Predicting Breast Cancer Based on Anthropometric Data

Machine Learning Foundation

Dialog Data Science Academy

By

Niroshan Balasuriya

Supervisor

Dr. Sumudu Thennakoon

Introduction

Machine Learning provides statistical, probabilistic and optimization techniques in detecting patterns from different data sets. This is one of the best approaches that can be used in detecting serious medical conditions such as cancer. Early detection, and prediction of a cancer is very important to save a life of a patient. Modern life style patterns are meant to cause a major impact on cancers. Since recent past machine learning is frequently adopted to detect and predict cancers.

Breast cancer is a common cancer type among women. In 2020, 2.3 million women have diagnosed with breast cancer and 685,000 deaths globally. Breast cancer is a type of cancer which can be cured if found early. Early detection is the key here. There are advance diagnosis techniques available with the modern medical technologies. But still the fatality rate is high. In this research, it is mainly focused on the predicting a breast cancer based on anthropometric data and parameters which can be easily collected from routine blood analysis. Since these data can be collected with routine medical checkups, it gives an added advantage to detect such a condition using machine learning.

Data Set

The selected data set is a collection of data that indicating the presence or absence of breast cancer. The data set was picked from the UCI Machine Learning Repository.

- There are 10 predictors, all quantitative, and a binary dependent variable.
- The predictors are anthropometric data and parameters which can be gathered in routine blood analysis.
- Number of Instances = 116.
- Number of Attributes = 10.
- 44% of the given data set consists of the records from healthy people and 56% are patients. Therefore, the data set distribution can be considered as an acceptable one.

X_Variables

- Age (years)
- BMI (kg/m2)
- Glucose (mg/dL)
- Insulin (µU/mL)
- HOMA
- Leptin (ng/mL)
- Adiponectin (µg/mL)

- Resistin (ng/mL)
- MCP-1(pg/dL)

Dependent Variable:

- Classification
  Labels:
  1 = Healthy Patients
  2 = Patients

Methodology

This is a binary classification problem. Two classes denote “1” for healthy patients and “2” denotes the patients.

Exploratory Data Analysis and Data Pre-processing

Data types of the data set is as follows. All variables are quantitative and no null values.

#	Column	Non-Null Count	Dtype
0	Age	116 non-null	int64
1	BMI	116 non-null	float64
2	Glucose	116 non-null	int64
3	Insulin	116 non-null	float64
4	HOMA	116 non-null	float64
5	Leptin	116 non-null	float64
6	Adiponectin	116 non-null	float64
7	Resistin	116 non-null	float64
8	MCP.1	116 non-null	float64
9	Classification	116 non-null	int64

Figure 1.0

Further the data set didn’t contain any missing values as per the below image.

```
print(data.isna().all())
```

```
Age                False
BMI                False
Glucose            False
Insulin            False
HOMA               False
Leptin             False
Adiponectin        False
Resistin           False
MCP.1              False
Classification     False
dtype: bool
```

Figure 2.0

Following correlation matrix shows the pair wise correlation between the variables.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
Age	1.00	0.01	0.23	0.03	0.13	0.10	-0.22	0.00	0.01	-0.04
BMI	0.01	1.00	0.14	0.15	0.11	0.57	-0.30	0.20	0.22	-0.13
Glucose	0.23	0.14	1.00	0.50	0.70	0.31	-0.12	0.29	0.26	0.38
Insulin	0.03	0.15	0.50	1.00	0.93	0.30	-0.03	0.15	0.17	0.28
HOMA	0.13	0.11	0.70	0.93	1.00	0.33	-0.06	0.23	0.26	0.28
Leptin	0.10	0.57	0.31	0.30	0.33	1.00	-0.10	0.26	0.01	-0.00
Adiponectin	-0.22	-0.30	-0.12	-0.03	-0.06	-0.10	1.00	-0.25	-0.20	-0.02
Resistin	0.00	0.20	0.29	0.15	0.23	0.26	-0.25	1.00	0.37	0.23
MCP.1	0.01	0.22	0.26	0.17	0.26	0.01	-0.20	0.37	1.00	0.09
Classification	-0.04	-0.13	0.38	0.28	0.28	-0.00	-0.02	0.23	0.09	1.00

Figure 3.0

Most of the variables in the above matrix do not show a perfect positive or negative correlation between them. However, for example the correlation between HOMA and Insulin is 0.93 and it shows a good positive correlation.

By examining the below heatmap it shows that the correlation between the variables is low.

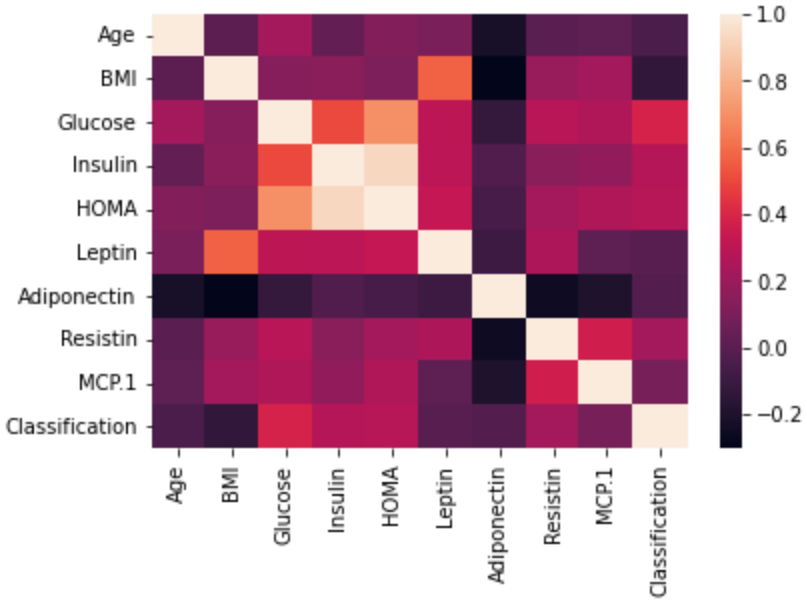

Figure 4.0

Therefore, didn't drop any of the columns at the beginning since no variable is showing a significant correlation over others.

Training Split

Data was split in 70%-30% proportion where 70% for the training set and 30% for the test set.

```
print(F"Train sample size = {len(X_train)}")
print(F"Test sample size  = {len(X_test)}")
```

```
Train sample size = 81
Test sample size  = 35
```

Figure 5.0

Model Training

Model was trained using the following classification algorithms.

Random Forest Classifier:

RandomForestClassifier(n_estimators=500, max_depth=10, n_jobs=3, verbose=1)

Decision Tree Classifier:

DecisionTreeClassifier(random_state=0, max_depth=10, min_samples_split=20)

Result

For both algorithms following were calculated.

Algorithm	Accuracy	F1 Score	Precision	roc_auc
Random Forest Classifier	0.657	0.657	0.647	0.794
Decision Tree Classifier	0.685	0.684	0.65	0.772

Table 1.0

According to the above table, it shows that Decision Tree Classifier shows better performance than the Random Forest Classifier.

Following table shows the performance after conducting the manual hyper parameter tuning for both algorithms.

Algorithm	Accuracy	F1 Score	Precision	roc_auc
RandomForestClassifier n_estimators=100, max_depth=None	0.657	0.631	0.656	0.789
RandomForestClassifier n_estimators=500, max_depth=None	0.657	0.647	0.657	0.781
RandomForestClassifier n_estimators=500, max_depth=10	0.685	0.666	0.685	0.795
RandomForestClassifier n_estimators=500, max_depth=20	0.714	0.733	0.712	0.789
DecisionTreeClassifier max_depth=5, min_samples_split=10	0.8	0.812	0.799	0.813
DecisionTreeClassifier max_depth=10, min_samples_split=20	0.685	0.65	0.684	0.772
DecisionTreeClassifier random_state=0, max_depth=20	0.628	0.591	0.622	0.66

Table 2.0

As per above stats it shows that both Random Forest Classifier and Decision Tree Classifier shows better performance with Hyper parameter tuning. Random Forest Classifier shows its best performance at n_estimators = 500 & max_depth = 20. Decision Tree Classifier shows its best performance at max_depth=5 & min_sample_split=10.

However, the stats in the above table also depicts that Decision Tree Classifier shows better performance in this case.

Further, Grid Search & Random Search was conducted on the Random Forest Classifier. In both occasions, “n_estimators=100, max_depth=None” selected as the best model but the performance pretty much similar to the above.

Feature importance was further examined.

```
feature_profile:
        feature  importance
2       Glucose    0.240099
0           Age    0.180082
4          HOMA    0.106374
7      Resistin    0.097825
1           BMI    0.086668
5        Leptin    0.074962
8         MCP.1    0.074871
3       Insulin    0.071537
6  Adiponectin    0.067583
```

Figure 6.0

In both searches, it showed that MCP.1, Insulin & Adiponectin as the least important features.

Model was retrained from the Random Forest Classifier and ran after dropping the above three columns from the data set. There is no significant improvement shown.

Conclusion

Breast cancer has become one of the major cancer type among the women throughout the world. Machine learning can be adopted for cancer detection and prediction. In this research, Breast Cancer Coimbra Data Set which consists of clinical features of 116 patients was used to develop a machine learning model to predict cancer patients. Random Forest Classifier and Decision Tree Classifier were used as the primary algorithms. Decision Tress Classifier was identified as the best performed algorithm for this model.