Machine Learning Engineer Nanodegree

Capstone Proposal

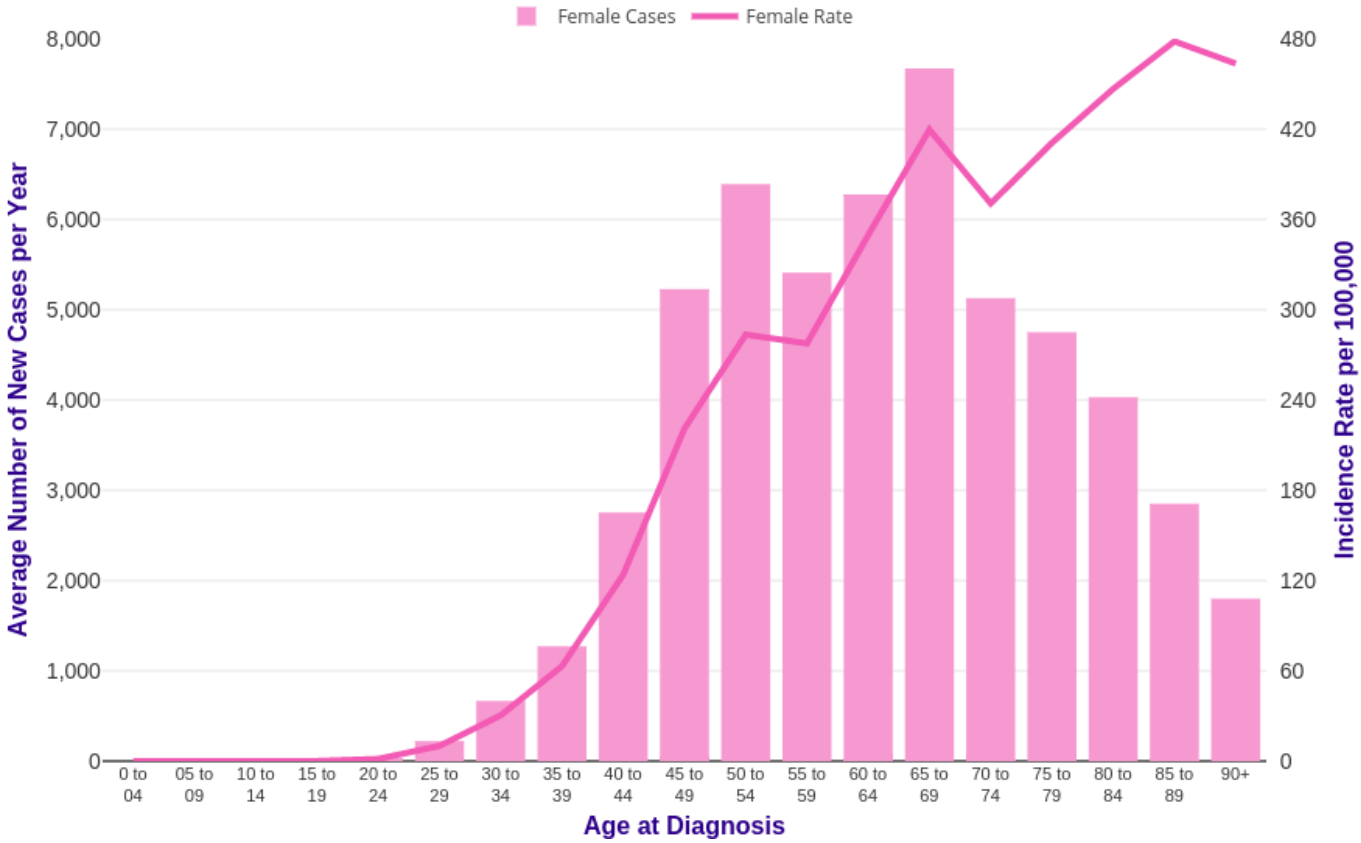
Breast Cancer Prediction

Nourhan Mohamed Fathy
April 30th ,2019

Proposal

Domain Background

Breast cancer (BC) is the most common invasive cancer in women, and it continues to be a worldwide medical problem since the number of cases has significantly increased over the past decades (~ 4.4 million), with a highly yearly incidence (~1.15 million new diagnosed).



Source: [Cancer Research UK]

Problem Statement

Breast tumor features are computed from digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image, using ML to predict whether the cancer is benign or malignant.

Datasets and inputs

The dataset(Breast Cancer Wisconsin(Diagnostic) Data set) provided on Kaggle and UCI Machine Learning Repository, it includes 570 rows (357 benign, 212 malignant), the attribute information are ID number and Diagnosis (M= malignant, B= benign), and ten -features computed for each cell nucleus:
1) Radius. 2) Texture. 3) Perimeter. 4) Area. 5) Smoothness. 6)Compactness. 7) Concavity. 8) Concave points. 9) Symmetry. 10) Fractal dimensions.
The mean, standard error, and worst of these features are computed for each image, resulting in 30 features.

Solution Statement

The solution is a classification model capable of predicting whether the tumor is malignant or benign. First, I will use pandas and NumPy to import and understand the data, I am inclined to use different machine learning algorithms, and compare their predictions.

Benchmark Model

Predict whether the tumor is malignant or benign, with R2- Score of 0.65 on the test dataset

Evaluation Metrics

Choose the model with highest R2- Score.

Project design

-Programming Language: Python 3.7

-Library: Pandas, NumPy, Matplotlib, Scikit-learn.

-Workflow:

-establish basic statistics and understanding of the dataset, perform processing if needed.

-train different classification models.

-test the model by testing set.

-calculate R2-score for each model.