Background

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.

Datasets and inputs

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, BMI, HOMA, Leptin, Insulin, Adiponectin, MCP-1.

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 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 or data imputation techniques may be used to deal with cases that were excluded by the original publication due to missing data.

Solution Statement

Supervised learning models such as Support Vector Machines (SVM), Random Forest or AdaBoost 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:

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

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.

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.

Evaluation 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)

Project Design

The proposed workflow for the prediction of breast cancer from the available variables in the dataset is shown in **Figure 1**.

Data exploration:

There are 9 quantitative variables in the dataset. At the outset, these variables will be investigated for skew and appropriate transformation strategy such as logarithmic transformation will be applied to reduce the range of values caused by outliers. Furthermore, scaling of the numerical features will be performed to ensure that each feature is treated equally when applying supervised learners. Functions in the "preprocessing" package in sklearn will be utilized to scale and/or normalize the data.

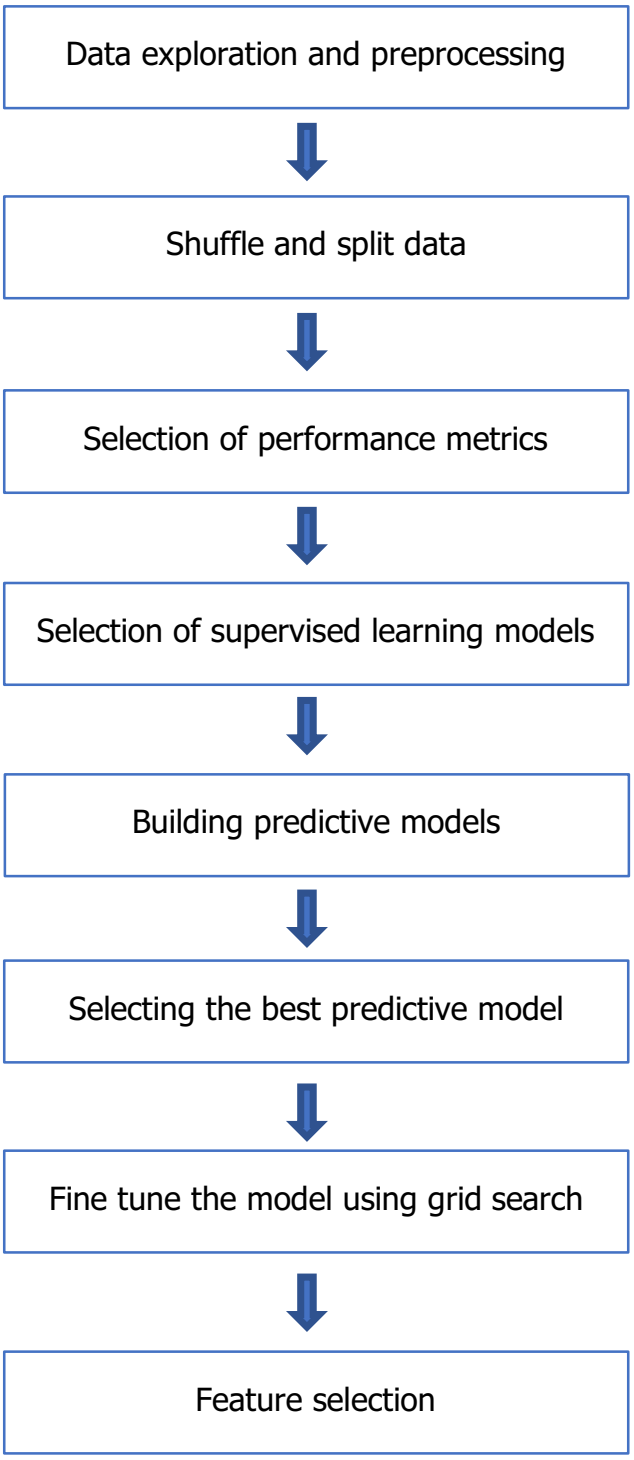

Figure 1 *Schematic of the proposed workflow to build a predictive model of breast cancer based on results from a routine blood screening*

Shuffle and split data:

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

Evaluation of Model Performance:

As has been mentioned in an earlier section various metrics but with a special emphasis on F-score will be employed to evaluate model performance. These functions are available in the "metrics" package in sklearn.

Supervised Learning Models:

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

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

The best model based on the chosen performance metric will be shortlisted for further exploration. The supervised learning models mentioned above are available in scikit-learn.

Model Tuning:

The model shortlisted from the above step will be further fine-tuned for performance using the "gridsearch" package and the "GridSearchCV" function.

Feature selection:

A supervised learning model that has the "feature_importance" attribute available for it (for example, AdaBoostClassifier) will be utilized to extract the weights that signify the relative importance of the feature in model prediction. The top features will be extracted from the dataset and the effect of feature selection on the predictive ability of the model will be evaluated.

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

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