BREAST CANCER PREDICTION

CONTENTS

Executive Summary

The major goal of this study work is to use machine-learning algorithms to predict and diagnose breast cancer, and to evaluate and compare the models and identify the best algorithm for the breast cancer in terms of confusion matrix, accuracy, and precision.

BREAST CANCER PREDICTION

Highlights of Project

We used the Wisconsin Breast Cancer Diagnostic dataset and applied the following machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, and Decision Tree (C4.5) . Once the results were available, we conducted a performance review and comparison between these different classifiers. The main objective of this study was to use machine learning algorithms to predict and diagnose breast cancer, and to evaluate and compare models, and identify the best algorithm for breast cancer in terms of matrix. confusion, precision and accuracy. The random forest classifier has been shown to outperform all other classifiers and achieve the best accuracy (90.6%). The entire project is done in the Anaconda environment, based on the Python programming language and the Scikitlearn library.

Abstract

Breast cancer is causing an alarming increase in the number of deaths each year. Any advancement in cancer illness prediction and detection is critical for a healthy life.

It's a type of tumor that develops in the breast tissues. It is the most frequent cancer in women worldwide, and it is one of the main causes of mortality in women. Any improvement in cancer sickness prediction and diagnosis is crucial for a healthy life. As a result, high cancer prediction accuracy is crucial for maintaining treatment and survival criteria current for patients.

This research compares and contrasts machine learning algorithms for breast cancer prediction. Many studies have focused on breast cancer diagnosis; however, each approach has a distinct accuracy rate, which varies depending on the circumstances, tools, and datasets used.

In this study, we used the Breast Cancer Wisconsin Diagnostic dataset and applied the following machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression and Decision tree (C4.5). After obtaining the results, we performed a performance evaluation and comparison between these different classifiers. The major goal of this study work is to use machine-learning algorithms to predict and diagnose breast cancer, and to evaluate and compare the models and identify the best algorithm for the breast cancer in terms of confusion matrix, accuracy, and precision. Random forest classifier was shown to outperform all other classifiers and reach the best accuracy (90.6%). The entire project is carried out in the Anaconda environment, which is based on the Python programming language and the Scikit-learn library.

Introductory Section

Each year, breast cancer is responsible for an increase in the number of deaths. For a healthy life, any improvement in cancer sickness prediction and diagnosis is crucial. Breast cancer is more common in women than in males. Breast cancer affects around 2.1 million women each year, and it has become one of the leading causes of cancer-related deaths among women.

As per the reports of 2018, World Health Organization (WHO), Breast cancer is responsible for roughly 15% of deaths among women. One of the key factors that has resulted in the detection of breast cancer at a later stage is a lack of awareness. Another major reason is access to limited health resources which make the problem worse. So, here I'm using machine learning models to predict and diagnosis and identify the best model which is giving higher efficiency.

GitHub : https://github.com/dvk69/Breast-Cancer-Prediction.git

Review of available research

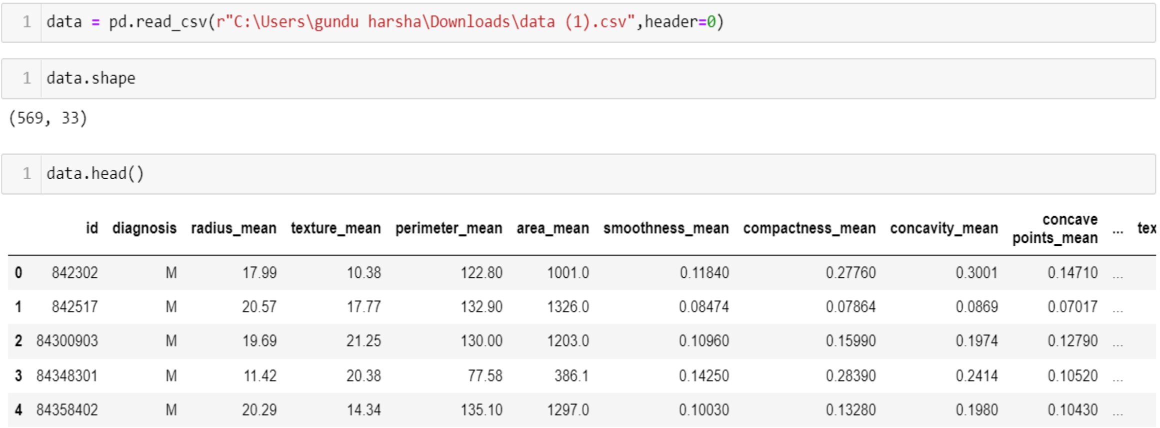
Healthcare is an open system for improvement through research on data mining and machine learning techniques. With 202,932 patient records, Delen et al. [15] investigated the prediction of breast cancer data. The dataset was separated into two groups: those who survived (93,273) and those who did not (109,659), and then the nave Bayes, neural network, and c4.5 decision tree methods were used. Williams et al. [10] used data mining classification algorithms to conduct research on breast cancer risk prediction. Breast cancer is the most prevalent cancer kind among Nigerian women. There are few services available to detect breast cancer before it is too late. As a result, they needed to find a reliable approach to predict breast cancer. The J48 decision trees and nave Bayes were two data mining approaches employed in their investigation. Dhar argues that computer expertise provides the ability to predict important outcomes and the future on the premise of data [16]. There are many studies on breast cancer datasets, most of which have sufficient classification accuracy [20,21].

Methodology

- Data is selected from the website ‘Kaggle’.
- Data is collected and sorted.
- Data is filtered.
- Label encoder is used to normalize the labels.
- We are plotting the count of the diagnosis column, the correlation graph and analyzing it.
- Now, implementing the model and finding out the accuracy of different machine learning algorithms.
- Used models: Logistic Regression, Random Forest, Decision Tree and SVM.
- We are doing the correlation and training the model to predict Breast cancer using different ML models
- Cross validation scores of each model are done in this project.

Results Section

- We are downloading the dataset from the Kaggle and importing it into Jupyter Notebook using pandas library



```
data = pd.read_csv(r"C:\Users\gundu harsha\Downloads\data (1).csv",header=0)
```

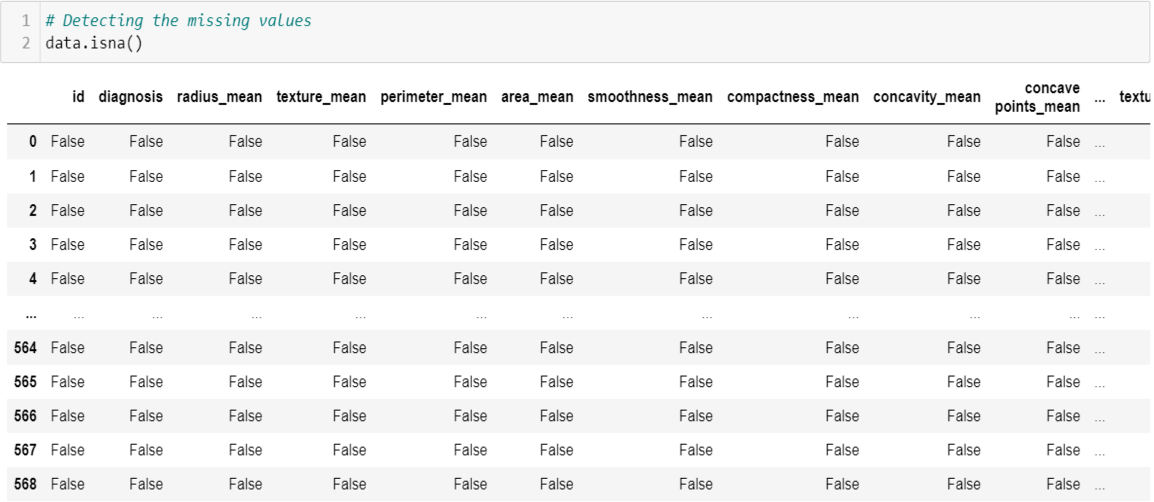
```
data.shape
```

```
(569, 33)
```

```
data.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	tex
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

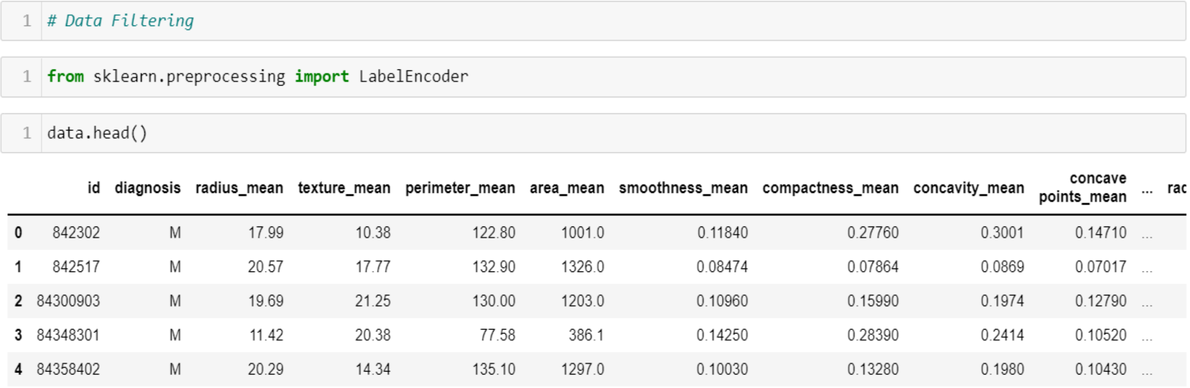
- Now, we are detecting the missing values



```
# Detecting the missing values
data.isna()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	textu
0	False	False	False	False	False	False	False	False	False	False	...	
1	False	False	False	False	False	False	False	False	False	False	...	
2	False	False	False	False	False	False	False	False	False	False	...	
3	False	False	False	False	False	False	False	False	False	False	...	
4	False	False	False	False	False	False	False	False	False	False	...	
...	...	...	...	...	...	...	...	...	...	...	...	
564	False	False	False	False	False	False	False	False	False	False	...	
565	False	False	False	False	False	False	False	False	False	False	...	
566	False	False	False	False	False	False	False	False	False	False	...	
567	False	False	False	False	False	False	False	False	False	False	...	
568	False	False	False	False	False	False	False	False	False	False	...	

- In Data filtering, by using Label Encoder we are finding values for unique variables

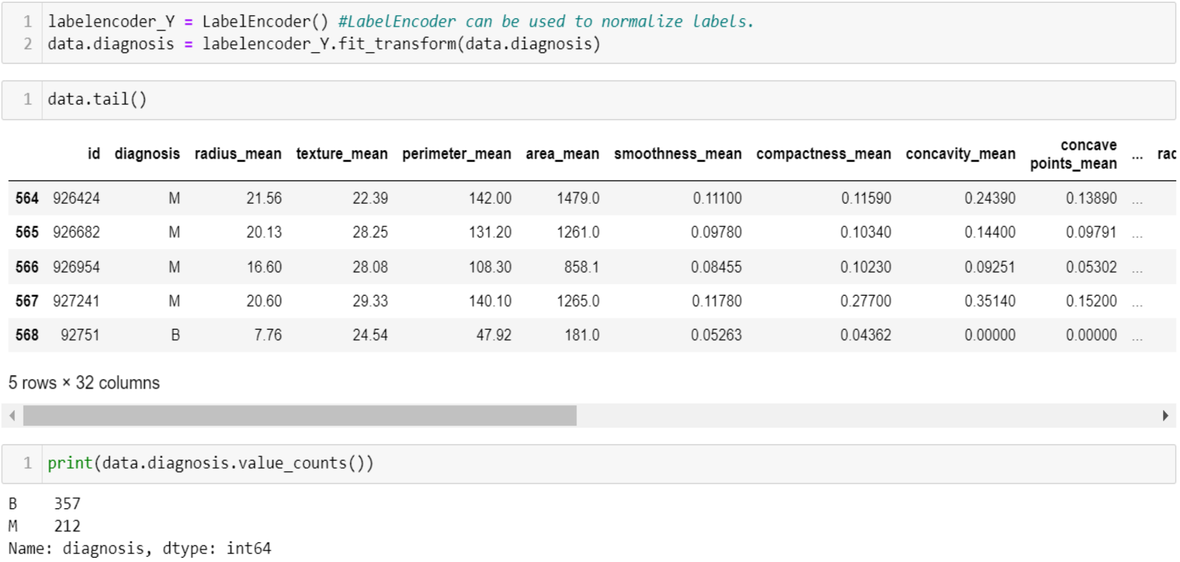


```
# Data Filtering
```

```
from sklearn.preprocessing import LabelEncoder
```

```
data.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	rac
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	



```
labelencoder_Y = LabelEncoder() #LabelEncoder can be used to normalize labels.
data.diagnosis = labelencoder_Y.fit_transform(data.diagnosis)
```

```
data.tail()
```

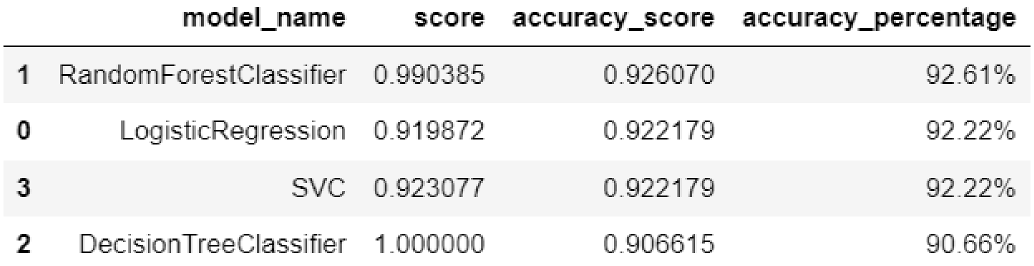
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	rac
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	...	
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	...	
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	...	
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	...	
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	...	

5 rows × 32 columns

```
print(data.diagnosis.value_counts())
```

```
B    357
M    212
Name: diagnosis, dtype: int64
```

- We have implemented machine learning models and finding the accuracy for each of the model.



	model_name	score	accuracy_score	accuracy_percentage
1	RandomForestClassifier	0.990385	0.926070	92.61%
0	LogisticRegression	0.919872	0.922179	92.22%
3	SVC	0.923077	0.922179	92.22%
2	DecisionTreeClassifier	1.000000	0.906615	90.66%

Discussion

we applied four main algorithms which are: SVM, Random Forests, Logistic Regression, Decision Tree to calculate and compare different results obtained based on confusion matrix, accuracy, sensitivity, precision. All algorithms have been programmed in Python using scikit-learn library in Anaconda environment. After an accurate comparison between our models, we found that Random Forest Classifier achieved a higher efficiency of 92.61%, Precision of 93% and f1-score of 93% and outperforms all other algorithms. In conclusion, Random Forest Classifier has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision.

Conclusion

In conclusion, Random Forest Classifier has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of efficiency and precision.
Random forest Classifier achieved a higher efficiency of 92.61%, Precision of 92%, f1-score of 93% and outperforms all other algorithms.

Contributions/References

[1] ‘WHO | Breast cancer’, WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2] Dataflow - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[3] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[4] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, ‘Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis’, Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[6] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 4225- 2/18/$31.00 ©2018 IEEE.

[7] L. Latchoumi, T. P., & Parthiban, “Abnormality detection using weighed particle swarm optimization and smooth support vector machine,” Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[8] A. H. Osman, “An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[9] Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.

[10] Williams T.G.S., Cubiella J., Griffin S.J. Risk prediction models for colorectal cancer in people with symptoms: A systematic review. *BMC Gastroenterol.* 2016;16:63. doi: 10.1186/s12876-016-0475-7.

[11] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag;2001.

[12] Quinlan JR. C4.5: Programs for Machine Learning.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[13] “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.”

[14] Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.

[15] Delen D., Walker G., Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif. Intell. Med.* 2005;34:113–127. doi: 10.1016/j.artmed.2004.07.002.

[16] Dhar V. Data science and prediction. *Commun. ACM.* 2013;56:64–73. doi: 10.1145/2500499

[17] Bazazeh D., Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis; Proceedings of the 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA); Ras Al Khaimah, UAE. 6–8 December 2016; pp. 1–4

[18] Aalaei S., Shahraki H., Rowhanimanesh A., Eslami S. Feature selection using genetic algorithm for breast cancer. 16 Computational and Mathematical Methods in Medicine diagnosis: An experiment on three different datasets. *Iran. J. Basic Med. Sci.* 2016;19:476