Feature selection using Genetic Algorithms for Brest Cancer Classification

Name: Aafaq Saleh Al Shidhani

Introduction

Breast cancer is the second most common cancer in women and men worldwide and it is the second leading cause of death in women today. It even cause an adverse effect when left unnoticed for a long time. However, its early diagnosis provides significant treatment, so it’s better to start cure before detection to improve chance of survival. Therefore building a good Classifier for breast cancer predication will be very useful. Machine learning algorithms and artificial intelligence concepts can help us create a classifier that result in accurate diagnosis.

In this project we are creating a classifier to predict breast cancer. We aim to implement Feature selection using Genetic Algorithm (GA) to find the best features for the classifier.

Method

To create a classifier that predict breast cancer we will need a classification algorithm, K Nearest Neighbor algorithm is a good algorithm for this problem. We also want to increase the accuracy of this classifier by using feature selection, and to find the best features we will use genetic algorithm.

classification

For the classification we will implement K Nearest Neighbor algorithm (KNN) with three different k values(1,3 and 5). KNN chooses the K closest neighbors and then based on these neighbors, assigns a class for a new observation. The K is the number of neighbors, which must be an odd number to avoid having equal votes. KNN uses Euclidean distance to measure the distance between points. The following formula is used to calculate Euclidean distance:

$$Dist(X^n,X^m)=\sqrt{\sum_{i=1}^{D}(X_i^n-X_i^m)^2}$$

Feature selection

The objective of the feature selection is to reduce the number of features by identifying those that perform best in the classification process. It’s important to use feature selection because having irrelevant features in the data can decrease the accuracy of the models. There are many algorithm for feature selection, Here we will choose the genetic algorithm.

Genetic algorithm

Genetic algorithm (GA) is an heuristic optimization method inspired by the laws of genetics and the procedures of natural evolution. GA explores the space of possible subsets to obtain the set of features that maximizes the predictive accuracy and minimizes irrelevant attributes.
The genetic algorithm was implemented with three features, obtained by using different k values(1,3 and 5) in KNN classification.

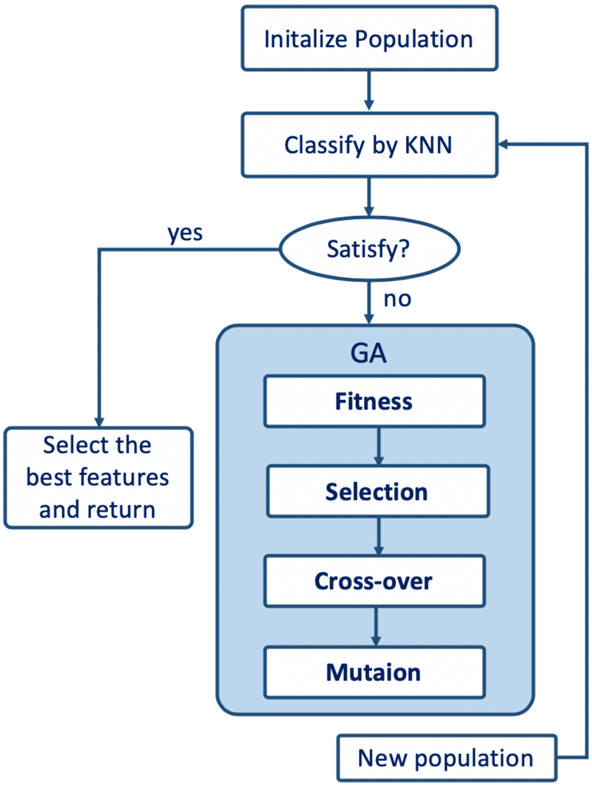


Figure 1 Flowchart of the proposed method

Dataset

The dataset that we will be using for our project is the Wisconsin Breast Cancer Database (WBCD) dataset . It was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples,taken from patients with solid breast masses and a graphical computer program called Xcyt. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.

The dataset include many features for Brest cancer diagnoses which make it very useful for this classification project.
The data set has 569 instances of 569 tumors and includes data on 30 features like the radius, texture, perimeter, area, etc. of a tumor. We will be using these features to train our model.

We will use the data from Scikit-learn which doesn’t require downloading.
link of database:
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

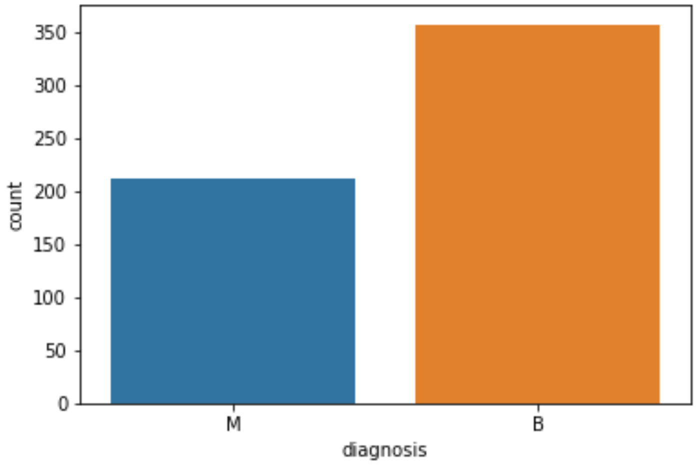
Data Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)

The ten real-valued features are computed for each cell nucleus:

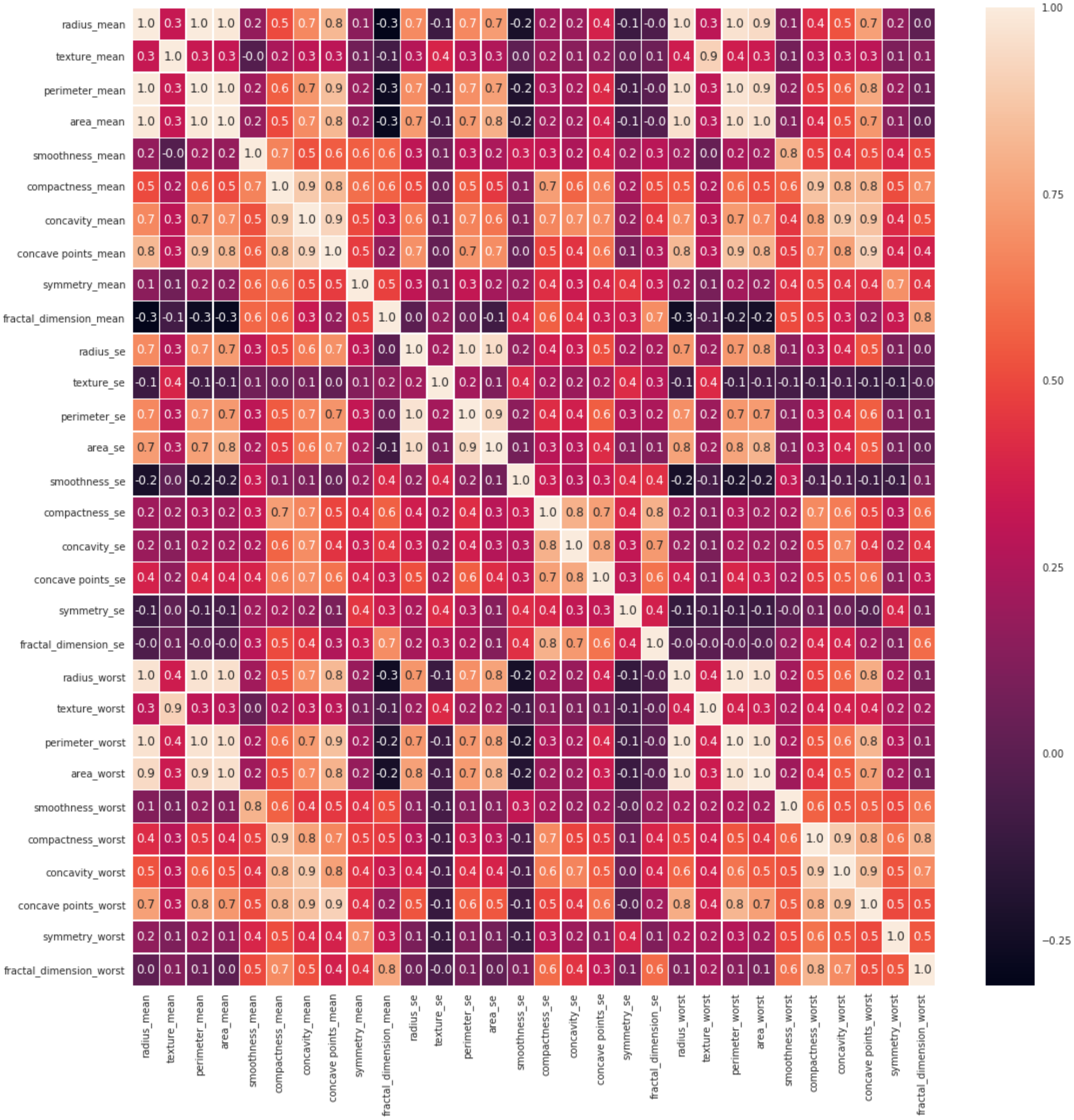
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Plot count of results values:



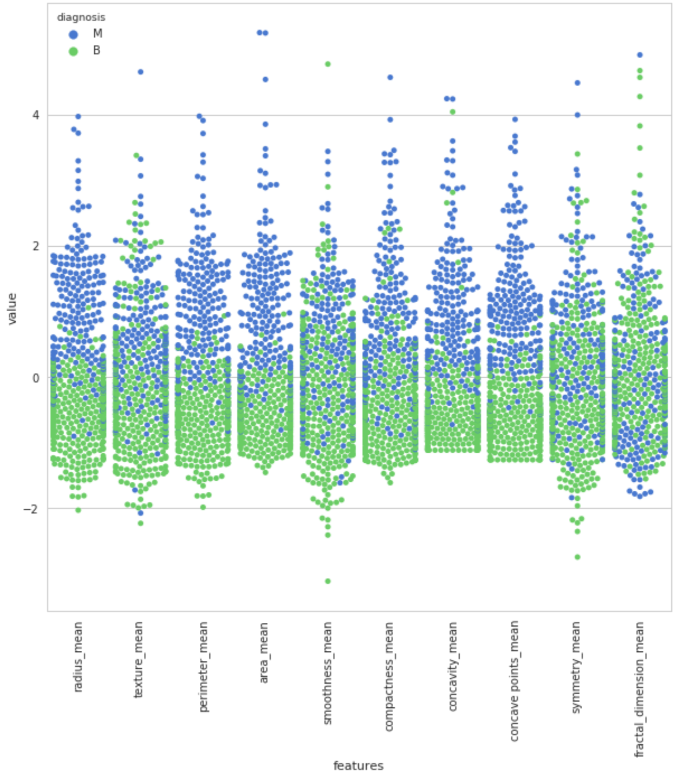
Histogram displays the distribution of data values. (M = malignant, B = benign)

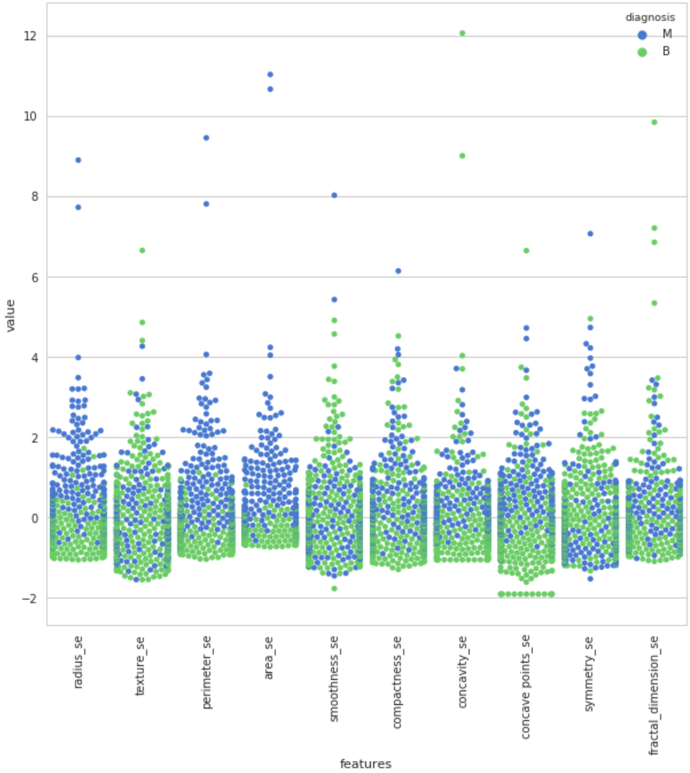
Heat map of the data (of all the 30 features):

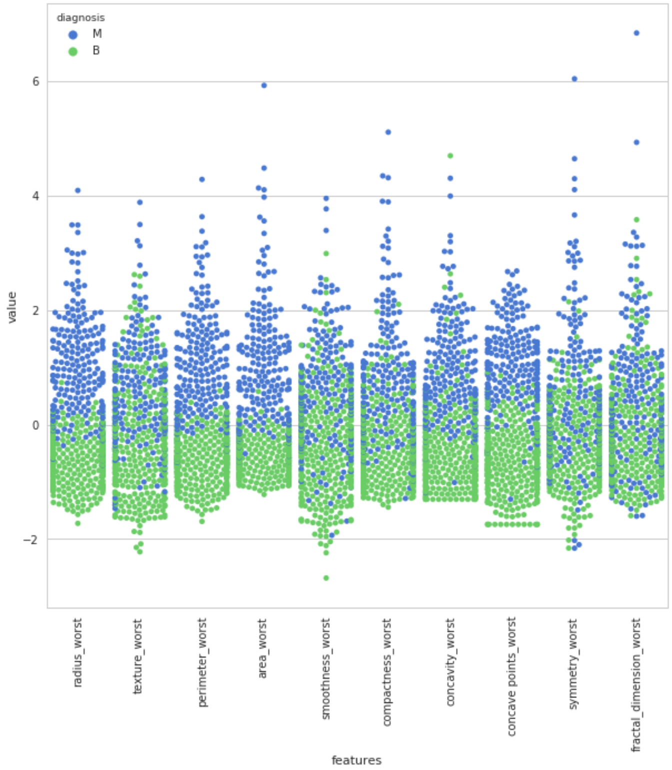


	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
radius_mean	1.0	0.3	1.0	1.0	0.2	0.5	0.7	0.8	0.1	-0.3	0.7	-0.1	0.7	0.7	-0.2	0.2	0.2	0.4	-0.1	-0.0	1.0	0.3	1.0	0.9	0.1	0.4	0.5	0.7	0.2	0.0
texture_mean	0.3	1.0	0.3	0.3	-0.0	0.2	0.3	0.3	0.1	-0.1	0.3	0.4	0.3	0.3	0.0	0.2	0.1	0.2	0.0	0.1	0.4	0.9	0.4	0.3	0.1	0.3	0.3	0.3	0.1	0.1
perimeter_mean	1.0	0.3	1.0	1.0	0.2	0.6	0.7	0.9	0.2	-0.3	0.7	-0.1	0.7	0.7	-0.2	0.3	0.2	0.4	-0.1	-0.0	1.0	0.3	1.0	0.9	0.2	0.5	0.6	0.8	0.2	0.1
area_mean	1.0	0.3	1.0	1.0	0.2	0.5	0.7	0.8	0.2	-0.3	0.7	-0.1	0.7	0.8	-0.2	0.2	0.2	0.4	-0.1	-0.0	1.0	0.3	1.0	1.0	0.1	0.4	0.5	0.7	0.1	0.0
smoothness_mean	0.2	-0.0	0.2	0.2	1.0	0.7	0.5	0.6	0.6	0.6	0.3	0.1	0.3	0.2	0.3	0.3	0.2	0.4	0.2	0.3	0.2	0.0	0.2	0.2	0.8	0.5	0.4	0.5	0.4	0.5
compactness_mean	0.5	0.2	0.6	0.5	0.7	1.0	0.9	0.8	0.6	0.6	0.5	0.0	0.5	0.5	0.1	0.7	0.6	0.6	0.2	0.5	0.5	0.2	0.6	0.5	0.6	0.9	0.8	0.8	0.5	0.7
concavity_mean	0.7	0.3	0.7	0.7	0.5	0.9	1.0	0.9	0.5	0.3	0.6	0.1	0.7	0.6	0.1	0.7	0.7	0.7	0.2	0.4	0.7	0.3	0.7	0.7	0.4	0.8	0.9	0.9	0.4	0.5
concave points_mean	0.8	0.3	0.9	0.8	0.6	0.8	0.9	1.0	0.5	0.2	0.7	0.0	0.7	0.7	0.0	0.5	0.4	0.6	0.1	0.3	0.8	0.3	0.9	0.8	0.5	0.7	0.8	0.9	0.4	0.4
symmetry_mean	0.1	0.1	0.2	0.2	0.6	0.6	0.5	0.5	1.0	0.5	0.3	0.1	0.3	0.2	0.2	0.4	0.3	0.4	0.4	0.3	0.2	0.1	0.2	0.2	0.4	0.5	0.4	0.4	0.7	0.4
fractal_dimension_mean	-0.3	-0.1	-0.3	-0.3	0.6	0.6	0.3	0.2	0.5	1.0	0.0	0.2	0.0	-0.1	0.4	0.6	0.4	0.3	0.3	0.7	-0.3	-0.1	-0.2	-0.2	0.5	0.5	0.3	0.2	0.3	0.8
radius_se	0.7	0.3	0.7	0.7	0.3	0.5	0.6	0.7	0.3	0.0	1.0	0.2	1.0	1.0	0.2	0.4	0.3	0.5	0.2	0.2	0.7	0.2	0.7	0.8	0.1	0.3	0.4	0.5	0.1	0.0
texture_se	-0.1	0.4	-0.1	-0.1	0.1	0.0	0.1	0.0	0.1	0.2	0.2	1.0	0.2	0.1	0.4	0.2	0.2	0.2	0.4	0.3	-0.1	0.4	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0
perimeter_se	0.7	0.3	0.7	0.7	0.3	0.5	0.7	0.7	0.3	0.0	1.0	0.2	1.0	0.9	0.2	0.4	0.4	0.6	0.3	0.2	0.7	0.2	0.7	0.7	0.1	0.3	0.4	0.6	0.1	0.1
area_se	0.7	0.3	0.7	0.8	0.2	0.5	0.6	0.7	0.2	-0.1	1.0	0.1	0.9	1.0	0.1	0.3	0.3	0.4	0.1	0.1	0.8	0.2	0.8	0.8	0.1	0.3	0.4	0.5	0.1	0.0
smoothness_se	-0.2	0.0	-0.2	-0.2	0.3	0.1	0.1	0.0	0.2	0.4	0.2	0.4	0.2	0.1	1.0	0.3	0.3	0.3	0.4	0.4	-0.2	-0.1	-0.2	-0.2	0.3	-0.1	-0.1	-0.1	-0.1	0.1
compactness_se	0.2	0.2	0.3	0.2	0.3	0.7	0.7	0.5	0.4	0.6	0.4	0.2	0.4	0.3	0.3	1.0	0.8	0.7	0.4	0.8	0.2	0.1	0.3	0.2	0.2	0.7	0.6	0.5	0.3	0.6
concavity_se	0.2	0.1	0.2	0.2	0.2	0.6	0.7	0.4	0.3	0.4	0.3	0.2	0.4	0.3	0.3	0.8	1.0	0.8	0.3	0.7	0.2	0.1	0.2	0.2	0.2	0.5	0.7	0.4	0.2	0.4
concave points_se	0.4	0.2	0.4	0.4	0.4	0.6	0.7	0.6	0.4	0.3	0.5	0.2	0.6	0.4	0.3	0.7	0.8	1.0	0.3	0.6	0.4	0.1	0.4	0.3	0.2	0.5	0.5	0.6	0.1	0.3
symmetry_se	-0.1	0.0	-0.1	-0.1	0.2	0.2	0.2	0.1	0.4	0.3	0.2	0.4	0.3	0.1	0.4	0.4	0.3	0.3	1.0	0.4	-0.1	-0.1	-0.1	-0.1	-0.0	0.1	0.0	-0.0	0.4	0.1
fractal_dimension_se	-0.0	0.1	-0.0	-0.0	0.3	0.5	0.4	0.3	0.3	0.7	0.2	0.3	0.2	0.1	0.4	0.8	0.7	0.6	0.4	1.0	-0.0	-0.0	-0.0	-0.0	0.2	0.4	0.4	0.2	0.1	0.6
radius_worst	1.0	0.4	1.0	1.0	0.2	0.5	0.7	0.8	0.2	-0.3	0.7	-0.1	0.7	0.8	-0.2	0.2	0.2	0.4	-0.1	-0.0	1.0	0.4	1.0	1.0	0.2	0.5	0.6	0.8	0.2	0.1
texture_worst	0.3	0.9	0.3	0.3	0.0	0.2	0.3	0.3	0.1	-0.1	0.2	0.4	0.2	0.2	-0.1	0.1	0.1	0.1	-0.1	-0.0	0.4	1.0	0.4	0.3	0.2	0.4	0.4	0.4	0.2	0.2
perimeter_worst	1.0	0.4	1.0	1.0	0.2	0.6	0.7	0.9	0.2	-0.2	0.7	-0.1	0.7	0.8	-0.2	0.3	0.2	0.4	-0.1	-0.0	1.0	0.4	1.0	1.0	0.2	0.5	0.6	0.8	0.3	0.1
area_worst	0.9	0.3	0.9	1.0	0.2	0.5	0.7	0.8	0.2	-0.2	0.8	-0.1	0.7	0.8	-0.2	0.2	0.2	0.3	-0.1	-0.0	1.0	0.3	1.0	1.0	0.2	0.4	0.5	0.7	0.2	0.1
smoothness_worst	0.1	0.1	0.2	0.1	0.8	0.6	0.4	0.5	0.4	0.5	0.1	-0.1	0.1	0.1	0.3	0.2	0.2	0.2	-0.0	0.2	0.2	0.2	0.2	0.2	1.0	0.6	0.5	0.5	0.5	0.6
compactness_worst	0.4	0.3	0.5	0.4	0.5	0.9	0.8	0.7	0.5	0.5	0.3	-0.1	0.3	0.3	-0.1	0.7	0.5	0.5	0.1	0.4	0.5	0.4	0.5	0.4	0.6	1.0	0.9	0.8	0.6	0.8
concavity_worst	0.5	0.3	0.6	0.5	0.4	0.8	0.9	0.8	0.4	0.3	0.4	-0.1	0.4	0.4	-0.1	0.6	0.7	0.5	0.0	0.4	0.6	0.4	0.6	0.5	0.5	0.9	1.0	0.9	0.5	0.7
concave points_worst	0.7	0.3	0.8	0.7	0.5	0.8	0.9	0.9	0.4	0.2	0.5	-0.1	0.6	0.5	-0.1	0.5	0.4	0.6	-0.0	0.2	0.8	0.4	0.8	0.7	0.5	0.8	0.9	1.0	0.5	0.5
symmetry_worst	0.2	0.1	0.2	0.1	0.4	0.5	0.4	0.4	0.7	0.3	0.1	-0.1	0.1	0.1	-0.1	0.3	0.2	0.1	0.4	0.1	0.2	0.2	0.3	0.2	0.5	0.6	0.5	0.5	1.0	0.5
fractal_dimension_worst	0.0	0.1	0.1	0.0	0.5	0.7	0.5	0.4	0.4	0.8	0.0	-0.0	0.1	0.0	0.1	0.6	0.4	0.3	0.1	0.6	0.1	0.2	0.1	0.1	0.6	0.8	0.7	0.5	0.5	1.0

swarm plot of data (for the 30 features split in 3 graphs each contain 10):







Experiments

Data Preprocessing:

1. Import needed libraries.
2. Read data.
3. Split the dataset into Training and Testing sets: 80% for training and 20% for testing data.

Classification process:

1. The classifier was trained the on training set and tested on the testing set using KNN.
2. Compute the accuracy performance by comparing the real label (y_test) to the labels predicted by the classifier (y_pred) .
3. Repeat the process for K = 1,3 and 5.

Feature selection process:

We used a genetic algorithm for the feature selection. Here is a description on how the implemented genetic algorithm works:

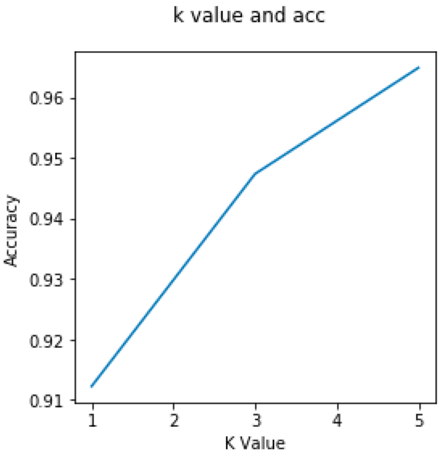
1. **Initial Population**– Initialize the population randomly based on the data.
2. **Fitness function**– Find the fitness value of the each of the chromosomes(a chromosome is a set of parameters which define a proposed solution to the problem that the genetic algorithm is trying to solve)
3. **Selection**– Select the best fitted chromosomes as parents to pass the genes for the next generation and create a new population
4. **Cross-over**– Create new set of chromosome by combining the parents and add them to new population set
5. **Mutation**– Perfrom mutation which alters one or more gene values in a chromosome in the new population set generated. Mutation helps in getting more diverse oppourtinity. Obtained population will be used in the next generation

Results:

1. **k-NN classification accuracy for k = 1, 3 and 5:**
   To find the relation between K value and KNN accuracy we tested KNN with three different values of K (1,3 and 5)

Table 1 Experimental results of accuracy of k values

k values	Accuracy
1	0.9122807
3	0.9473684
5	0.9649122



From the results we can say that the accuracy increases as K value increase. The best accuracy is 0.965 when k=5. And he lowest accuracy is 0.912 When k= 1.

2. **The best features for each k using GA:**

Table 2 Experimental results of k values using GA

k values	Accuracy Before GA	Accuracy After GA
1	0. 95614	0. 98245
3	0. 92982	0. 97368
5	0. 95614	0. 96491

Selected Features K = 1:

1. mean radius
2. mean texture
3. mean perimeter
4. mean smoothness
5. mean symmetry
6. texture error
7. smoothness error
8. concavity error
9. symmetry error
10. worst radius
11. worst texture
12. worst perimeter
13. worst compactness
14. worst concave points

Selected Features K = 3:

1. mean texture
2. mean smoothness
3. mean compactness
4. mean concave points
5. radius error
6. perimeter error
7. smoothness error
8. compactness error
9. concave points error
10. symmetry error
11. fractal dimension error
12. worst perimeter
13. worst smoothness
14. worst compactness
15. worst fractal dimension

Selected Features K = 5:

1. mean radius
2. mean smoothness
3. mean compactness
4. mean symmetry
5. texture error
6. compactness error
7. worst texture
8. worst perimeter
9. worst smoothness
10. worst compactness
11. worst concavity
12. worst fractal dimension

Table 2 shows the experiment results of k values before and after applying GA. From the results it's clear that Genetic Algorithm improved the Accuracy . the best increase in the accuracy was when k=3, GA increased the accuracy by 4.7%. The average number of selected features by GA is 14.

Conclusion

This project represent a Brest Cancer classifier with feature selection using Genetic Algorithm (GA). The classifier was based on kNN method. The project showed that on the Wisconsin Breast Cancer dataset it's possible to find features that improve the accuracy of predicting breast cancer. We also found that choosing the right value of k is important to get the best accuracy.

The project showed an implementation of Genetic Algorithm (GA) that could be used to to improve the accuracy of kNN classifier when used for breast cancer diagnoses . Therefore GA can be very useful in improving the accuracy when creating a Breast Cancer classifier.

References:

- Saha, Suraj. (2019). DATA MINNING PROJECT BREAST CANCER CLASSIFICATION USING DATA MINNING APPROACH.
- Convolutional Neural Network for Breast Cancer Classification. https://www.kdnuggets.com/2019/10/convolutional-neural-network-breast-cancer-classification.html
- Genetic Algorithm in Machine Learning using Python https://datascienceplus.com/genetic-algorithm-in-machine-learning-using-python/
- https://www.kaggle.com/kanncaa1/feature-selection-and-data-visualization