HDSC August ’22 Capstone Project

By

Team Keras

Topic: Breast Cancer Prediction In Women

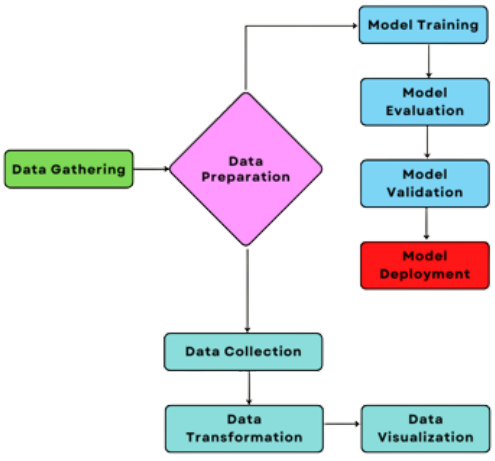
Given that October is Breast Cancer Awareness Month, now is a good opportunity to support the cause. Most cancer-related deaths in women are caused by breast cancer. With over 30% of all female malignancies, it is the most prevalent cancer in women globally and is regarded as a complex illness (i.e. 1.5 million women are diagnosed with breast cancer each year, and 500,000 women die from this disease in the world). While the death rate has reduced over the previous 30 years, this condition has become more prevalent. Mammography screening is thought to have a 20% reduction in mortality and a 60% improvement in cancer therapy. Early detection, nevertheless, can save lives.

Aim and Objectives

The goal of this project is to determine when cancer has the potential to cause harm, including death and to deploy a machine learning model that predicts the benignity or malignancy of a cancer based on the dataset provided.

Flow Process

The steps taken are illustrated with the flowchart below:



Data Gathering

The dataset was obtained from Kaggle via the link below:
[https://www.kaggle.com/datasets/merishnasuwal/breast-cancer-prediction-dataset]

Data Preparation

The following procedures were used to prepare the data:

- Data collection: The data gleaned was structured data, and it consisted of 570 rows and 6 columns namely; mean radius, mean texture, mean perimeter, mean area, mean smoothness and diagnosis.
- Data transformation: The data acquired led to the finding that a tumour is "Malignant" when the diagnosis is zero, and "Benign" when the diagnostic is one. To correct the existing mislabelling in the dataset, the diagnosis column's output was inverted to indicate that " 1 " represents cancer or malignant tumour.
- Data visualisation: Here, information was displayed using histograms, box plots, pair plots, and other visual aids to facilitate clear and simple interpretation.

Exploratory Data Analysis

This step was taken to better understand the data that had been gathered, give a more full picture of the data, and uncover and comprehend patterns that would explain unexpected results.

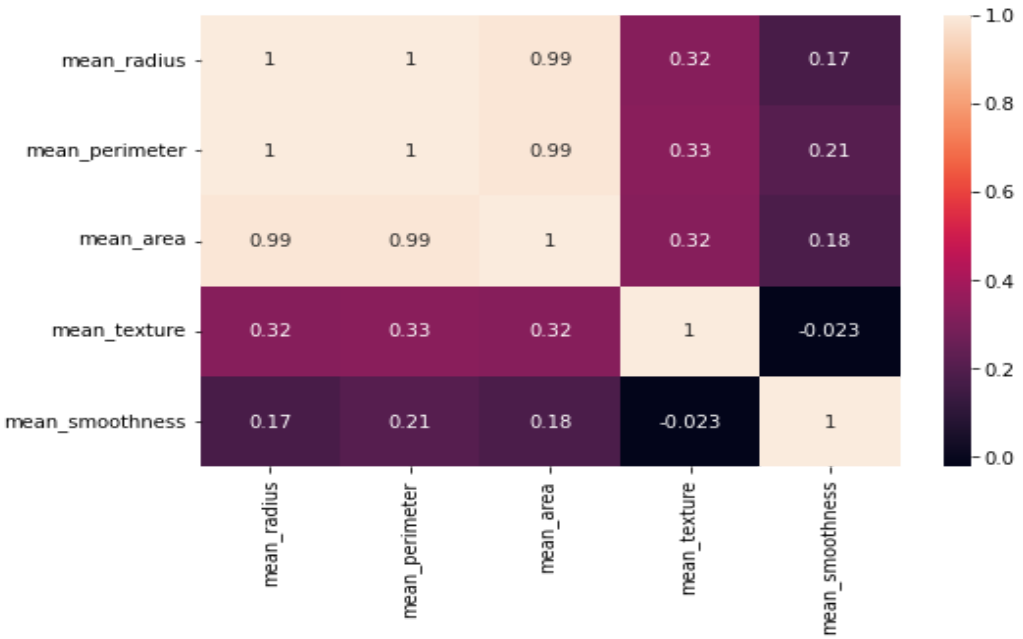


Figure 1: Correlation Heatmap

The relationship between the features is depicted in the correlation heatmap above. The mean perimeter, mean radius, and mean area can be inferred to have a strong positive association with one another.

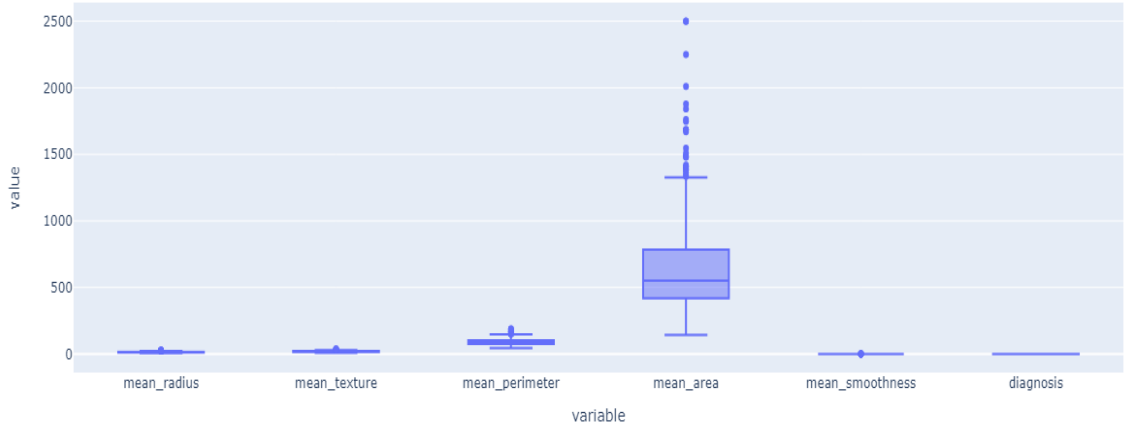


Figure 2: Box Plot Showing Outliers

The outliers in the data are shown in the figure above. It can be observed that there are no outliers in the diagnosis feature.

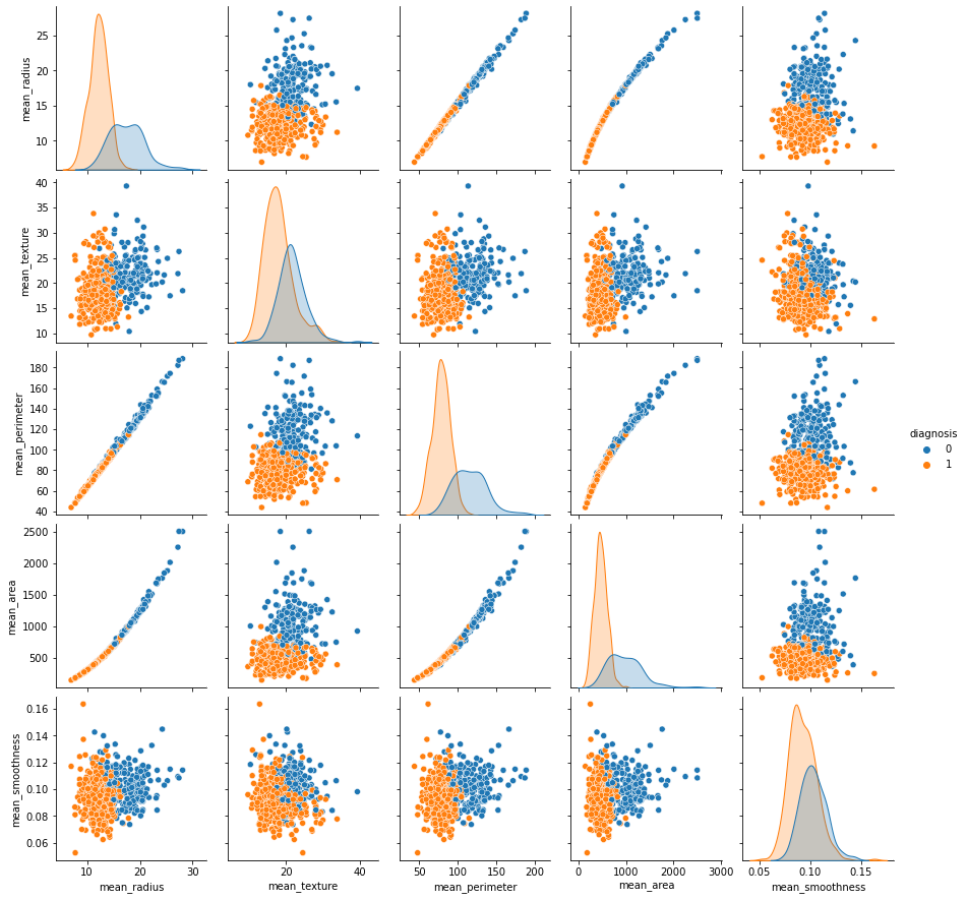


Figure 3: Pairplot

According to the plot, benign tumours (0) have low values for mean perimeter, mean radius, mean area, mean texture, and mean smoothness, whereas malignant tumours (1) have larger values for these parameters.

Model Training

In order to facilitate the creation and validation of the model, the dataset was divided into two parts: the training dataset and the testing dataset.

Model Evaluation

In the evaluation of the model, various supervised machine learning algorithms such as Logistic Regression Algorithm, Nearest Neighbour Algorithm, Support Vector Machine Algorithm, Naïve Bayes Algorithm, Decision Tree Algorithm, and Random Forest Classification Algorithm were used. Logistic Regression was the best performing algorithm, with the fewest false negatives and a high recall value.

Model Validation

The train/test split was employed to validate the model.

Model Deployment

The Django web framework was used to deploy the model to end users, and it was integrated with a user interface (UI) frontend. The link to the deployed model, including execution codes can be found in the links below:
Heroku: https://kcapstone.herokuapp.com/
GitHub: https://github.com/Keras-Capstone

Results

The vast majority of breast cancer cases are discovered when the illness is still curable and in its initial stages. Early detection is thus the most effective method of preventing breast cancer deaths and is critical to reducing the number of women who die from the disease each year. Fortunately, this model will help increase survival rates while raising knowledge of the factors that affect a tumour's benignity or malignancy going forward.

Conclusion and Recommendation

The ultimate aim of machine learning is to create algorithms that enable a system to automatically collect data and use that data to learn more. To avoid inaccurate results or forecasts and to guarantee the accuracy of the

breast cancer prediction model for any future innovation, it is crucial to always gather trustworthy, accurate, current, and relevant data.

Team Members

Etieneabasi Kingsley Effiong	Obinna Nwachukwu
Lilian Chidinma Nwafor	Akinlabi Akinkunmi Taofeek
Esther Chizitere Amadi	Abdulakeem Yusuf A
Ayisha Parveen	Adesina Lekan Samuel
Richard Orimolade	Adesokan Sulaimon Adewale
Akinrotimi Feyisola	Esther Opeyemi Ogundoyin
Orina Tolulope	Abdus-Salaam AbdulHafiz
Egunjobi Tunde	Sakeenat Adesina
Triumph Urias	Adetola Adebanwo
Princess Amadi	Joy Abiodun