Project Overview

Cancer is an abnormal growth of cells. There are more than 100 types of cancer, including breast cancer, skin cancer, lung cancer and prostate cancer("WebMD," n.d.). It is estimated that there is a lag period of 20 years between the development of the first cancer cell and the onset of end-stage metastatic disease(Alberts & Hess, 2014). Cancer that is diagnosed at an early stage, before it has had the chance to become too big or spread is more likely to be treated successfully. For example, in breast cancer 90% of the women diagnosed at the early stage survive their disease for at least 5 years compared to around 15% of women diagnosed with the most advanced stage of disease. ("Cancer Research UK," n.d.). Unfortunately, effective screening tests for early detection do not exist for many cancers. And, for cancers for which there are widely used screening tests, many of the tests have not proven effective in reducing cancer mortality. While there is a clear benefit to screening it has its downside too in that there is a risk of over-diagnosis and overtreatment—the diagnosis and treatment of cancers that would not threaten life or cause symptoms. Over-diagnosis and overtreatment expose patients unnecessarily to the potential physical harms of unneeded and often invasive diagnostic tests and treatment, as well as to the psychological stresses associated with a cancer diagnosis. Therefore, non-invasive tests that identify biomarkers through routine blood analysis can be a powerful tool for cancer diagnosis.

Problem Statement

Breast cancer screening is an important strategy to allow for early detection and ensure a greater probability of having a good outcome in treatment. The research goal of this project is to utilize a publicly available dataset to identify and validate a generic blood screening tests for early breast cancer detection in the general population. In this project, we will assess how models based on the input variables collected during routine blood analysis may be used to predict the presence of breast cancer. Furthermore, the applicability of the predictive model to larger and more heterogeneous populations can be subsequently assessed. Since this is a bi-class classification problem, supervised learning algorithms that classify observations based on a set of input variables will be

used to assess the suitability of the algorithm for predicting breast cancer in the given dataset. The goal of this project is two-fold:

1. Reproduce the results obtained by the authors of the publication
2. Improve the accuracy of the model: This is relevant because the authors mention that that the focus of their work was not in optimizing the accuracy of the classifiers, but rather assessing the predictive value of the set of predictors. Therefore, with the data used in this study to build the prediction models, it is possible to try to achieve better diagnosis accuracy. This project will explore different classifiers or ensemble methods, the amount of data allocated to the training or test sets may be altered.

Metrics

The F-score is a useful metric to compare classifiers. It is computed using the harmonic mean of precision and recall. Precision is the ability of the classifier to precisely identify the positives while recall (or sensitivity) is the true positive rate. The F-score can range from 0 to 1, with 1 being the best possible F-score. The F-beta-score is a useful metric that considers both precision and recall with an additional parameter, beta that serves as a weight of precision in the harmonic mean. This can be useful parameter in a disease diagnostic because we would like to favor recall over precision since we value a highly sensitive diagnosis. The formula for F-beta score is as follows:

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall})$$

where precision = True Positives / (True Positives + False Positives)

recall = True Positives / (True Positives + False Negatives)

Analysis

Data Exploration

The dataset used for this project was obtained from the UCI machine learning dataset (Breast Cancer Coimbra Data Set, n.d.). The data was made available publically by a research group at the Faculty of Medicine, Coimbra, Portugal (Patrício et al., 2018). The original study proposed a model for breast cancer detection based on biomarkers. Clinical,

demographic and anthropometric data were collected for a total of 64 women with breast cancer and 52 healthy volunteers. The putative biomarkers assessed as part of the study were Glucose, Resistin, Age, Body Mass Index (BMI), Homeostasis Model Assessment (HOMA), Leptin, Insulin, Adiponectin, Monocyte Chemoattractant Protein 1 (MCP-1). Leptin, Insulin, Adiponectin and Resistin are important hormone proteins that are involved in the regulation of inflammation, glucose metabolism and/or insulin resistance. MCP-1 is a protein, a cytokine, and an important mediator of inflammation. HOMA is a method used to quantify insulin resistance.

In summary, there are 9 quantitative predictors and a binary dependent variable, indicating the presence or absence of breast cancer. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer. The mean values for patients and control is shown in Table 1. Statistical differences between the two categories was evaluated using the Mann-Whitney test and the p-value of this test is shown in Table 1. There was a statistically significant difference between the two groups (p-value < 0.05) for Glucose, HOMA, Insulin and Resistin. The values for these metabolic parameters were consistently higher for the patients. There was no significant differences between the patients and controls for the other parameters.

Table 1 Descriptive statistics and assessment of differences between patients and controls.

Variable	Control	Patient	p-value
Adiponectin	10.33	10.06	0.383
Age	58.08	56.67	0.239
BMI	28.32	26.98	0.101
Glucose	88.23	105.56	< 0.001
HOMA	1.55	3.62	0.001
Insulin	6.93	12.51	0.013
Leptin	26.64	26.6	0.475
MCP.1	499.73	563.02	0.252
Resistin	11.61	17.25	0.001

Exploratory Visualization

Figure 1 presents the box plot which is a visual representation of the five number summary (minimum, first quartile, median, third quartile and the maximum value) for patients and controls for each of the variables. Even though the median value for Age differs quite significantly between patients and controls, it is interesting to note that there is no statistical difference between the two.

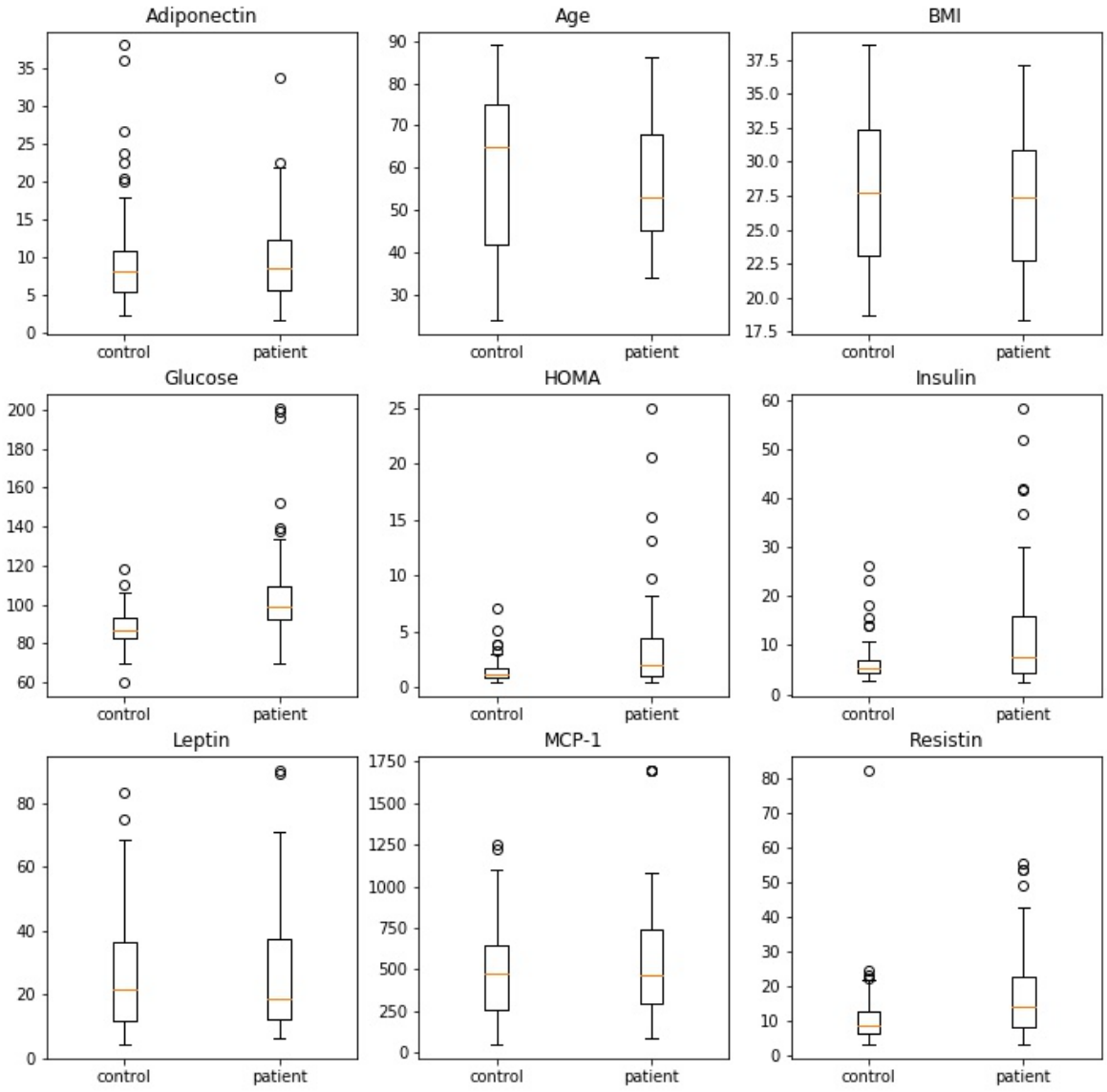

Figure 1 Boxplot depicting the distribution of values for each variable between patients and controls.

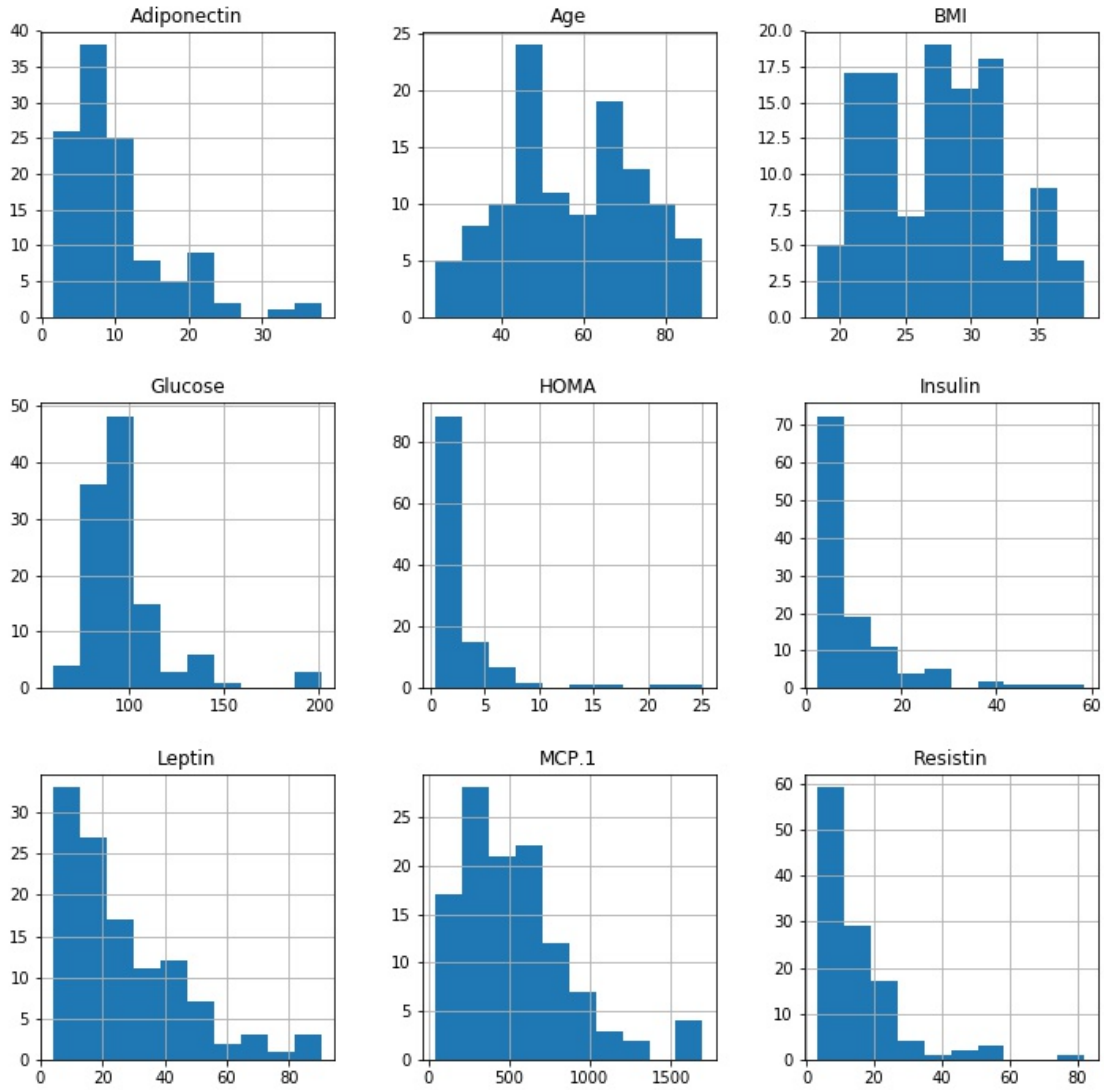

Figure 2 Histogram of the raw data for the nine variables present in the dataset.

Figure 2 shows the distribution of values for each variable. The following variables show a slight skew and may have to be transformed appropriately in order to prevent any performance issues: Insulin, Adiponectin, Glucose, HOMA, Resistin, Leptin and MCP-1. Furthermore, the variables have different ranges that need to be scaled so that each feature is treated equally by the supervised learners.

Algorithms and Techniques

Supervised learning models such as Support Vector Machines (SVM), Random Forest or AdaBoost Classifiers and XGBoost classifiers will be employed to assess the utility of variables measured during routine blood analysis to predict the presence of breast cancer. Specifically, we propose to predict the presence of breast cancer based on the variables: Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP-1.

SVM has been successfully used for cancer classification by several groups. The mathematical formulation of SVM is as follows:

Given a labelled training dataset:

$$(x_1, y_1), \dots, (x_n, y_n)\ , x_i \in R^d \ and\ y_i \in (-1,1)$$

where x_i is a feature vector representation and y_i the class label (-1 or 1) of a training compound i.

The optimal hyperplane can then be defined as:

$$WX^T + b = 0$$

where w is the weight vector, x is the input feature vector, and b is the bias.

The w and b would satisfy the following inequalities for all the elements of the training set:

$$wx_i{}^T + b \geq +1\ if\ y_i = 1$$

$$wx_i{}^T + b \leq -1\ if\ y_i = -1$$

The objective of training an SVM model is to find the w and b so that the hyperplane separates the data and maximizes the margin:

$$1/_{||w||^2}$$

Vectors x_i for which $|y_i|(wx_i{}^T + b) = 1$ will be termed support vector.

Similarly, AdaBoost is a type of "Ensemble Learning" where multiple "weak classifiers" are employed to build a single "strong classifier". A weak classifier is simply a classifier that performs poorly, but performs better than random guessing. AdaBoost works by choosing a base algorithm, for example, Decision tree, and iteratively improving it by accounting for the incorrectly classified examples in the training set. At the outset, equal weight is assigned to all the training examples and a base algorithm is chosen. At each

step of iteration, the base algorithm is applied to the training set and AdaBoost assigns an increased "weight" to the incorrectly classified examples. This in turn determines the probability that each example will appear in the training set. Thus, after n iterations, each time applying base classifier on the training set with updated weights, the final model is determined to be the weighted sum of the n "weak" classifiers.

XGBoost is a tree learning algorithm and is an implementation of the gradient-boosted decision trees. Assembly algorithms create and combine a high number of individually weak but complementary classifiers, to produce a robust estimator. A decision tree allows making prediction on an output variable based on a series of rules arranged in a tree-like structure. They consist of a series of split points, the nodes, in terms of the value of an input feature. The last node is a leaf and gives us the specific value of the output variable. Tree learning algorithms do not require linear features or linear interactions between features. Moreover, XGBoost, has two major improvements: (a) speeding up the tree construction and (b) proposing a new distributed algorithm for tree searching (Torlay, Perrone-Bertolotti, Thomas, & Baciu, 2017).

Benchmark Model

The solution will be compared with the results obtained by the publication of Patricio et al (Patrício et al., 2018). This group was responsible for collecting the data for the 116 participants and building predictive models to aid in the diagnosis of breast cancer. The group employed machine learning models such as Support Vector Machines, Random Forest, Logistic Regression etc. using Glucose, Resistin, Age and BMI to predict the presence of breast cancer in women. The resulting models were assessed with a Monte Carlo Cross-Validation approach to determine 95% confidence intervals for the sensitivity, specificity and AUC of the models. They obtained sensitivity ranging between 82 and 88% and specificity ranging between 85 and 90% using a SVM model. The 95% confidence interval for the AUC was [0.87, 0.91]. The paper concluded that metabolic parameters may be a powerful tool for a cheap and effective biomarker of breast cancer.

Methodology

The proposed workflow for the prediction of breast cancer from the available variables in the dataset is shown in **Figure 3**.

Data Preprocessing

There are 9 quantitative variables in the dataset. At the outset, these variables were investigated for skew and it was determined that Insulin, Adiponectin, Glucose, HOMA, Resistin, Leptin and MCP-1 are slightly skewed. Therefore, logarithmic transformation was applied to reduce the range of values caused by outliers. Furthermore, the numerical features were scaled using 'MinMaxScaler' available in 'preprocessing' package in sklearn. This was done to ensure that each feature is treated equally when applying supervised learners. Functions in the "preprocessing" package in sklearn were utilized to scale and/or normalize the data.

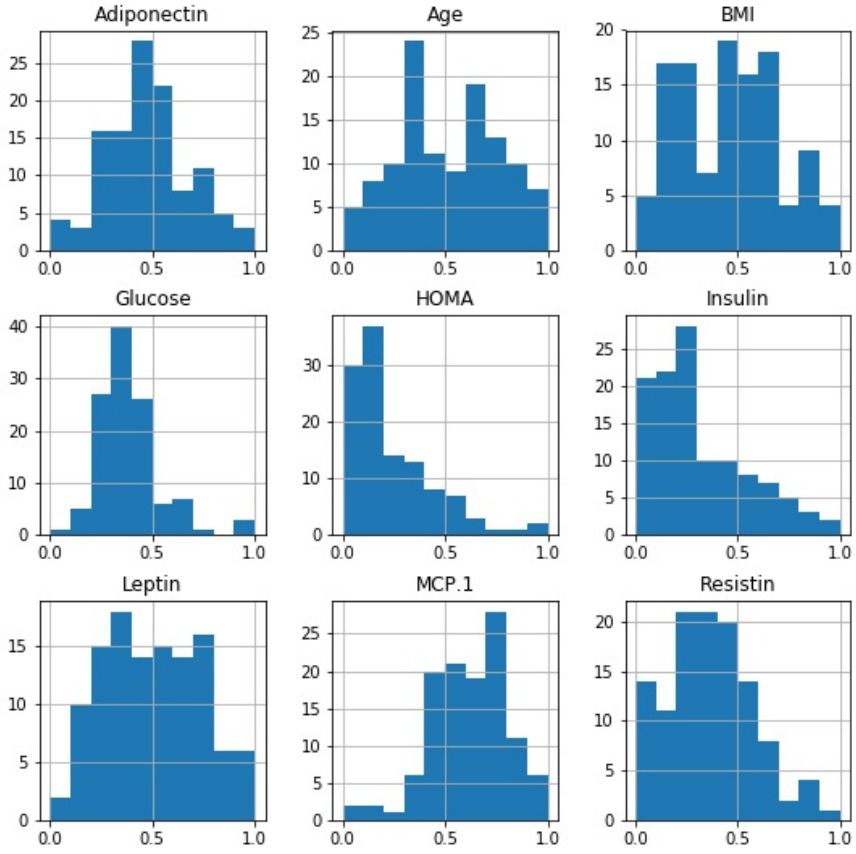

Figure 3 Histogram of the variables after the relevant transformations have been applied to the input data.

Implementation

Figure 4 depicts the steps that were utilized to implement the model. The following sections will describe each of the step in greater detail:

Shuffle and split data:

The data was split into training and test sets. 80% of the data was used for training and 20% for testing. The function "train_test_split" from "cross_validation" package was used for performing this split.

Evaluation of Model Performance:

As has been mentioned in an earlier section Fbeta-score was employed to evaluate model performance. ROC curves were also plotted to serve as a comparison to the benchmark model. These functions are available in the "metrics" package in sklearn.

Supervised Learning Models:

The following supervised learning models were employed to build the predictive model:

- Gaussian Naïve Bayes (GaussianNB)
- Ensembl methods
    - AdaBoost
    - Random Forest
- Support Vector Machines
- Logistic regression
- XGBoost

XGBoost classifier was the best model based on the chosen performance metric and was therefore taken up for further exploration. The supervised learning models mentioned above are available in scikit-learn with the exception of XGBoost which was downloaded and imported separately.

Refinement

Model Tuning:

The XGBoost model shortlisted from the above step was further fine-tuned for performance using the "gridsearch" package and the "GridSearchCV" function. A KFold cross-validation generator with 5 folds was applied towards the splitting strategy. The

parameters that were fine-tuned during this process were the 'learning_rate' and 'max_depth' for the XGBClassifier that was chosen as the final supervised learning model based on its performance during initial model training and evaluation.

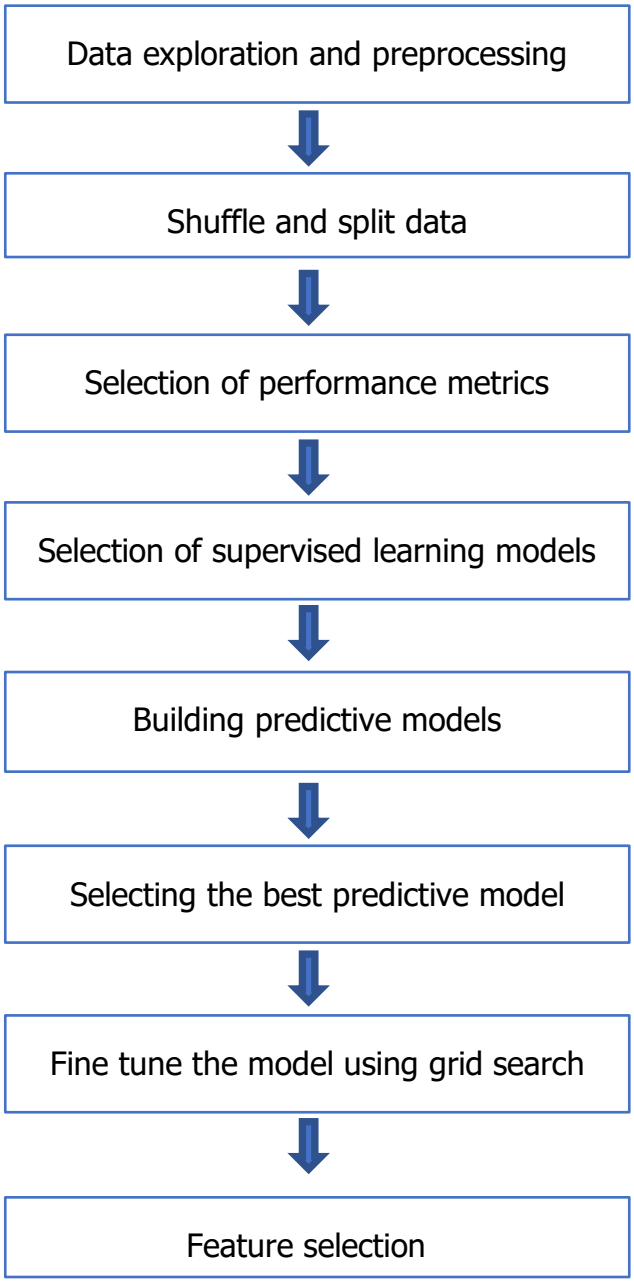

Figure 4 Schematic of the workflow that was used to build a predictive model of breast cancer based on results from a routine blood screening.

Feature selection:

Several supervised learning models have a method called "feature_importance" that can be utilized to extract the weights that signify the relative importance of the feature in model prediction. The top few features were extracted from the dataset and the effect of feature selection on the predictive ability of the model was evaluated and visually depicted using the "feature_plot" function in the python script called "visuals.py".

Results

Model Evaluation and Validation

The goal of this project was to develop a model that could accurately predict the presence of breast cancer based on a set of anthropometric and metabolic parameters. Among the supervised learning algorithms that were evaluated XGBClassifier performed the best during the model evaluation phase with an Fbeta score of 0.928. The shortlisted model was further fine-tuned using the 'GridSearchCV' function available in sklearn. The optimized model was tested against the test set that was set aside for this purpose. The optimized model returned a fbeta_score of 0.946. Table 2 depicts the confusion matrix that was obtained using this classifier to predict the label for the 24 test samples. This indicates that the model worked well for unseen data. More importantly, the false negative rate was very low as indicated in Table 2. The latter is important in the context of the diagnostic model that we are trying to develop. Furthermore, the hyperparameter tuning with 'GridSearchCV' applies a KFold cross-validation for obtaining parameter estimates which helps in generating a more robust model.

Table 2 Confusion matrix (or contingency table) obtained for the best classifier (learning_rate=0.1 and max_depth=5) using XGBClassifier for the 24 samples present in the test set.

True class \ Prediction class	Control	Patient
Control	9	1
Patient	0	14

Furthermore, the process of feature selection identified 5 features that contributed most to the prediction variable as shown in Figure 5. While the smaller variable set did not improve the accuracy of the model, it is still desirable because simpler models are easier to explain and it also potentially reduces overfitting. Interestingly, the benchmark model identified the same subset of variables with the exception of HOMA. The best combination of sensitivity and specificity, in the benchmark study, were obtained for a model that contained Glucose, Age, BMI and Resistin. The results obtained in this project can be trusted because they conform to the general intuition in the field.

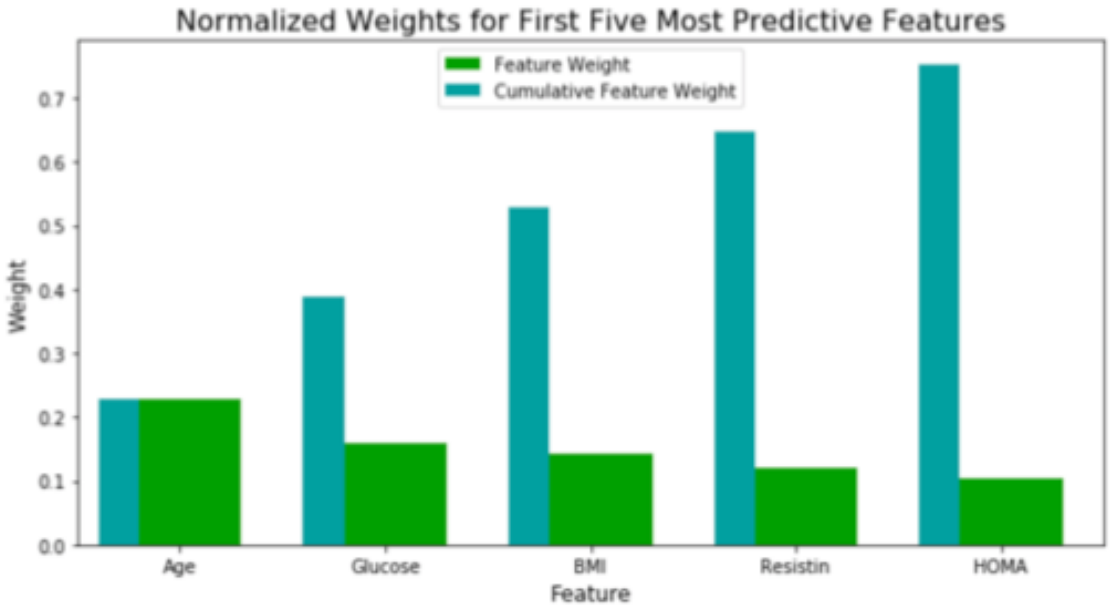

Figure 5 Normalized weights of the most predictive features

Justification

The result obtained by applying the supervised learning model developed in this project is similar to the benchmark model (Patrício et al., 2018). In particular, the ROC curve obtained for our model, as shown in Figure 6, is similar to the 'best' curves provided by the original publication. By the standards of modern machine learning datasets, our dataset had a small sample size which is a limitation. However, the result of this project shows that non-invasive and inexpensive tests are promising and in the long run they can at least complement breast mammograms if not serve as an alternative.

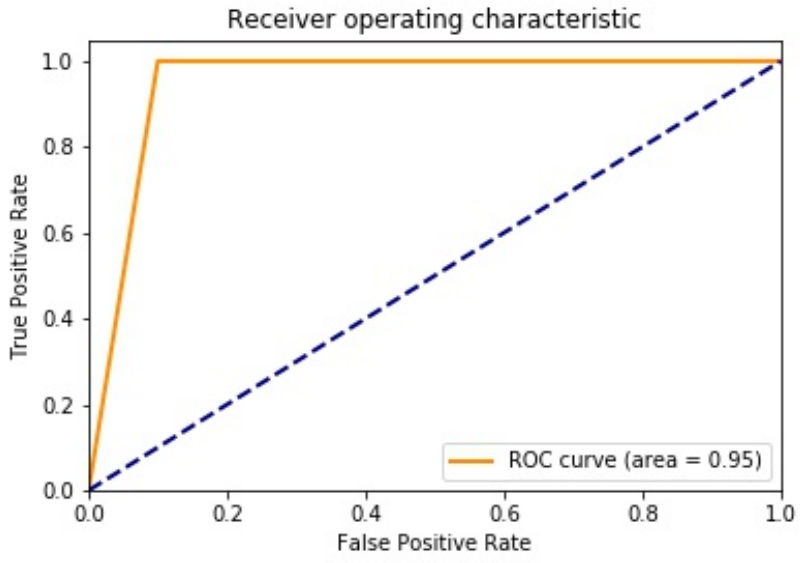

Figure 6 ROC curve corresponding to the XGBClassifier for the model with all the variables in the dataset.

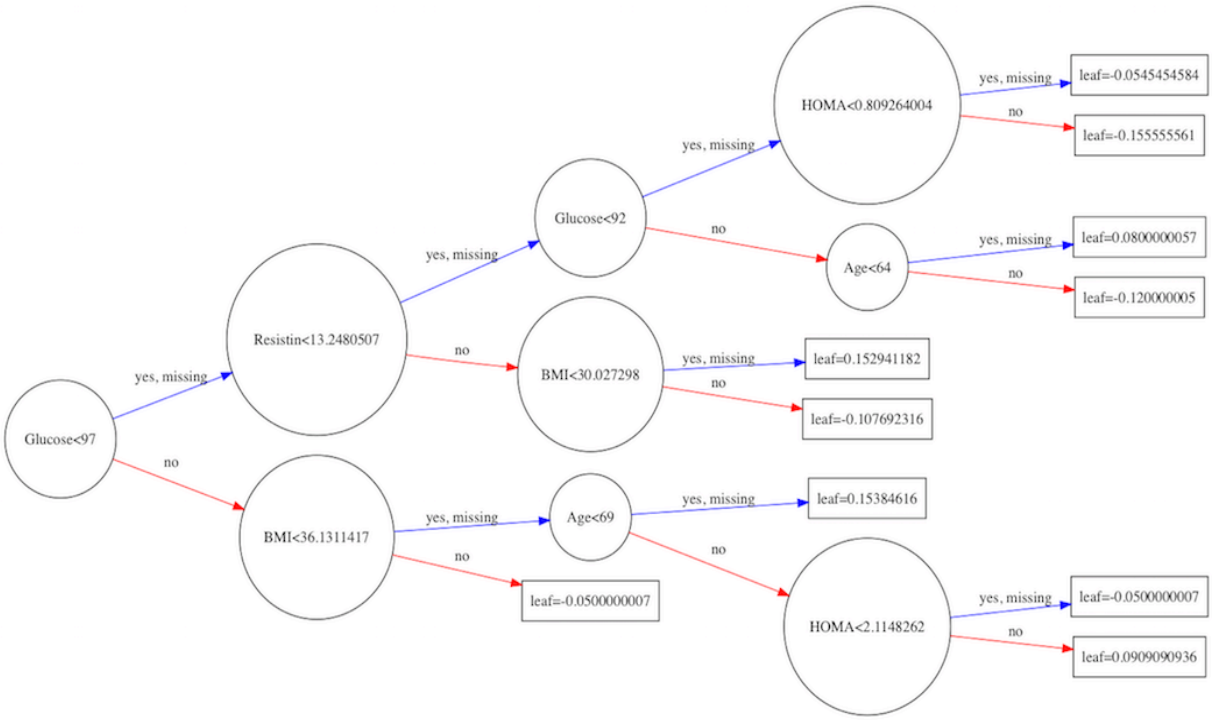

Figure 7 Visualizing Gradient Boosting Decision Trees for the selected five features

Conclusion

Free-Form Visualization

The XGBoost Python API provides a function for plotting decision trees within a trained XGBoost model. The function 'plot_tree' creates a plot of the decision tree in the model showing the features and feature values for each split as well as the output leaf nodes. Interestingly, XGBoost performs identically for raw as well as transformed data values and is immune to issues surrounding skew or scaling. This is useful because a tree that is based on a model trained on raw values provides us with an idea of threshold values for variables that potentially determine the disease state as shown in Figure 7. For example, traversing the tree at the bottom of the figure it can be determined that a high Glucose value coupled with a high BMI is a reasonable predictor of disease state, this is in line with literature where these parameters are seen as important predisposing factors for cancer. On the other hand, even with relatively lower Glucose and BMI, a higher Resistin level is a predisposing factor for cancer as shown in the figure.

Reflection

The final model fits with my expectations for the problem and also aligns well with the ideas proposed in the original publication as well as the literature in the field. I find it extremely interesting that such simple tests can be utilized as a diagnostic for breast cancer. The small sample size was a limitation in my opinion, however, this seems to be the norm in healthcare data where the cost of data acquisition is very high. In the future, with more samples and additional metabolic parameters, such models can be an attractive alternative to the more invasive and expensive diagnostics for breast cancer. In addition, with sequencing data becoming increasingly available, mutation information can be added as parameters to improve the predictability of supervised learning models.

Improvement

I believe that the algorithms used in this project were ideally suited for the dataset in hand. If the sample size was much larger, deep learning techniques could have been employed to develop a diagnostic for breast cancer.

Reference

Alberts, D., & Hess, L. M. (2014). *Fundamentals of Cancer Prevention*. Retrieved from https://books.google.com/books?isbn=364238983X

Breast Cancer Coimbra Data Set. (n.d.). UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Cancer Research UK. (n.d.). Retrieved July 20, 2018, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survival#By

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1), 29. https://doi.org/10.1186/s12885-017-3877-1

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, *4*(3), 159–169. https://doi.org/10.1007/s40708-017-0065-7

WebMD. (n.d.). Retrieved July 20, 2018, from https://www.webmd.com/cancer/default.htm