

**BREAST CANCER PREDICTION**

Machine Learning Documentation

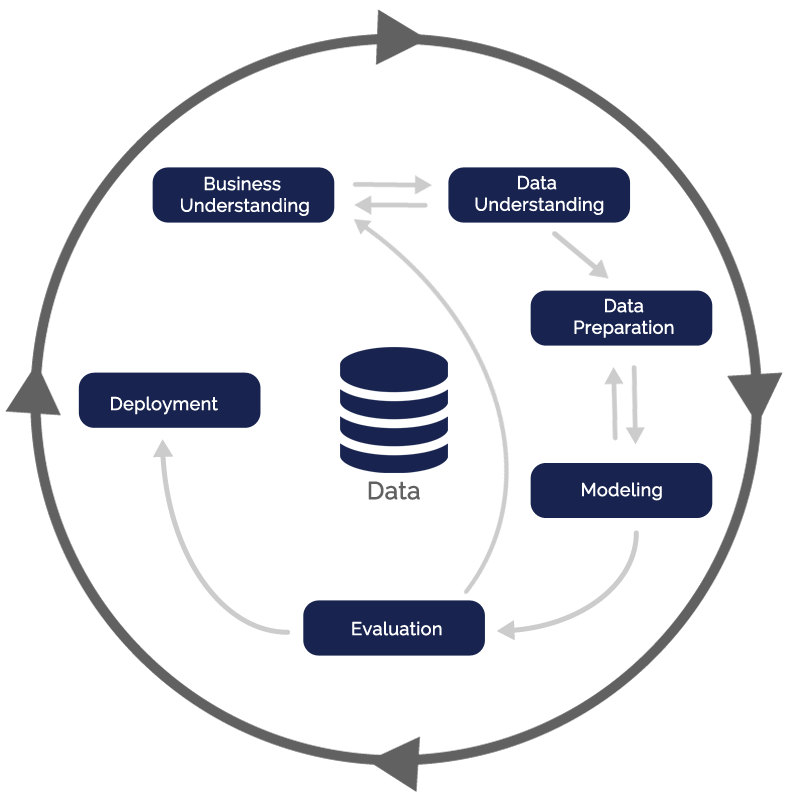


August 17, 2021

Women TECHSTERS

Lagos, Nigeria

**PROJECT LIFE CYCLE**



1. **INTRODUCTION**

1.1. **BACKGROUND AND PROJECT OBJECTIVE**

Cancer is a significant public health issue that has taken millions of lives, affecting both men and women. Research shows that women accounts for 99% of breast cancer cases. Research shows that early diagnosis can help save lives, improve prognosis and increase survival of patients greatly; as it bring about timely action to be taken by medical practioners.

The lack of good prognosis machines or models reduces the ability of medical personnels to quickly classify a patients tumor and make a treatment plan as such. Hence this research to develop a model that accuractely predicts a patients tumor; benign or malignant.

**2. PROBLEM STATEMENT**

The second major cause of women death is breast cancer with very low survival rate for diagnosed patients. Although cancer has been seen in both men and women, women accounts 98.80% of the total cases worldwide and most newly found cases are women.

This research seeks to observe features that contribate most to this cancer and classify patients with breast tumors accurately into benign or malignant tumors. This can help prevent patients with benign tumors from going through unnecessary treatments and help patients with malignant tumours get immediate medical care.

**3. HYPOTHESIS:**

Due to the high rate of deaths from breast cancer especially in women, with this research we aim to deploy our model into a hospital system and help health workers easily predict breast cancer thereby; reducing mortality rate.

**4. RESEARCH AIM:**

The aim of this research to classify breast tumors into benign or malignant thereby predicting breast cancer. This aim is to be achieved through building machine learning algorithms that seeks the best contributing features and predicts accurately, the type of tumor a patient has. By doing this, death toll as a result of breast cancer will be reduced as early diagnosis increases patients’ survival.

**Begin Tumors:** benign tumors are non cancerous cells or unsual growth that grows in the breast or breast tissue. Although benign tumors aren’t cancerous, some can eventually develop into breast cancer.

**Malignant Tumors**: malignant tumors are cancerous cells that grows in the breast or issues in the breast. Once a patient is diagnosed of having malignant tumors, he or she is to have immediate medical attention.

**5. METHODOLOGY OF RESEARCH**

The general methods we carried out for this research includes;

Data Cleaning

Data Preprocessing

Data Collection

Evaluation

Modelling

Feature Selection /

Dimensionality Reduction

Deployment

Method 1 – Data collection.

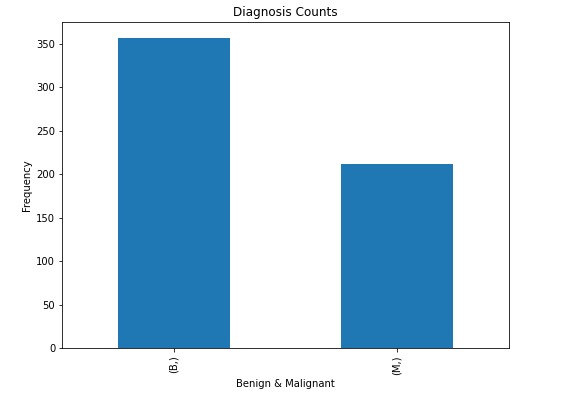
The data for this research was gotten from Kaggle <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>.

Method 2 – Data Cleaning.

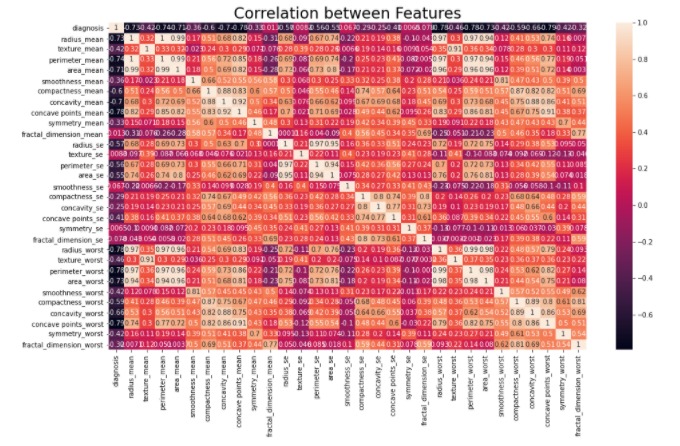
The data was cleaned to remove null values that will affect the modelling.

Method 3 – Data Preprocessing

The first thing we did here was to understand the basics of our data by carrying out an exploratory data analysis. We checked the shape, sum and count of null values, simple statistics of the data. The count of the diagnosis;



Then our target was mapped to an understandable format as it was in a categorical format. It is important to process this data such that they have consistent types that can be used in our model. This is the main objective for preprocessing to transform the data so that it can be fit for use in our models. During this stage, we created visualizations of our data to understand it more; visualizations like the general correlation of all the features we have;



Method 4 – Feature Selection - Dimensionality reduction:

Here, we dropped columns that were not useful or not going to add value to our model such as the Id. Also, we identified and removed our target variable from our features, and named them accurately.

Because this model is going into production, we understand how impatient users can be hence the need to reduce our features to only the most important features that are highly correlated to our target feature which is diagnosis. Out of 33 features we had 14 features selected excluding our target.

To do this, we used the LASSO variable selection method.

Method 4 – Model Selection

Model selection is an iterative process while building machine learning algorithms. We chose a model by examining many models used for classification problems until we found a model that minimizes loss most with good accuracy – random forest classifier. This model uses an ensemble method of classification. Ensemble meaning collection of smaller models (decision trees) due to the collection of trees, overfitting is avoided.

Model 5 – Deployment

To deploy our model, we saved our model as a pickle file, then created an API with POST request using flask and finally deployed on Heroku.

Link to notebook - <https://www.kaggle.com/nwankwokateogochukwu/aces-g4>

Link to API and deployment - <https://github.com/saniamomercy/aces-app.git>

**Tools used for this analysis includes;**

Kaggle cloud platform

Visual studio code

Heroku

Python and libraries like Scikit-Learn, flask, requests, pandas, NumPy, matplotlib, json

Microsoft Excel