A MINI PROJECT REPORT ON

"Breast Cancer Classification"

Submitted by

Piyush Jha (17IT1027)

Kamaljit Kaur (17IT2039)

Prarthana Dokh (17IT2035)

Prateek koul (17IT1008)

Isha Gujar (17IT2042)

Mansi Lambat (17IT2036)

Under The Guidance Of

Prof. Reshma Gulwani

Department of Information Technology

Ramrao Adik Institute Of

Technology,Nerul,Navi Mumbai

(Affiliated to University of Mumbai)

(2020)

CERTIFICATE

This is to certify that the project entitled ` **Breast Cancer Classification '** being submitted by Piyush Jha (17IT1027), Kamaljit Kaur (17IT2039), Prarthana Dokh (17IT2035), Prateek Koul (17IT1008), Isha Gujar (17IT2042) And Mansi Lambat (17IT2036) to the University of Mumbai in partial fulfilment of the requirement for the award of the degree Of 'T.E. I.T' in "BUSINESS INTELLIGENCE LAB".

Project Guide	**External Examiner**	**Head of Department**
(Prof. Reshma Gulwani)	( )	(Dr. Ashish Jadhav)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others ideas or words have been included; we have adequately cited and Referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or Falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke Penal action from the sources which have thus not been properly cited or from whom Proper permission has not been taken when needed.

1. **Piyush Jha (17IT1027)**
2. **Kamaljit Kaur (17IT2039)**
3. **Prarthana Dokh (17IT2035)**
4. **Prateek koul (17IT1008)**
5. **Isha Gujar(17IT2042)**
6. **Mansi Lambat(17IT2036)**

Date:

Place:

ACKNOWLEDGEMENT

The project "**Breast Cancer Classification**" is a creative work of many minds. A proper synchronization between individual is must for any project to be completed successfully. One cannot imagine the power of the force that guides us all and neither can we succeed without acknowledging it.

We would like to express our gratitude to Principal **Dr. Mukesh D. Patil** and **Dr. Ashish Jadhav**, our Head of the department, Information Technology Engineering for encouraging and inspiring us to carry out the project in the department lab.

We would also like to thank our Guide **Prof. Jyoti Kundale** Department of the Information technology engineering for her expert guidance, encouragement and valuable suggestions at every step.

We also would like to thank all the staff members Department of the Information Technology Engineering for providing us with the required facilities and support towards the Completion of the project.

Last but not the least, we are thankful to our parents and friends for their constant Inspiration, encouragement and well wishes by which we have made a challenging project.

Piyush Jha (17IT1027)

Kamaljit Kaur (17IT2039)

Prarthana Dokh (17IT2035)

Prateek koul (17IT1008)

Isha Gujar (17IT2042)

Mansi Lambat (17IT2036)

PREFACE

We take great opportunity to present this Mini Project report on "**Breast Cancer Classification**" and put before readers some useful information regarding our project. We have made sincere attempts and taken every care to present this matter in precise and compact form, the language being as simple as possible. We are sure that the information contained in this volume certainly proves useful for better in sight in the scope and dimension of this project in it true perspective. The task of the completion of the project though being difficult was made quite simple, Interesting and successful due to deep involvement and complete dedication of our group members.

TABLE OF CONTENTS

1.Declaration..... I

2.Acknowledgement..... II

3.Preface.....III

4. Table of Contents..... IV

TABLE OF CONTENTS

1. PROBLEM STATEMENT……1

2.PROPOSED SYSTEM……2

3.DATASET……3

4. PLATFORM USED……4

5. IMPLEMENTATION DETAIL……5

6. RESULT……6

7. BUSINESS INTELLIGENCE……7

8. CONCLUSION……8

PROBLEM STATEMENT

Predicting if the cancer diagnosis is benign or malignant based on several observations/features.

30 features are used, examples:

- Radius
- Texture
- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness (perimeter^22/area - 1.0)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension (“coastline approximation” - 1)

PROPOSED SYSTEM

In this work, a data mining technique based breast cancer classification system classifies tumors into malignant or benign tumors using features of pain from several cell images with data mining technology. This analysis go aims to observe which features are most helpful in predicting malignant or benign cancer using different data sets and to see general trends that may aid us in model selection and hyper parameter selection.

The goal is to classify whether the breast cancer is benign or malignant using different data sets and different data mining classification algorithms & predict the discrete class of new input. Based on the results of this system, we compare the accuracy rate of various data mining algorithms or techniques.

DATA SET

- We have used to classic dataset from sklearn datasets available in sklearn machine learning library which is breast cancer wisconsin dataset (classification).
- The breast cancer dataset is a classic and very easy binary classification dataset.
- In this data set, we have :
    1. radius (mean of distances from centre to points on the perimeter)
    2. texture (standard deviation of grey-scale values)
    3. perimeter
    4. area
    5. smoothness (local variation in radius lengths)
    6. compactness (perimeter^2 / area - 1.0)
    7. concavity (severity of concave portions of the contour)
    8. concave points (number of concave portions of the contour)
    9. symmetry
    10. fractal dimension ("coastline approximation" - 1)

- Classes: 212 Malignant, 357
- Benign Target class:
    1. Malignant
    2. Benign

PLATFORM USED

1. Python Jupyter Notebook
2. Kaggle Dataset
3. Machine learning libraries
4. Data Analysis libraries for data frame and visualization

IMPLEMENTATION DETAILS

Breast cancer starts when malignant lumps which are cancerous begin to grow from the breast cells. Doctors may wrongly diagnose benign tumour (which is non-cancerous) as malignant tumour. There is need for a computer aided detection (CAD) systems which uses machine learning approach to provide accurate diagnosis of breast cancer. These CAD systems can aid in detecting breast cancer at an early stage. When, breast cancer is detected early enough, the survival rate increases because better treatment can be provided. This paper aims at detecting the tumour by providing certain parameters and using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset .The following algorithms were used in the system to obtain the desired results .Logistic Regression Training ,K Nearest Neighbour Training ,Support Vector Machine (Linear Classifier) Training ,Support Vector Machine (RBF Classifier) Training ,Gaussian Naive Bayes Training ,Decision Tree Classifier Training ,Random Forest Classifier Training ,AdaBoost Classifier Training ,XGBoost Classifier Training .

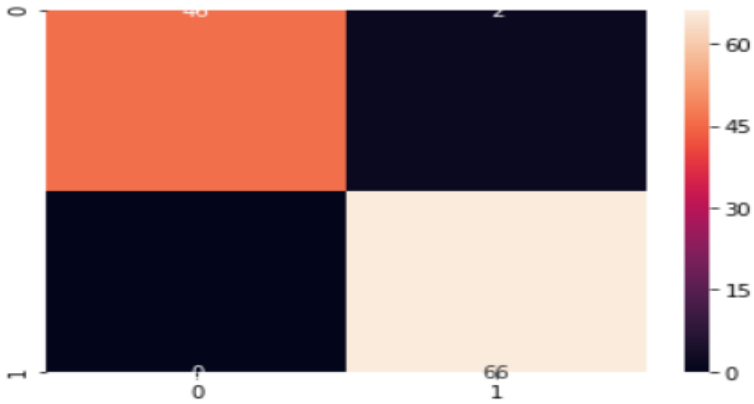
RESULT

1. The following classifiers were used to get the result of the accuracy

- Support Vector machine(RBF classifier): 96.49%
- Logistic Regression : 97.36%
- K nearest neighbor : 96.40%
- Naïve Bayes : 93.85%
- Random Forest Classifier Training Accuracy: 96.49 %
- AdaBoost Classifier Training Accuracy: 94.73%
- XGBoost Classifier Training Accuracy: 98.24%

XGBoost Classifier:

```
In [8]: model = XGBClassifier()
        model.fit(X_train, Y_train)
        Y_pred = model.predict(X_test)

In [9]: from sklearn.metrics import confusion_matrix
        cm = confusion_matrix(Y_test, Y_pred)
        cm

Out[9]: array([[46,  2],
               [ 0, 66]], dtype=int64)

In [11]: ans = model.score(X_test, Y_test)
         print('Accuracy',ans*100,'%')

         Accuracy 98.24561403508771 %

In [15]: sns.heatmap(cm,annot=True)

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x22c41eab048>
```



Logistic Regression:

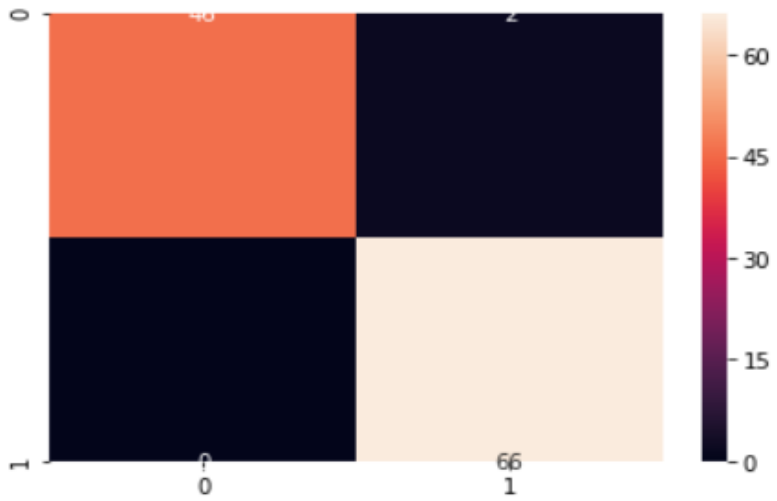
[17]:
```
from sklearn.linear_model import LogisticRegression
model2 = LogisticRegression(random_state = 0)
model2.fit(X_train, Y_train)
Y_pred2= model.predict(X_test)
ans = model2.score(X_test, Y_test)
print('Accuracy',ans*100,'%')
```

```
C:\Users\kamal\Anaconda3\lib\site-packages\sklearn\linear_mod
to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

```
Accuracy 97.36842105263158 %
```

[18]:
```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred)
cm
sns.heatmap(cm,annot=True)
```

[18]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22c41e93b88>
```



Random Forest Classifier:

```
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(n_estimators = 50,random_state = 0)
forest.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=None, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=50,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

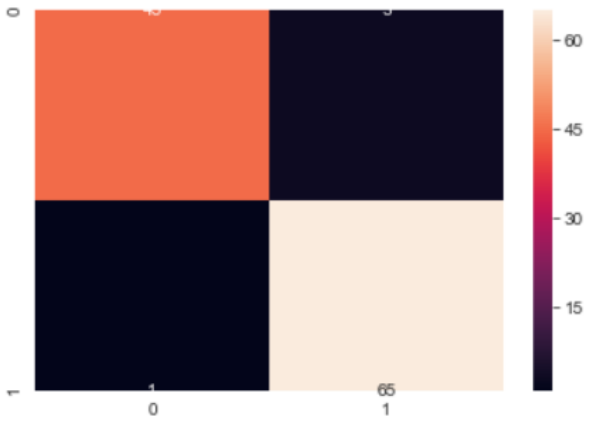
```
print('Random Forest Classifier Training Accuracy:', forest.score(X_test,y_test)*100,'%')
```

```
Random Forest Classifier Training Accuracy: 96.49122807017544 %
```

```
y_predict =forest.predict(X_test)
cm = confusion_matrix(y_test, y_predict)
cm
```

```
array([[45,  3],
       [ 1, 65]], dtype=int64)
```

```
In [105]: sns.heatmap(cm,annot=True)
Out[105]: <matplotlib.axes._subplots.AxesSubplot at 0x1e0f5770988>
```

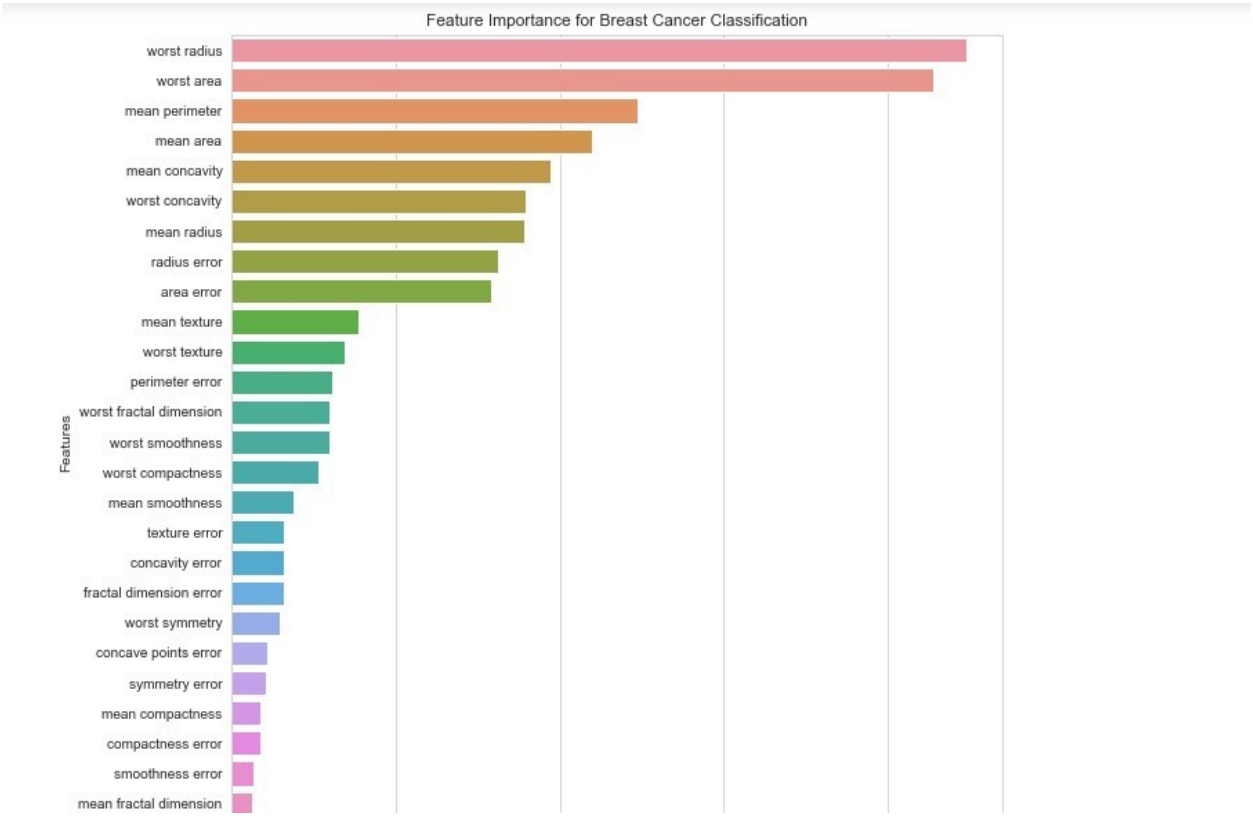


Importance of each feature is stated the graph:

In [109]:
```
rfc_features = pd.DataFrame(zip(forest.feature_importances_, df_cancer.columns[:-1]), columns = ['Importance', 'Features'])

# Sort in descending order for easy organization and visualization
rfc_features = rfc_features.sort_values(['Importance'], ascending=False)
```

In [108]:
```
plt.figure(figsize=(10,10))
sns.barplot(x = 'Importance', y = 'Features', data = rfc_features[3:], )
plt.title('Feature Importance for Breast Cancer Classification')
sns.set_style("whitegrid")
plt.show()
```

Feature Importance for Breast Cancer Classification



HOW BUSINESS INTELLIGENCE IS TO BE USED

In current scenario, doctors detect the type of breast cancer with 79% accuracy. Assume within a year, 30 lakh people are suffering from breast cancer & are treated by doctors with 79% accuracy. Now, remaining 21% inaccurate treatment will not be reliable. Our system provides results with 98% accuracy by using different datasets and different data mining algorithms.

Feature importance shows that the "worst_radius" as the most important feature.Therefore we recommended that worst_radius features should be extracted from each future biopsy as a strong predictor for diagonsing breast cancer

Thus, people suffering from breast cancer are saved in lakhs even cores.

CONCLUSION

In this system, we have studied different datamining and machine learning algorithms to predict the breast cancer using different data sets and different data mining algorithms. Based on the results of this system, most of the research works are concerned on comparing the accuracy rate of data mining various algorithms or techniques. Unfortunately, there is no tool that automatically diagnoses breast cancer. Further, there is no research work which applies personalized features for proposing the best treatment for patients.

In the future work, we will attempt to develop a tool with the help of intelligent agents and applying data mining tools with the capability of automatically breast cancer diagnosis and proposing the best treatment for patients.