Predicting Breast Cancer Within Cuban Women

Muhammad Razi Mahardika  
Computer Science Department  
School of Computer Science  
Bina Nusantara University  
Jakarta, Indonesia  
[muhammad.mahardika004@binus.ac.id](mailto:muhammad.mahardika004@binus.ac.id)

Marcell Kurniawan Sutanto  
[DEPARTMENT]  
[AFFILIATION]  
Bina Nusantara University

Jakarta, Indonesia  
[marcell.sutanto@binus.ac.id](mailto:marcell.sutanto@binus.ac.id)

***Abstract*—**Text Here

***Keywords***—Breast Cancer, Cuba, Binary Classification, Random Forest, Logistic Regression, KNN, SVM

1. Introduction

TEXT HERE

1. Literature Review

TEXT HERE

1. Methodology

The methodology section can be split into the following sections described by the flowchart below.

A diagram of a process flow

Description automatically generated

Fig. 1. Methodology Flowchart

The methodology section will start by obtaining the dataset. Once the dataset has been obtained, the data will be analysed for any issues that may affect the machine learning models. If any issues are identified, it will be resolved at the data preprocessing section. Once the dataset has been pre-processed, feature selection will be done on the dataset using several feature selection methods to identify the most significant feature for predicting the target label. Once the features are selected, each model chosen will be experimented on using the dataset that only has the chosen features and the dataset that still has all its features. This means that there will be a total of eight experiments being done. Once all eight experiments have been conducted, each of the experiment’s performance will be evaluated using several evaluation metrics.

*A. Dataset Gathering*

The dataset that this research uses is the “Cuban Breast Cancer” dataset. The dataset contains data of patients from Cuba that consulted with a hospital in Havana for breast cancer. The dataset contains 23 variables used to represent the risk factors of breast cancer and 1697 data instances. The target label that will be used in this dataset is the “cancer” variable that includes boolean values “yes” or “No”.

*B. Dataset Preprocessing*

Preprocessing the dataset is a crucial step in machine learning that can help increase the machine learning model’s performance and avoid having issues when training the machine learning model.

The dataset contains several issues that can hinder the machine learning model process. First, a lot of data instances still contain missing values and will need to be handled with. Second, some columns contain unique values that contains unnecessary suffixes such as “1 month” or “2 months” where the suffix “month(s)” can be removed so that the column can be turned into an integer datatype. Third, some columns contain multiple unique values that all represents the same meaning such as the string value “No” and the integer value 0 in some columns. Lastly, some data instances combine multiple unique values of a column to a single data instance. How the dataset combines multiple values is by appending additional values at the end of the previous string value separated by a slash (/). One example of this is in one of the columns, a string value “Mother/Sister” combines the unique value “Mother” and “Sister” into one new unique value and assigns it to a data instance. Because every new unique combination of values created in this manner can cause the machine learning model to consider more unique values then they are supposed to, a new way to represent data instances with multiple values needs to be considered.

With the issues within the dataset identified, specific preprocessing methods can be utilized to resolve the issues. The preprocessing steps that were taken are as follows:

1. Replacing unique values that represents the same meaning with a uniformed value using replace functions.
2. Removing unnecessary suffixes using string replace function accompanied by regex syntax.
3. Creating new binary columns for columns that can contain a combination of values and assigning a “True” or “False” to the binary column by checking the string value for each unique value.
4. Handling missing values by:
   1. Removing the data instances with missing values if there aren’t much missing values to handle in the column.
   2. Replacing missing values with the median of the column if the value of the data isn’t significant.
   3. Imputing missing values with the results of the random forest regressor if the value of the data is significant.
5. Encoding categorical data into integers using one hot encoding or label encoding

*C. Feature Selection*

Each model being used in this experiment will be tested on two types of datasets. The first dataset will have all the features available in the original dataset, and the second will have only a few chosen features that were considered after seeing the results of three feature selection methods. The three feature selection methods that will be utilized are as follows:

1. Tree-based feature importance using random forest classifier.
2. Recursive feature elimination using logistic regression.
3. Select K-best using the chi2 score.

The results of each of the feature selection methods vary from each other. For this research, the most consistent top performing features across all methods will be chosen to be used for the second dataset. The selected features are: “biopsies”, “histologicalclass”, “consumed\_alcohol”, “menopause”, and “is\_sad”.

*D. Experimental Setup*

Binary classification is one of the many tasks that can be achieved in machine learning. The goal of binary classification is to classify each data instance to one of two possible classes. This research is trying to classify its data instances to identify if a patient is suffering from breast cancer or not.

Several machine learning models can be used to perform this task. Each machine learning model has its own algorithms, advantages, and limitations. This research will be using four models and making a comparison between them to find which model is best suited for this task and dataset.

The first model is the random forest classifier. The random forest classifier works by creating many different decision trees which are trained on different parts of the dataset and conducts voting between all the trees to find the most likely predictions. This model is known to provide accurate predictions with large datasets and helps reduce the risks of overfitting. However, this model is limited due to the high computational cost to run it.

The second model is the logistic regression model. The logistic regression model is usually used for binary classification tasks. This model uses the sigmoid function that uses variables as inputs to produce a float value between 0 and 1. If the resulting value is closer to 0 then the data instance will be assigned to class 0, and such is the case for class 1. The logistic regression model is known to be extremely fast when classifying new data instances. The model is limited in its ability to capture complex relations within the dataset.

The third model is the K-nearest neighbour (KNN) model. The KNN model is a supervised learning model that can be used for both classification and regression tasks. This model works by considering the k amount of the nearest datapoints to the new datapoint that is trying to be classified and assigns a class to the new datapoint depending on what the majority population is among its neighbouring datapoints. KNN is widely used in machine learning due to its versatility, simplicity, and ease of use. However, KNN is prone to overfitting, does not scale well, and does not perform well when the dataset used has a high dimensionality.

The fourth model is the support vector machine (SVM) model, specifically the support vector classification (SVC) version of the model. SVM is widely used for linear and nonlinear classification tasks as well as regression tasks. SVM works by finding a maximum separating hyperplane between each class available in the target label. The separating hyperplane is created by trying to find the maximizing distance between the nearest datapoint of each class and the hyperplane itself. SVM models are known to perform well in high dimensional datasets and are memory efficient when performing their task due to their focus on support vectors. The model is limited due to its slow training time, sensitivity to noise, and feature scaling sensitivity.

All the above models will be trained, tested, and evaluated on the two types of datasets, the normal dataset and the dataset with only the most significant features. With four models, there is a total of eight experiments that will be conducted. The experiments are as follows:

1. Experiment list

|  |  |  |
| --- | --- | --- |
| Experiment | Model | Dataset |
| Experiment 1 | Random Forest | Normal |
| Experiment 2 | Random Forest | Best Features |
| Experiment 3 | Logistic Regression | Normal |
| Experiment 4 | Logistic Regression | Best Features |
| Experiment 5 | KNN | Normal |
| Experiment 6 | KNN | Best Features |
| Experiment 7 | SVM SVC | Normal |
| Experiment 8 | SVM SVC | Best Features |

*E. Evaluation Metrics*

To identify if the models performed well, an evaluation needs to be performed using evaluation metrics. This research will be using four scoring metrics to help evaluate each experiment. These metrics include accuracy score, precision score, recall score, and f1-score.

1. Results and Discussion

TEXT HERE

1. Conclusion

TEXT HERE

References

1. REFERENCES HERE