Machine Learning Application on Prediction of Male Breast Cancer with PLCO Dataset

Juntao Li[1] and Ganesh Mani[#]

[1]Cannon High School, Concord, NC, USA
[#]Advisor

ABSTRACT

The objective of paper is to explore and examine the applicability of machine learning models on Male Breast Cancer with PLCO dataset. People who are unaware of the potential danger of getting breast cancer like males would not have the medical awareness beforehand for predictions. Therefore, the PLCO trials dataset consisting of ages, prostate status, marriage status etc. from National Institute of Cancer is used in this research for detection. The main purpose of using PLCO test is to discover the potential risk of getting an Male Breast Cancer (MBC) as soon as possible with low cost and easy collection. It is the rarity of MBC that imposes the threat for males who are unaware of the danger. To explore the relatively most suitable models to use for detecting MBC using non-traditional PLCO test dataset, different existing models including decision tree, random forest, DBSCAN, One Class SVM and so on were used to fit the data. Due to its extremity of imbalance, evaluation comes from the combination of standard accuracy and Area Under the Receiver Operating Characteristics(AUROC) for the overall accuracy of those models mentioned above. K-means and Logistic Regression models performed best with the AUC score of 0.62 and 0.67. Results suggested that more efficient approaches for common male breast cancer diagnosis or more advanced models and algorithms are needed in further study.

Introduction

Breast cancer is one of the most commonly diagnosed cancers in women, second to skin cancer (Ormene, 2010). Due to Male Breast Cancer (MBC) 's rarity, it usually is detected in a later stage. Randomized studies are not feasible in this topic since the cases of MBC are small. Male breast cancer is rare and constitutes 0.5–1% of all patients with breast cancer (Sasco, 1993). MBC is much less studied since it's a rare disease in males. Although there have been some data in male breast cancer, most approaches for this disease in males come from the extrapolation of information about female patients with breast cancer (Ormene, 2010). Even though male and female are completely different genders, they still share similar structures of breasts. Both genders may have inherited mutations in their BRCA1 and BRCA2 genes that may increase cancer risk (Markman, 2021).

So far, the practices on analyzing MBC dataset are uncommon. One of the studies does have access to large data across 12 years with mammography (Yiming Gao et al. 2019). However, machine learning model application was not mentioned in the study.

Methodology

The Data

Due to the rarity of MBC, many datasets have limited access. The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial is a large-scale clinical trial to determine whether certain cancer screening tests can help reduce

deaths from prostate, lung, colorectal, and ovarian cancer ("PLCO Q&A", 2009). The Male Breast dataset is a comprehensive dataset that contains nearly all the PLCO study data available for male breast cancer incidence and mortality analyses. The dataset contains one record for each of the approximately 77,000 men in the PLCO trial.

Data Preprocessing

At first, Data samples are collected with several features and different values. The collected data can possess many issues like noisy data, outliers, missing data, duplicate data and biased data (Shalina et al. 2020). To overcome these data related issues data preprocessing should be performed. Originally, there were 76678 rows and 136 columns in the csv file. Out of these 77, 000 men, only 38 of them have confirmed breast cancer. To improve the accuracy, the data should be carefully cleansed. Instead of dropping columns with null values directly, each category was examined to improve the data in general. See Appendix A for the details of the data cleansing process. After deleting all the rows that contain blank values and irrelevant columns, there are 54708 rows and 45 columns left. The tool used in this section is Excel.

Feature Selection

The generation of heatmap from seaborn was used to find out the correlation between each feature, especially to the diagnosis of MBC. None of the features is more than 10% directly related to the factor "mbreast_cancer", which is the diagnosis of MBC. This section highlights the low correlation of the PLCO test dataset related to MBC.

Table 1. Top 5 most correlated features to mbreast_cancer variable

Features	Correlation
agelevel	0.010387
enlpros_f	0.008981
cigar	0.008794
age	0.008279
prosprob_f	0.006640

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Figure 1. A table for interpreting the correlation in general (credit: Parvez Ahammad)

Above all, the features in the dataset have negligible correlation to the male breast cancer detection. The low correlation of different features were expected due to the fact that PLCO trials are based on less technical data from patients than other tests like mammography.

Splitting & Scaling

Using train_test_split function from sklearn, the dataset was split into 0.75 training set and 0.25 testing set. To further let model, such as logistic regression, adapt to the data more efficiently, standardization of the data was necessary.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning (Goonewardana, 2019). Due to the general large data of about 41000 rows, some training models like one Class SVM and Spectral Learning were not able to finish the runtime in reasonable time span.

Training Models

This research utilized a various classfiers from scikit-learn tool. To fully comprehend the dataset, both supervised learning and unsupervised learning were used. The major difference between Supervised and Unsupervised learning algorithms is the absence of data labels in the latter. With unsupervised learning, data features are dumped into the learning algorithm, which determines how to label them and based on what, which dictates which unsupervised learning algorithm to follow (AI-Masri, 2019). To also compare the results of supervised learning and unsupervised learning, the number of clusters for all the clustering methods were all set to 2.

Decision Tree

Decision tree classifiers are able to work with numerical and categorical features. Since most data in PLCO tests contain categorical data, the decision tree would probably lead to a decent result. In addition, it is easy to understand and interpret, perfect for visual representation. There will be a plot to show the classifying process of the decision tree in this data set. Interestingly, this data set is extremely unbalanced.

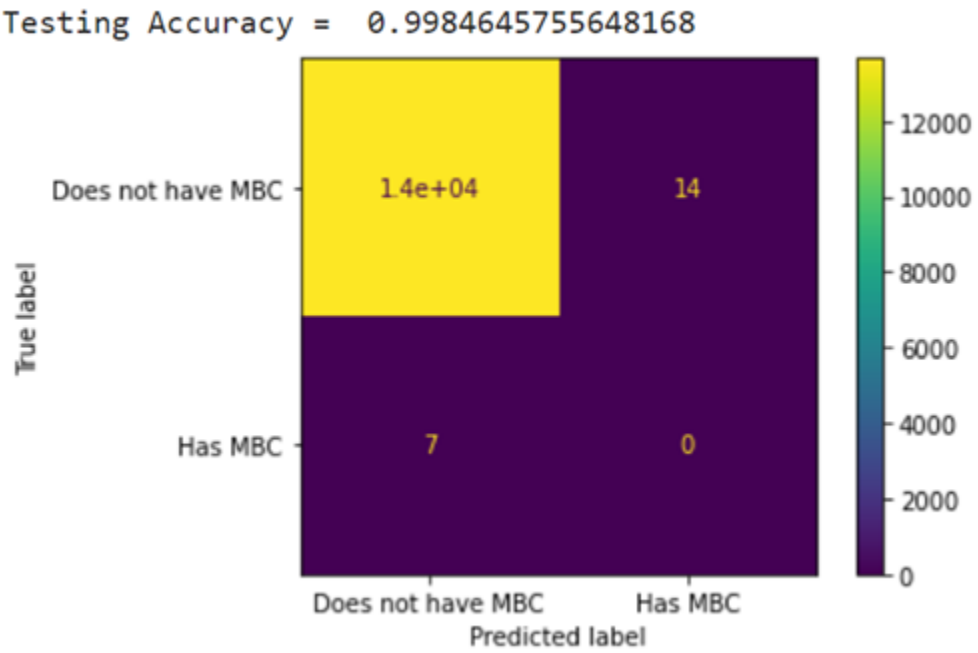

Figure 2. Confusion Matrix of Preliminary Decision Tree Classifier (pDTC)

Decision tree classifiers tend to overfit (Innab, 2019). One of the methods is to prune the tree by using cost complexity pruning (ccp) parameter "alpha". Through 5 fold cross validation and optimization, the alpha comes down to 1.4E-5. Pruning is a necessary process for this model. However, due to the unbalanced data, alpha barely affects the dataset.

Random Forest

Random forest is a technique used in modeling predictions and behavior analysis and is built on decision trees. It contains decision trees representing a distinct instance of the classification of data input into the random forest (Sklearn.ensemble.RandomForestClassifier, n.d.). The parameter "max_depth" = 2 refers to The maximum depth of the tree. The "random_state" was set to 42 for random number generation. Additionally, "class_weight" was set to "balanced". The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)).

Xgboost

XGB consists of a number of hyper-parameters that can be tuned — a primary advantage over gradient boosting machines. However, like any other boosting method, XGB is sensitive to outliers. The weak learners are considered to be regression trees while using gradient boosting for regression, in which each of the regression trees maps an input data point to one of its leaves that includes a continuous score. XGB minimizes a regularized objective function (regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting [Regularization (mathematics), n.d.]) that combines a convex loss function, which is established on the variation between the target outputs and the predicted outputs. The training then proceeds iteratively, adding new trees with the capability to predict the residuals as well as errors of prior trees that are then coupled with the previous trees to make the final prediction (Choudhury, 2021).

Logistic Regression

Logistic Regression is a classic algorithm to solve a classification problem. Logistic regression was used over the linear regression since the linear regression is susceptible to the outlier. When the logistic regression model comes across an outlier, the logistic equation will take care of it (Narkhede, 2018).

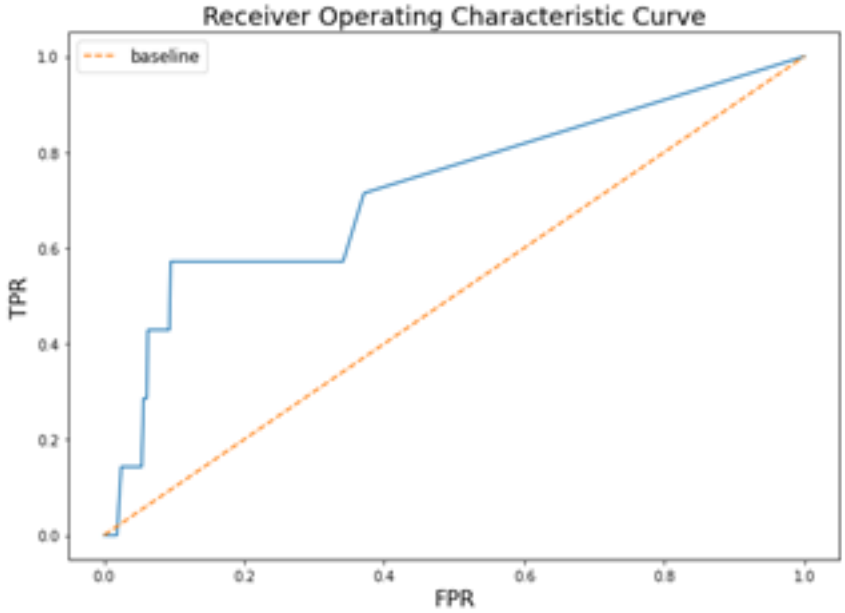

Figure 3. The visualization of AUC curve for Logistic Regression
The roc_auc_score is 0.6664071480823492.

Isolation Forest

Isolation forest works on the principle of the decision tree classifier. It isolates the outliers by randomly choosing a feature from the existing set of features and then randomly picking a split value between the maximum and minimum values of the chosen feature. This random partitioning of features will produce small paths in trees for the anomalous data values and distinguish them from the normal set of the data (Verma, 2020).

Spectral Clustering (Not Used)

In spectral clustering, the data points are considered as nodes of a graph. Thus, clustering is treated as a graph partitioning method. The nodes are then connected to a low-dimensional space that can be easily segregated to form clusters. It is worthy to note that no assumption is made about the shape/form of the clusters (Doshi, 2019). This model wasn't used to run through the dataset due to its large time complexity. Because it is a flexible class of clustering algorithms that can produce high-quality clustering on small data sets, but which has restricted applicability to largescale problems due to its computational complexity of O(n^3) (Karatsalos, 2018). However, a suggested model of fast approximate spectral clustering could be performed efficiently (Yan et al. 2009).

One Class SVM

It basically splits all the data points from the origin and maximizes the distance from this hyperplane to the origin. This leads to a binary function which captures regions in the input space where the probability density of the data lives. Thus the function returns +1 in a "small" region (capturing the training data points) and −1 elsewhere (Scholkopf et al. 2000).

K-means

The idea behind k-Means is that the k new points are added to the target data. Each one of those centroids will center itself in the middle of one of the k clusters. Once those points stop moving, the clustering algorithm stops. K is a hyper-parameter that should be set before training, which specifies the number of clusters for the algorithm to yield. This number of clusters is actually the number of centroids going around in the data (Al-Masri, 2019).

Local Outlier Factor (LOF)

Local Outlier Factor (LOF) is a score that tells how likely a data point is an outlier/anomaly. If LOF ≈1 ⇒ no outlier. If LOF ≫1 ⇒ outlier. Parameter k is introduced to indicate the number of neighbors the LOF calculation is considering. The LOF is a calculation that examines the neighbors of a certain point to find out its density and compare this to the density of other points later on. Determining a correct number k is not straight forward. While a small k has a more local focus, i.e. looks only at surrounding points that are close, it is more erroneous when having much noise in the data. A large k, however, can potentially miss local outliers (Wening, 2018). The parameter "n_neighbors" was set to 10, which represents the number of neighbors to use for "kneighbors" queries.

DBScan

DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions separated by regions of lower density. It groups data points that are considered "densely grouped" into a single cluster. It can recognize clusters in huge spatial datasets by inspecting the local density of the data points. The most notable feature of DBSCAN clustering is that it is robust to outliers. Therefore, it does not require the number of clusters to be told beforehand, unlike K-Means, where the number of centroids has to be specified (Sharma, 2020). The parameter eps was set to 3, which specifies how close points should be to each other to be considered a part of a cluster (Prado, 2017). Another modified parameter is min_sample = 10, which refers to the number of samples (or total weight) in a neighborhood for a point to be considered as a core point (Sklearn.cluster.DBSCAN, n.d.).

Results

The results obtained after applying the models to the PLCO trials dataset are presented, analyzed, and discussed in this section.

To briefly address the performance of methods, the column 'ENSEMB' was added to conclude the scores of 'ISOFOR', 'SVM-1C','KMEANS', and 'LOCOUT'.

Table 2. The results of all classifiers that successfully performed.

Classifiers	Accuracy Score	AUROC Score
Isolation Forest	0.74	0.46
Local Outlier Factor	0.83	0.5
SVM-1Class	0.44	0.39
DBSCAN	0.15	0.49

KMEANS	0.66	0.62
ENSEMB	0.81	0.49
XGBOOST	0.995	0.5
Decision Tree	0.9985	0.4995
Logistic Regression	0.9	0.67

The accuracy rates are high in general, although AUROC values are not as high as the accuracy, which is a likely side effect of the highly imbalanced data and many false positives/negatives. The “ENSEMB” is the mode of the predictions from 'ISOFOR','SVM-1C','KMEANS','LOCOUT', and 'DBSCAN'.

To show the distribution of the classfiers’ results on each class, boxplots of the prediction results are presented below.

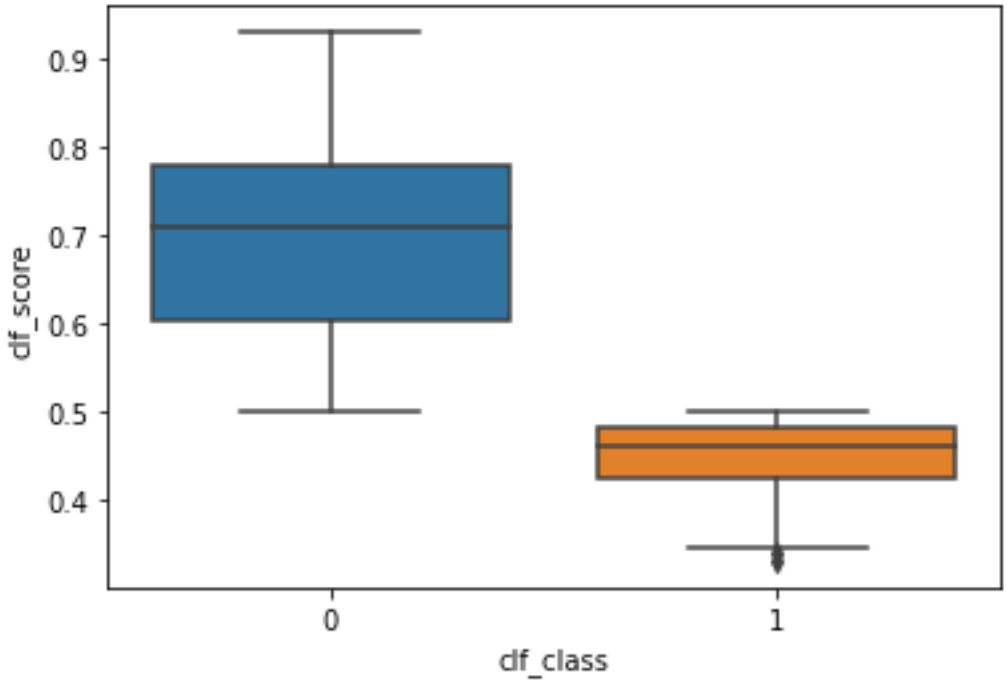

Figure 4. The boxplot from other predictions of “clf_class”

The “clf_class” represents the prediction results of the model Random Forest. X axis represents the prediction results of “0” or “1”. “0” represents the non-existence of MBC, whereas “1” represents the existence of MBC. “clf_score” represents the prediction accuracy of all the individual samples.

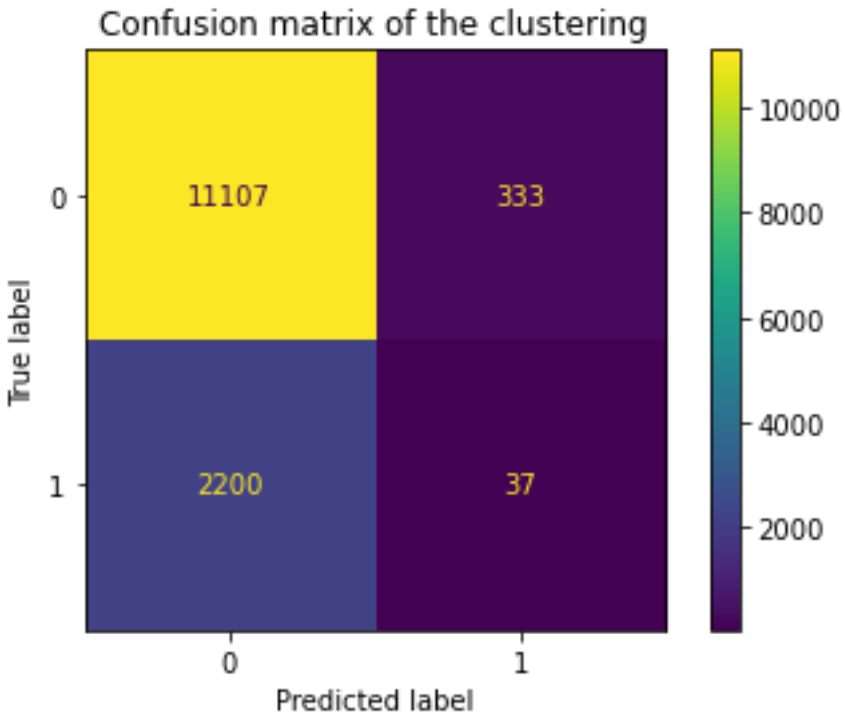

Figure 5. Confusion Matrix for “ENSEMB” versus “clf_score”

Any binary classifier classifies the results in terms of positive or negative, hence the name. So the results that the classifier predicts to be positive and are actually positive are the True Positives, the results that were falsely classified as positive but are in fact negative are False Positives. True Negatives are actual negative class predicted as negative and False Negatives are the falsely predicted negative but are in fact positive (Kunanbaeva, 2019).

To evaluate the quality of clustering methods, a Box-plot of the random forest classification scores against the predicted classes (0 and 1) from the "ensemble" cluster was performed.

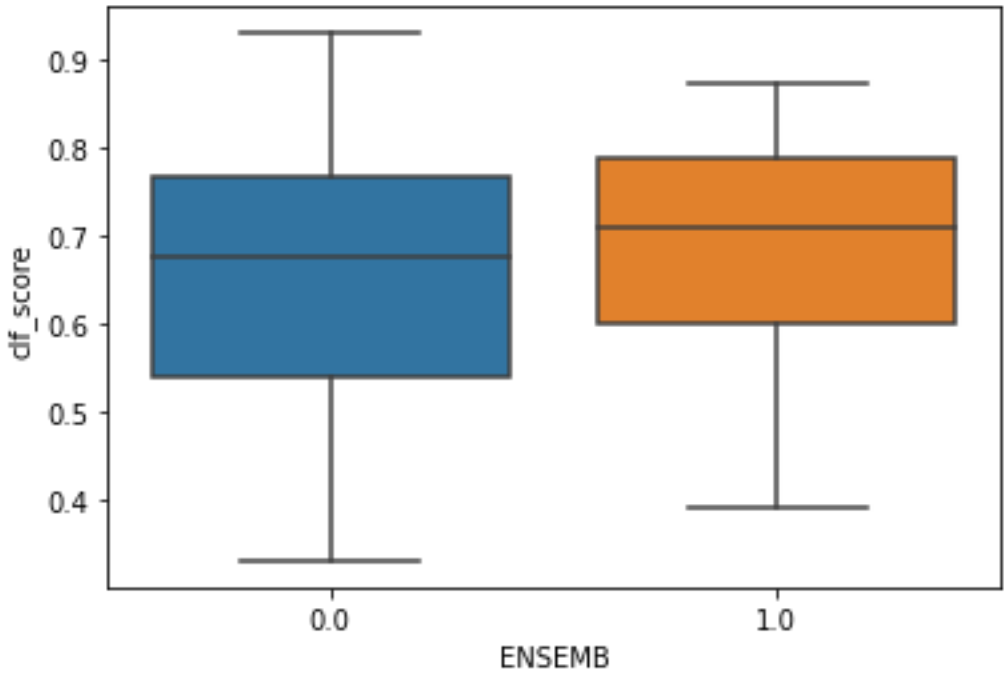

Figure 6. Box-plot displaying distribution of the modes of the prediction from "ENSEMB" against the prediction from "clf_score"

This indicates that the quality of clustering is poor, which leads to the concern of potential flaw for the clustering models when being used to operate on this particular dataset.

Discussion

The practice of machine learning on Male Breast Cancer was performed above, which proposed two potential model, K-means and Logistic Regression, among all models above for prediction of MBC based on PLCO dataset. Even though the application of models could be preliminary compared to other brest cancer study, the application on male breast cancer is specifically unprecedented. Especially with the non-traditional dataset from NIH, the fitting and training appear to be difficult to approach. In general, the parameters in all the models mentioned were mostly set to default or automatic values. Nevertheless, the parameters executed by the models would significantly affect the results, which suggest that the accuracy of this output may not be accurate.

Limitations & Flaw

PCA

PCA should be used mainly for variables which are strongly correlated. If the relationship is weak between variables, PCA does not work well to reduce data. Refer to the correlation matrix to determine. In general, if most of the correlation coefficients are smaller than 0.3, PCA will not help (17.7.1 Principal Component Analysis, n.d.).

Decision Tree

This model tends to overfit in some cases, especially for the unbalanced dataset of PLCO trials. Hence, the random forest model was performed to implement the potential flaw.

Related Studies

Age level

Shalini M has applied machine learning techniques on various breast cancer datasets (Shalini M et al. 2020). It suggests that Breast cancer detection and identification can be done from gene expression and large datasets. It also mentions that Breast cancer may be due to age factor and other gene mutation (Shalini M et al. 2020, p. 1), which corresponds to the highest correlation of age_level to mbreast_cancer in PLCO trail dataset in the section "Feature Selection" of this research. Overall, this paper helps strenthen the high correlation of age level with MBC in this dataset, albeit it didn't apply machine learning on MBC.

So far, the similar studies specifically for application of ML model on MBC could not be found, which indicates the significance of this topic.

Further Research

The application of these models on the dataset was simple and easy to replicate. Therefore, there is much space to improve on, especially the robustness of model training. To further produce better results on the existing data, the focus should be on selecting more proper features and tuning the parameters for a better fitting model.

Conclusion

The necessity of applying ML model on MBC has been addressd and the practice of application has also been attempted. With the training of various models, K-Means model produced the highest AUROC score among the unsupervised clustering methods mentioned above. The basic model of Logistic Regression has the highest AUROC score in general. However, for the accuracy, Local Outlier Factor has the highest score. The results and the approach used in this research should be noted that this is one of the few attempts to practice machine learning models on MBC topic especially using PLCO trials. Analysis of PLCO dataset can give new pattern that can be used as early prediction of disease, drug invention, clinical practice, personalized medicine and so on. Most importantly, the majority of PLCO dataset is based on various practices of questionaire, such as Baseline Questionnaire, which would largely decrease the cost of detection. For further study, the improvement on model's parameter tuning and data preprocessing is suggested.

Acknowledgments

The author thank the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. The author would also like to express the gratitude to research supervisor and teacher. Without their guidance and support this research article would not have been possible.

References

Al-Masri, A. (2019). How does k-means clustering in machine learning work? *TowardsDataScience*. https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0

Breast cancer wisconsin (diagnostic) data set. (1995). *UCI Machine Learning Repository*.

Cardoso, F. (2017). Characterization of male breast cancer: Results of the eortc 10085/tbcrc/big/nabcg international male breast cancer program. *ScienceDirect*. https://doi.org/10.1093/annonc/mdx651

Choudhury, A. (Ed.). (2021, January 14). *Top xgboost interview questions for data scientists*. Retrieved August 31, 2021, from https://analyticsindiamag.com/top-xgboost-interview-questions-for-data-scientists/

Decision tree learning pros and cons. (n.d.). *Orelly*. https://www.oreilly.com/library/view/machine-learning-with/9781787121515/697c4c5f-1109-4058-8938-d01482389ce3.xhtml

Doshi, N. (2019). Spectral clustering. *Towards Data Science*. https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7

Gao, Y. (2019). Breast cancer screening in high-risk men: A 12-year longitudinal observational study of male breast imaging utilization and outcomes. *Radiology*. https://doi.org/10.1148/radiol.2019190971

Goonewardana, H. (2019). PCA: Application in machine learning. *Apprentice Journal*. https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db

Hill, T. D. (2005). Comparison of male and female breast cancer incidence trends, tumor characteristics, and survival. *ScienceDirect*. https://www.sciencedirect.com/science/article/abs/pii/S1047279705000128?via%3Dihub

Innab, R. (2019, October 31). *Why do decision trees have a tendency to overfit to the training set?* [Online forum post]. Quora. https://www.quora.com/Why-do-decision-trees-have-a-tendency-to-overfit-to-the-training-set

Karatsalos, C. (2018, March 27). *What is the time complexity of spectral clustering and why is it so?* [Online forum post]. StackExchange. https://stats.stackexchange.com/questions/348512/what-is-the-time-complexity-of-spectral-clustering-and-why-is-it-so

Kunanbaeva, A. (2019). What is ROC AUC and how to visualize it in python. *Medium*. https://medium.com/@kunanba/what-is-roc-auc-and-how-to-visualize-it-in-python-f35708206663

M, S., & Radhika, S. (2020). *Machine learning techniques for prediction from various breast cancer datasets*. IEEE. https://sci-hub.st/https://ieeexplore.ieee.org/abstract/document/9167657/

Male breast cancer. (2020). *National Breast Cancer*. https://www.nationalbreastcancer.org/male-breast-cancer

Markman, M. (2021). BRCA1 and brca2. *Cancer Treatment Center of America*. https://www.cancercenter.com/cancer-types/breast-cancer/risk-factors/brca1-and-brca2

Narkhede, S. (2018). Understanding logistic regression. *Towards Data Science*. https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102

Omene, C. (2010). Chapter 42 - the differences between male and female breast cancer. *ScienceDirecta*. https://doi.org/10.1016/B978-0-12-374271-1.00042-3

Prado, K. (2017). How DBSCAN works and why should we use it? *TowadsDataScience*. https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

Prostate cancer screening results from the prostate, lung, colorectal, and ovarian cancer randomized screening trial: Questions and answers. (2009, March 19). Retrieved August 3, 2021, from https://www.cancer.gov/types/prostate/research/plco-screening-results-qa#:~:text=Cancer%20Screening%20Trial%3F-,The%20Prostate%2C%20Lung%2C%20Colorectal%2C%20and%20Ovarian%20(PLCO),%2C%20colorectal%2C%20and%20ovarian%20cancer.

Regularization (mathematics). (n.d.). Wikipedia. Retrieved August 31, 2021, from https://en.wikipedia.org/wiki/Regularization_(mathematics)

Sasco, A. (1993). Review article: Epidemiology of male breast cancer. A meta-analysis of published case-control studies and discussion of selected aetiological factors. *International Journal of Cancer*. https://onlinelibrary.wiley.com/doi/10.1002/ijc.2910530403

Scholkopf, B. (2000). Support vector method for novelty detection. *MIT Press*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.675.575&rep=rep1&type=pdf

Sharma, A. (2020). How to master the popular dbscan clustering algorithm for machine learning. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

Sklearn.cluster.DBSCAN. (n.d.). Scikit-learn. Retrieved August 31, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Sklearn.ensemble.RandomForestClassifier. (n.d.). Sklearn. Retrieved August 31, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Verma, P. (2020). Isolation forest algorithm for anomaly detection. *Heatbeat*. https://heartbeat.fritz.ai/isolation-forest-algorithm-for-anomaly-detection-2a4abd347a5

Vermeulen, M. A. (2017). Pathological characterisation of male breast cancer: Results of the eortc 10085/tbcrc/big/nabcg international male breast cancer program. *European Journal of Cancer*. https://doi.org/10.1016/j.ejca.2017.01.034

Wening, P. (2018). Local outlier factor for anomaly detection. *TowardsDataScience*. https://towardsdatascience.com/local-outlier-factor-for-anomaly-detection-cc0c770d2ebe

Yalaza, M. (2016). Male breast cancer. *US National Library of Medicine National Institutes of Health*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5351429/#b3-jbh-12-1-1

Yan, D. (2009). Fast approximate spectral clustering. *Association for Computing Machinery*. https://doi.org/10.1145/1557019.1557118