Problem Statement

Breast cancer is one of the most common cancers among women in the world. Early detection of breast cancer is essential in reducing their life losses.

Build a predictive model using machine learning algorithms to predict whether the tumor is Benign or malignant.

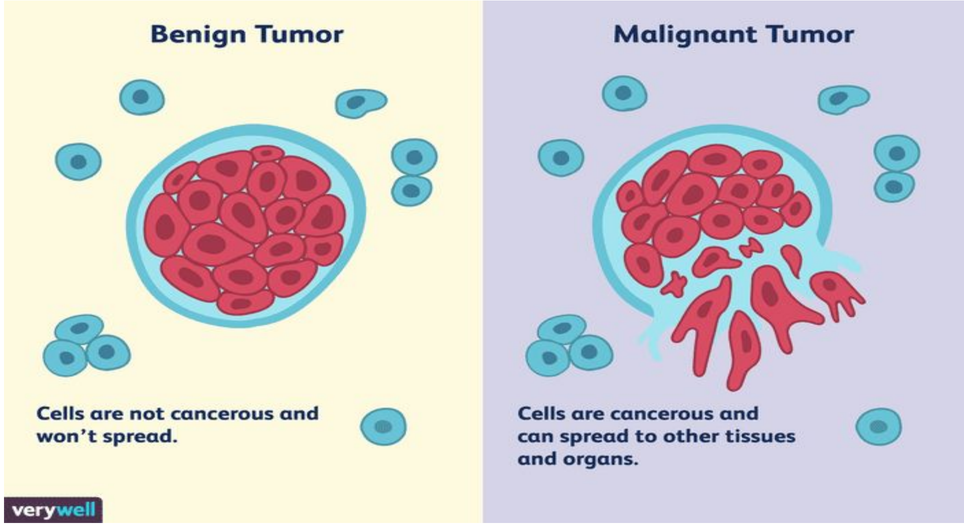
AIM:

To find the optimal k value having the lowest misclassification error, highest accuracy and highest AUC.

Brief Intro

Benign Tumor: If the tumor is benign , the cells are not cancerous. It won't invade nearby tissues or spread to other areas of the body (metastasize).Most grow slowly. Benign tumors usually don't recur once removed, but if they do it is usually in the same place.

Malignant tumor:Malignant means that the tumor is made of cancer cells, and it can invade nearby tissues. Some cancer cells can move into the bloodstream or lymph nodes, where they can spread to other tissues within the body this is called metastasis. Usually grow fairly rapidly.May recur after removal, sometimes in areas other the original site


Benign Tumor
Malignant Tumor
Cells are not cancerous and won't spread.
Cells are cancerous and can spread to other tissues and organs.
verywell

Model Selection

Implement KNN algorithm on training data, predicting labels for dataset and printing the accuracy of the model for different values of K.

What is KNN?

- KNN : K-Nearest Neighbour Algorithm
- Supervised Learning Algorithm
- Classification and Regression Algorithm
- Used for Binary and Multiclass Classification
- Works on Noisy Data
- Works efficiently for datasets with less observations(5000-10000)
- Known as Lazy Algorithm
- Not affected by outliers

ALGORITHM

- Read the training data from a file .
- Set K to some value
- Find the K nearest neighbors in the training data set based on the Euclidean distance
- Predict the class value by finding the maximum class represented in the K nearest neighbors
- Calculate the accuracy

DATA DESCRIPTION

- Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.
- Ten real-valued features are computed for each cell nucleus:
- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

Steps

1. Feature Selection : All variables are significant
2. No missing Values
3. No outlier imputation as KNN is not affected by outliers
4. Checking for presence of skewness in data
5. Applying Log Transformation on skewed variables
6. Converting Categorical variables into Numeric using LabelEncoder()
7. Scaling the data using StandardScaler()
8. Creating Train and Test Data
9. Applying KNN model
10. Select K value by Square Root method
11. Predicting Values
12. Saving to Excel File for client feasibility

Photo paste hotoy

Konta??

Nay disat   ha thhikke

First slide

Acha thamb whatsapp var photo pathavte

Aiknaaaa