for the dataset, methods such as logistic regression and support vector machines will be used. Using the concept of logistic regressions, predicted values can be determined from factors within the dataset, specifically by basing the factors against one another.

The process of Support Vector Machine can be used for data classification in determining the boundaries between different variables and levels within the dataset and is useful for complicated datasets like breast cancer diagnostic dataset. In regard to the information that this project hopes to explore, the method of Support Vector Machine (SVM) will be used to determine the nature of the tumors as well as using SVM to find the approximation rations and determine the accuracy and implementation of the aforementioned data. Using this alongside the methods of K-means and logistic regression, the minimized and full variable sets can be derived in order to assess the data and formulate accurate classifieds and predictive models regarding the cancer. The utilization of these methods will allow for the formulation of a full project, in which related work, methodology, experiment setup, results, and comparison of results can be explored and detailed.

In these various sections, discussion of similar projects as well as referencing the initial project and dataset will be explored to determine how the dataset varies to other experiments as well as the conducted experiment, with a deeper dive being made into the logistic regression and SVM methodology. An overview of the code behind the experiment, an outline of the methodology that will be used for the experiment, discussion of results, and comparison of these predictions will be conducted in order to determine the effectiveness of the methods used within the project and to see how they compare to the initial dataset. In order to begin this project, the topic must be investigated further, the code must be created, and the data will be analyzed and compared. This will culminate together to create a complete project regarding the analysis of the aforementioned dataset.

Related Work

Observing the experimental work conducted by the University of California Irvine, through the utilization of their Machine Learning system, breast cancer diagnoses were pulled from Wisconsin from an older nuclear extraction of tumors obtained from breast cancer. In the original dataset obtained by authors of a research paper from 1993, various image processes were used for diagnoses related to breast tumors which resulted in the development and improvement of breast cancer imaging methods and technologies. Based on the data obtained from this original dataset obtained from physical imaging technology, these features were then processed into images to create various variables regarding the tumors, masses, and various other features. This data obtained using the imaging tech was then inputted into machine learning systems, in which attributes were created such as the diagnosis of the cancer type, size of the tumor, texture, symmetry, and various other factors describing the cancer diagnosis providing various numerical and boolean analysis and information.

As per the CDC data on breast cancer statistics in Wisconsin from the years 2016 to 2020, there were 24,685 new breast cancer diagnoses. This is on average 135 diagnoses per 100,000 people. Of those diagnosed, 3,610 died due to cancer related complications. There are four types of breast cancer as stated by the cancer.org breast cancer research, which includes Luminal A, Luminal B, Basal-like, and HER2-enriched. Luminal A is by far the most common and is a slow growing cancer that can be treated with hormonal therapy. Luminal B cancers are more difficult to treat in comparison with Luminal A, due to being HR+ and HER2+. Basal-like cancers are even more difficult to treat due to a lack of advancement in this specific subcategory. HER2-enriched is by far the least common however is more treatable due to various therapies.

Benign breast tumors are categorized into three groups according to cancer.org's breast cancer research. These groups include nonproliferative lesions, proliferative lesions, and proliferative lesions with atypia. Nonproliferative lesions are the standard tumor that does not present itself in a dangerous way, the risk of cancer with these is little. Proliferative lesions include symptoms of hyperplasia and increase the cancer risk to 1.5 times those with nonproliferative lesions. Proliferative lesions with atypia present a risk of cancer 4 times greater than nonproliferative lesions and can include atypical ductal hyperplasia.

Proposed Methodology

The accurate assessment of all the associated data is incredibly important, and particularly for utilizing the aforementioned methods to determine various variables and analysis criteria. There will be a total of 3 different methodologies which will be utilized to determine various aspects of the experiment, that being the method of Support Vector Machines, K-means, and logistic regression. The process of Support Vector Machines can be used for data classification in determining the boundaries between different variables and levels within the dataset and is useful for complicated datasets like breast cancer diagnostic dataset. In regard to the information that this project hopes to explore, the method of Support Vector Machine (SVM) will be used to determine the nature of the tumors as well as using SVM to find the approximation rations and determine the accuracy and implementation of the aforementioned data. After the data is collected using the method of SVM to determine precision, recall, and F1-score, the data must be minimized in order to utilize the methodologies of K-means and logistic regression.

After the SVM model is trained on the dataset, another SVM classifier will be trained, after which the original data and minimized data will be evaluated, and the process of k-means, which entails the utilization of an algorithm which will determine distances for point assignment within a cluster, which will provide insight into the data structure. After using the process of k-means for a comparison of clustering results before and after feature reduction, and assessing the impact of dimensionality reduction on clustering quality, logistic regression will be conducted. Using logistic regression, the minimized and full variable sets can be derived in order to assess the data and formulate accurate classifieds and predictive models regarding the cancer. After introducing logistic regression as an alternative classification method, training the models on both the original and minimized datasets will allow further conclusions to be determined about the feature minimization process to obtain more accurate information regarding the cancer types and sizes.

In order to begin exploration of this dataset in greater depth and attempt to find ways to minimize the amount of data required for an accurate classification of tumors as benign or malignant. By leveraging point-based correlation coefficients, we can identify the most significant variables and eliminate those with minimal correlations. This is crucial as it reduces the number of observations needed, enabling faster diagnoses. Once the variables have been minimized, we can utilize a Support Vector Machine (SVM) to determine the malignancy or benign nature of tumors. We will then compare the approximation ratio of the dimensionally minimized SVM with the SVM using all variables to assess its adequacy and meaningfulness in implementation. Additionally, we can employ other tools such as K-Means and logistic regression to compare the minimized and full set of variables, aiming to identify the most effective approach for tumor classification.

Experiment Setup

The experiment is set up in the Python programming language, utilizing the sklearn library for convenient access and utilization of the breast cancer dataset. Additional libraries, including seaborn, pandas, numpy and scipy, are imported to facilitate various calculations.

The initial phase of the experiment involves applying a Support Vector Machine (SVM) classifier to the original dataset, which is then split into 70% training and 30% testing subsets. The SVM model is trained on the training set and evaluated on the testing set, employing classification metrics such as precision, recall, and F1-score. Subsequently, a point-biserial correlation analysis is conducted on the original dataset with a high correlation threshold of 0.70. This analysis identifies features strongly correlated with the target variable, and a subset of these features is isolated to create a new dataset referred to as the minimized-data.

The second part of the experiment entails training another SVM classifier on the minimized-data set, and its performance is evaluated using the same procedures applied to the original dataset. To further explore and understand the differences between the minimized-data and the original dataset, k-means clustering is employed, providing an additional perspective on the underlying structure of the data. The k-means analysis allows for a comparison of clustering results before and after feature reduction, aiming to uncover any noticeable changes in grouping patterns and assess the impact of dimensionality reduction on clustering quality. Additionally, logistic regression is introduced as an alternative classification method, training models on both the original and minimized datasets to draw further conclusions about the feature minimization process. This comprehensive, multi-faceted approach contributes to a thorough understanding of the breast cancer dataset and sheds light on the behavior of machine learning models in distinct feature spaces.

Results Discussion

This section showcases an in depth analysis of the results and accuracy of these results obtained from performing SVM, k-means analysis, and logistic regression on the breast cancer data. Each test was conducted twice for both datasets, the original and minimized. The original dataset used 30 features, whereas the minimized dataset only used 8 features. This was to find a proper diagnosis more efficiently and to remove noise from the dataset.

When training and fitting a SVM model for the original dataset, the accuracy was found to be 0.95, along with a precision of 0.95, a recall of 0.95, and a f1 score of 0.95. Training and fitting another SVM model to the minimized dataset obtains the same classification metrics as the original dataset. These classification metrics are of great importance when dealing with medical investigations. The precision score is used to measure the proportion of positively predicted labels to the total amount that are expected to be positive. The higher the precision of the model, the greater the amount of minimized false positives. Since the SVM model produced a precision of 0.97, there are minimized false positives. This means that the model performs well with correctly identifying if a patient has a malignant cancer tumor. Recall score is the model's ability to predict the positives out of actual positives correctly. A high recall score means the model is better at identifying both positive and negative examples. The SVM model produced in this experiment had a recall of 0.97, meaning that the model does well in minimizing the chance of not detecting a malignant cancer tumor. An alternative to the accuracy metric is the f1-score which measures the performance of the model in terms of accuracy by giving equal weights to both the precision and recall. When the model has a high f1-score it shows that the model has produced few false positives and false negatives or in the case of this experiment few false benign cancer tumor classifications and few malignant cancer tumor classifications.

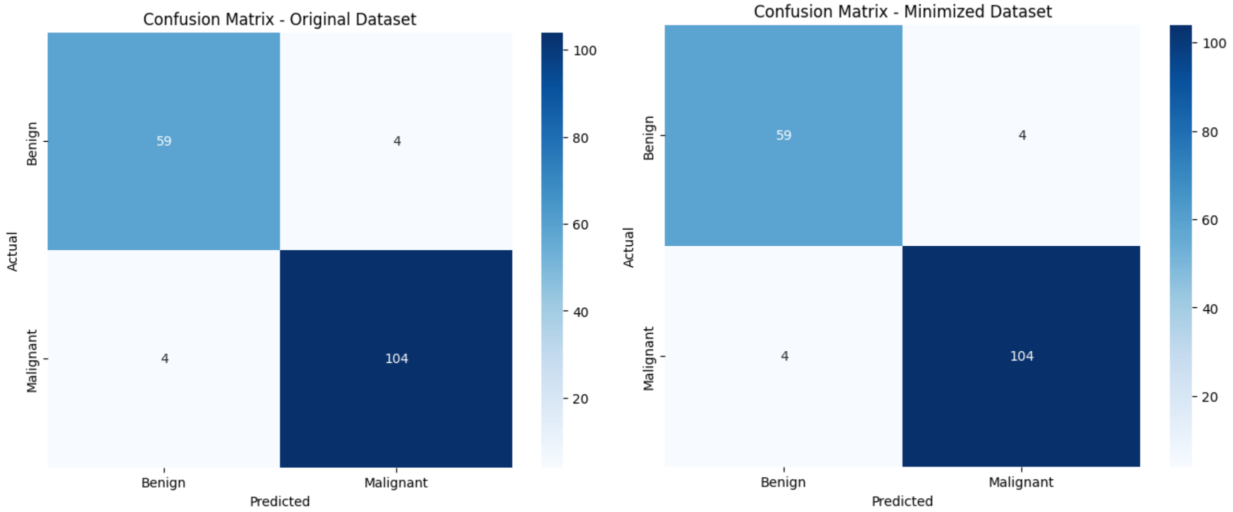

FIG 1. These images show the confusion matrices for the original and minimized datasets for SVM. The minimized performs equally well the the SVM utilizing the original dataset. This is likely due to how the SVM assigns weights in the original dataset, a process performed manually in the minimized version.

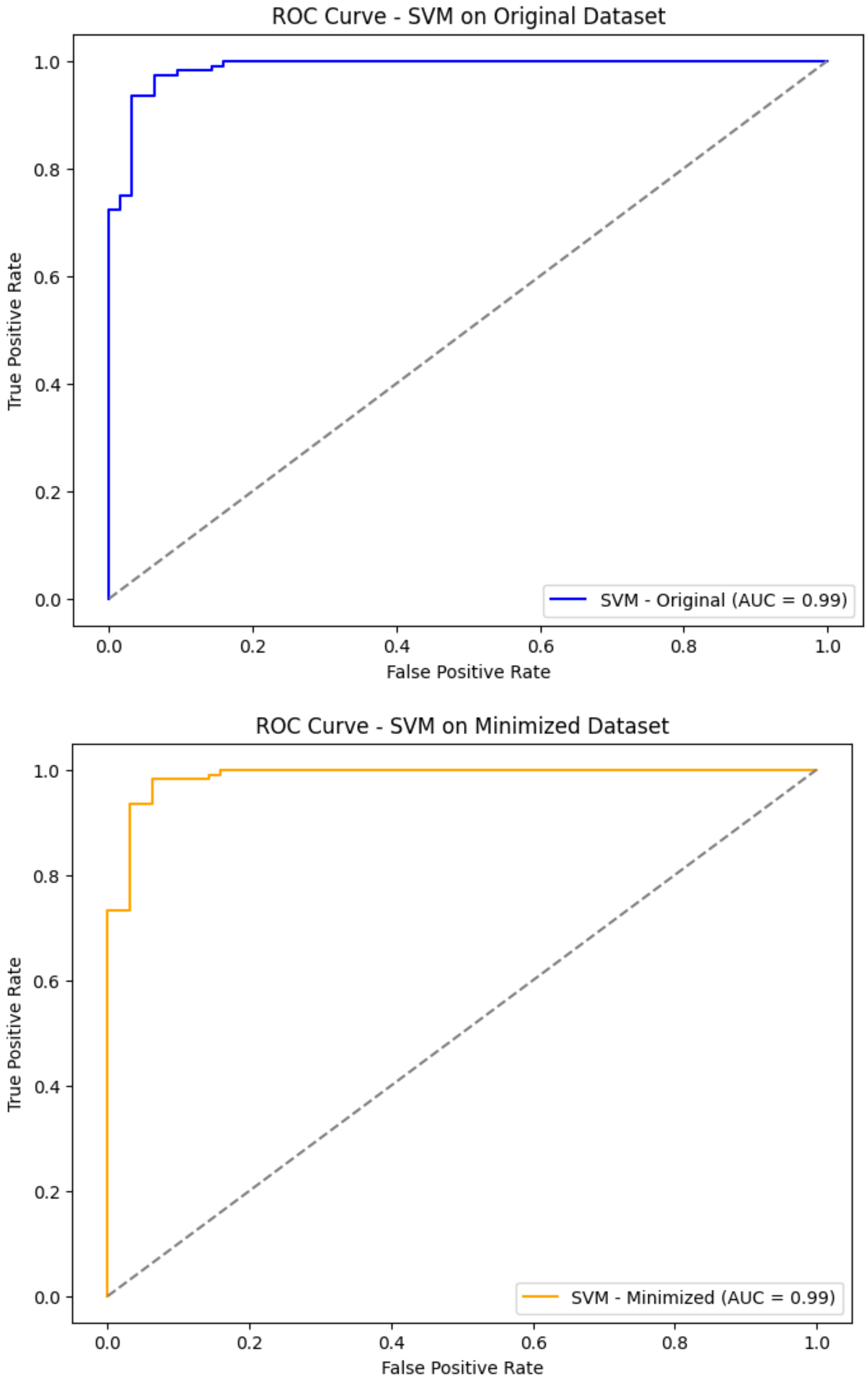

FIG 2. These images show the ROC curve for both the original and the minimized dataset for SVM. It is observed in the minimized ROC curve that the minimized version performs nearly similar to the original dataset in logistic regression at all thresholds. This is likely due to how the SVM assigns weights in the original dataset, a process performed manually in the minimized version.

A method used to better understand the structure of the data was k-means analysis was used of both the original and minimized dataset. Finding the number of clusters to be two, a silhouette and Adjusted Rand Index (ARI) score was used to better understand the nature of the two clusters. The original dataset had a silhouette score of 0.697 and the minimized dataset has a score of 0.698. A silhouette score measures the distance of separation between clusters, which ranges from -1 to 1. The higher the silhouette score is the further away the clusters are. If the score is 0 it means that the two neighboring clusters are close to one another. Since the silhouette score for both datasets was around 0.69, it signifies that the clusters are moderately separate from each other in both datasets. The ARI score for the original dataset was 0.491 and the score for the minimized dataset was 0.486. The ARI predicts the similarity between true labels and predicted clusters. The ARI score of 0 indicates random labeling and a score of 1 indicates identical labeling. Both ARI scores for the datasets were around 0.5 which shows that clusters aren't neither random nor are they identical, they are adequate.

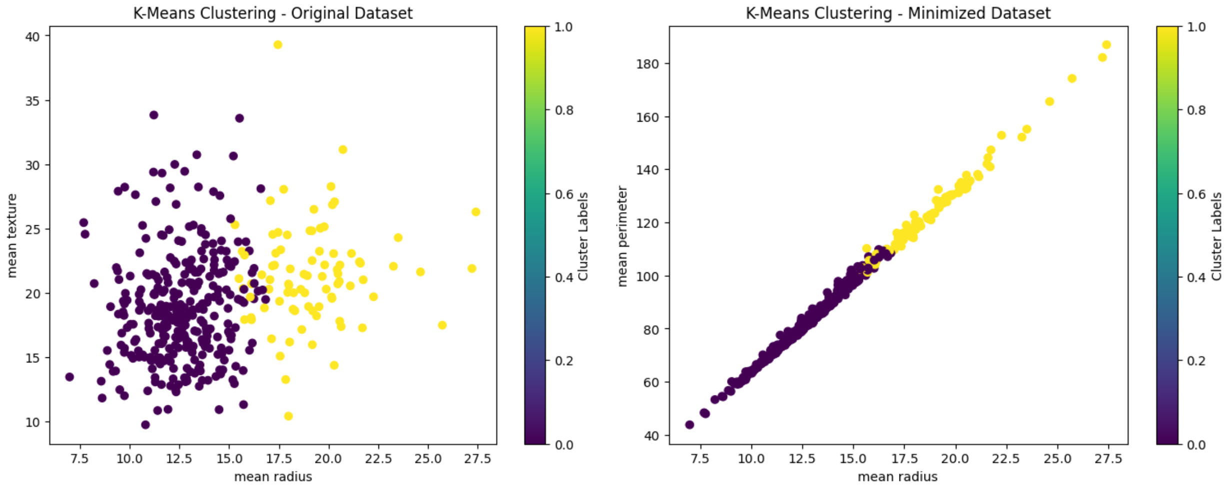

FIG 3. These images show a k-means clustering for two features from the original dataset, and the minimized dataset. This plot is not intended to represent approximations that can be made using specific values, but rather how the "shape" of the data changes when removing a feature, in this case, mean texture was removed.

Logistic regression was also used as an alternative classification method. Using the training model for the original dataset the following classification metrics were obtained: an accuracy of 0.97, a precision of 0.97, a recall of 0.96, and a f1-score of 0.97. Another similar training model was constructed for the minimized dataset and the following classification metrics were found: an accuracy of 0.96, a precision of 0.96, a recall of 0.95, and a f1-score of 0.96. Although the logistic regression model had higher classification metrics than the SVM model used, the difference wasn't significant. Both models did well in being able to identify whether a can tumor was benign or malignant.

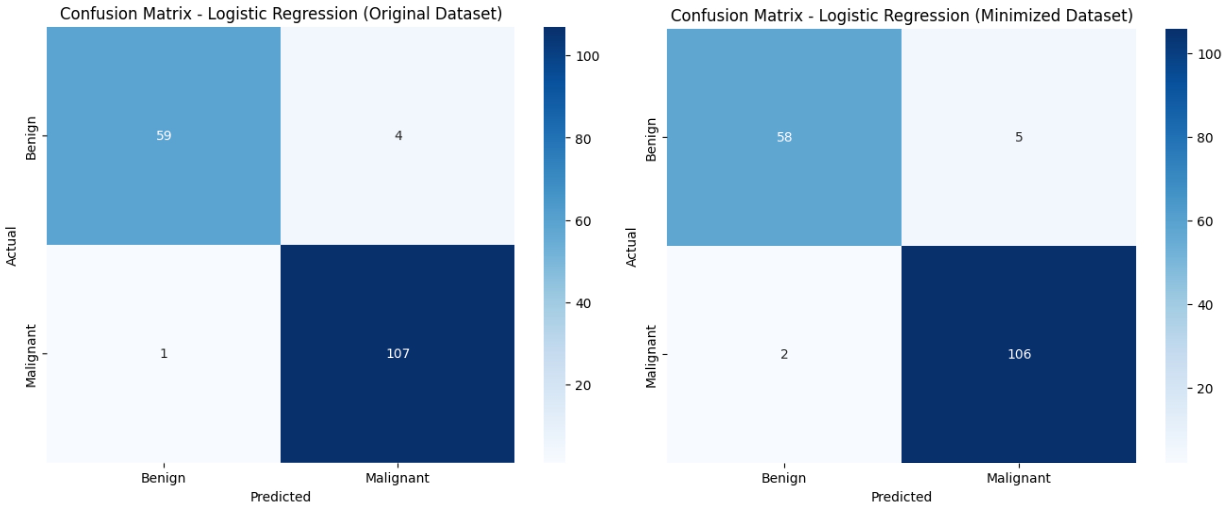

FIG 4. These images show the confusion matrices for the original and minimized datasets for logistic regression. The minimized makes false classifications more frequently than the logistic regression using the original dataset, though the margin is small.

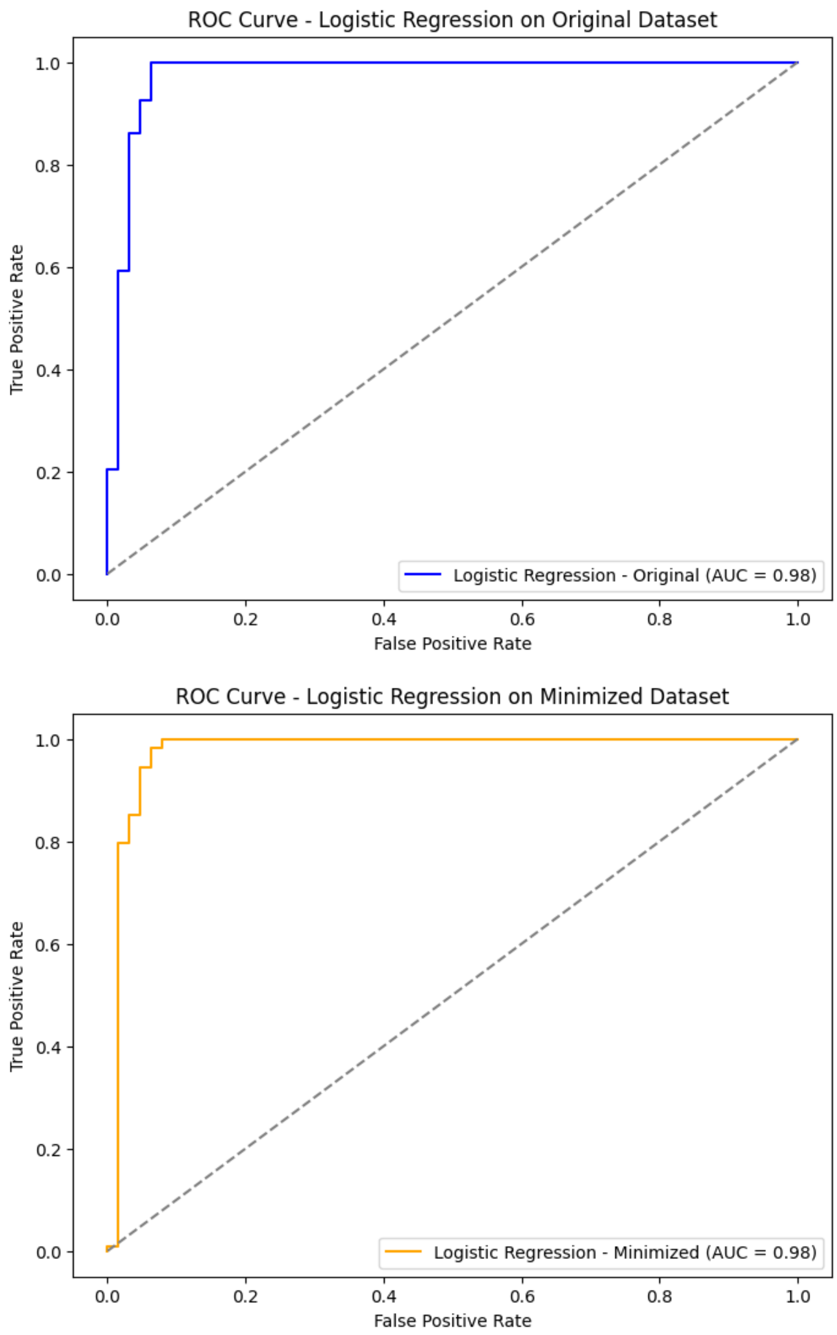

FIG 5. These images show the ROC curve for both the original and the minimized dataset for logistic regression. It is observed in the minimized ROC curve that the minimized version performs better than the original dataset in logistic regression at low thresholds. This is likely because the minimized-dataset was already computed at a 0.7 threshold.

Comparison

Comparing the model performance of our support vector machine tests and those from the UCI paper shows that our tests yielded an accuracy of 0.95 and a precision of 0.95, while those in the UCI paper yielded 0.944 accuracy and 0.947 precision. That is a 0.006 difference in accuracy and a 0.003 difference in precision. Comparing the model performance of our logistic regression tests with those published by UCI shows that our tests had an accuracy of 0.97 and a precision of 0.97, while the other paper had a 0.958 accuracy and a 0.955 precision. That is a 0.012 difference in accuracy and a 0.015 difference in precision. Our models had a higher accuracy and precision on average compared to those shown in the UCI paper, however these higher accuracy and precision levels are insignificant as both papers have a high level of accuracy and precision.

Another comparison can be made between the confusion matrices in which the original and minimized datasets seem to be quite similar in the distribution in regard to the predicted sets of benign and malignant tumors. Along with this, observing the K-mean clustering, it can be seen that in the original dataset, the points are clustered quite closely together, with no discernable shape, however when the K-means clustering was conducted for the minimized data set which had quite a linear distribution with the mean perimeter increasing as the mean radius increases. This plot is not intended to represent approximations that can be made using specific values, but rather how the "shape" of the data changes when removing a feature, in this case, mean texture was removed when using the k-mean cluster method.

A comparison of the methods used within our own paper shows that the logistic regression classification method had a higher accuracy and precision than the SVM classification method. The difference in accuracy and precision are 0.02 and 0.02 respectively. However, as stated earlier in the paper, these differences are negligible enough and both classification models can be used effectively.

Conclusion

There are two types of breast cancer tumors, benign and malignant. Benign tumors are not cancerous, but certain types of benign tumors can increase the cancer risk. Malignant tumors are cancerous and invade and damage surrounding tissue. To better understand whether a tumor is benign or malignant we will first leverage point-based correlation coefficients to find the most significant variables. Using these variables, we will use a Support Vector Machine (SVM) to identify whether a tumor is benign or malignant. We will then ensure that the SVM makes a good approximation ratio for minimized variable count by comparing it to a SVM using all the variables provided. We will also use other variables such as K-mean and logistic regression to compare minimized and normal variables.

Using the aforementioned methodologies, results, and comparisons that were derived, it was determined that the conducted experiment yielded an incredibly high degree of accuracy in determining whether a tumor is malignant or benign, thus training and fitting the SVM model was successful as the model was trained to handle the original and modified datasets with high accuracy. This modeling and other methodologies which were used allowed for the predicted and actual determination of what the diagnosis type for the tumors would be, and considered other factors which include but are not limited to tumor size, perimeter, and radius, which found in some instances that the the perimeter was found to be as high as 180mm and the radius was found to be as high as 27.5mm for the minimized dataset, which in the original dataset had a similar radius, but the perimeter had some variation.

Overall, it was found that the experiment conducted within this report shared many similarities to those derived from the original report and dataset, with the data and results obtained from our paper being more accurate and precise due to the SVM and minimization processes used within our code for analyzing the dataset. However this difference is very miniscule and is negligible, however this does indicate success in the created code and analysis methods used. This found that overall, the accuracy, precision, recall, and F1-score for both the original and minimized datasets were all quite close to scores of 1, with all of them being more than a baseline of a score of 0.9.

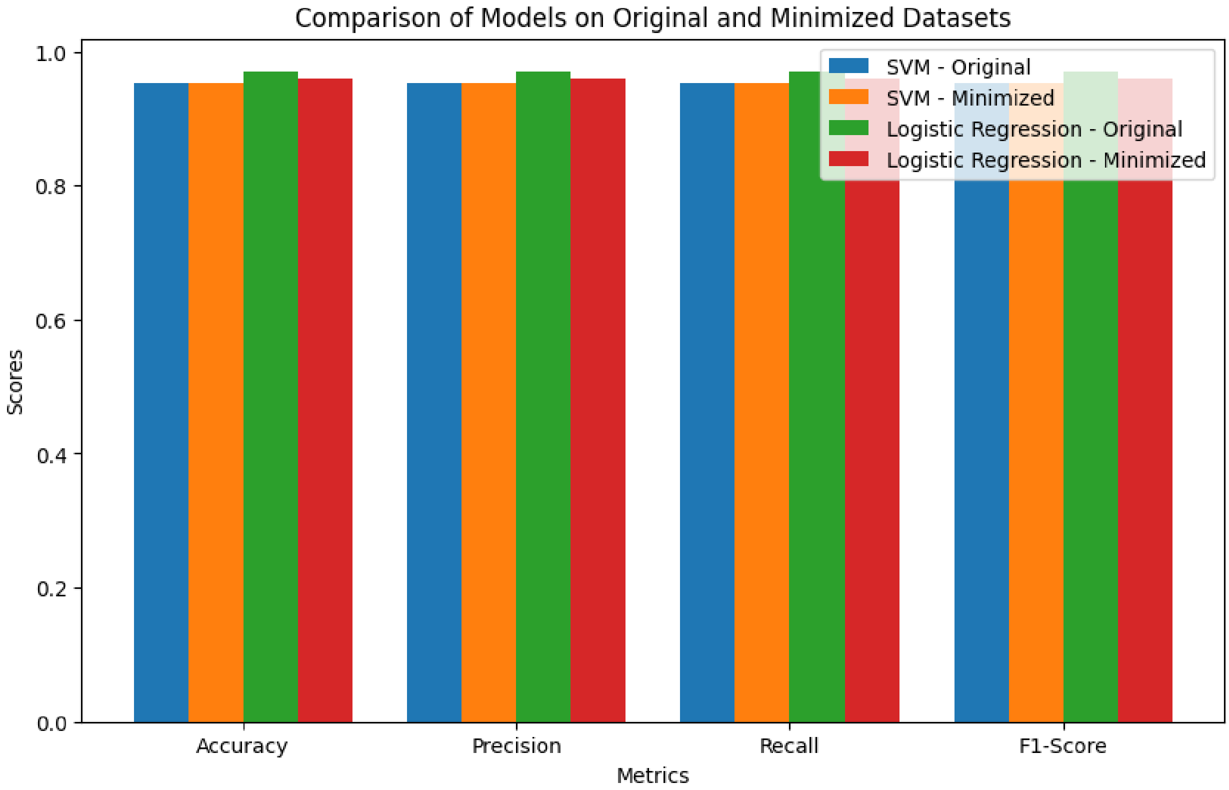

FIG 6. This image shows the scores for both the original and minimized datasets in SVM and logistic regression. It is important to note that they have near similar results despite the minimized having 8 features and the original dataset having 30 features.

Acknowledgements

The authors would like to thank Professor Wang for his efforts and work throughout the semester.

Author Contributions

This paper is the culmination of collaborative work amongst all the authors.

Ethical Standard

This experiment does not contain any performed studies pertaining to animals or human participants.

Data Availability

The authors declare that all data and code supporting the findings of this research are available below:

https://github.com/k4404c/MAT-422/blob/ebcc3bf7a7f7611943441ffc4661f50af8fa49af/MAT422Proj.ipynb

References

Breast Cancer Facts & Figures 2019-2020 - American Cancer Society, www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf. Accessed 21 Oct. 2023.

"Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. Accessed 22 Oct. 2023.

Learning, UCI Machine. "Breast Cancer Wisconsin (Diagnostic) Data Set." Kaggle, 25 Sept. 2016, www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data.

"USCS Data Visualizations - CDC." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, gis.cdc.gov/Cancer/USCS/#/StateCountyTerritory/. Accessed 21 Oct. 2023.

Nuclear Feature Extraction for Breast Tumor Diagnosis - Semantic Scholar, www.semanticscholar.org/paper/Nuclear-feature-extraction-for-breast-tumor-Street-Wolberg/53f0fbb425bc14468eb3bf96b2e1d41ba8087f36. Accessed 25 Nov. 2023.