Breast Cancer Project

Dataset includes the diagnosis of breast cancer for **569 patients** based on **30 different factors** (tumor’s features) excluding the ID since it seems to be a useless factor that doesn’t affect the diagnosis. Based on these factors variables, the diagnosis of the tumor is either **Malignant (M) or Benign (B).** Therefore the dependent variable in this machine learning model is the Diagnosis, which is her of string type (M) or (B). On the other hand, there are 30 different independent variables that determine the diagnosis case of the tumor.

The 30 variables are in short are:

|  |  |  |
| --- | --- | --- |
| Independent Variable | Description in English | Description in Arabic |
| **Radius\_mean** | mean of distances from center to points on the perimeter | متوسط المسافات من المركز إلى النقاط على المحيط |
| **Texture\_mean** | standard deviation of gray-scale values | الانحراف المعياري لقيم مقياس الرمادي |
| **perimeter\_mean** | Mean size of the core tumor | متوسط حجم الورم |
| **area\_mean** | Mean of area of tumor | متوسط مساحة الورم |
| **smoothness\_mean** | Mean of local variation in radius lengths | متوسط التباين في أطوال نصف القطر |
| **compactness\_mean** | Mean of perimeter^2 / area - 1.0 | متوسط المحيط تربيع / المساحة-1 |
| **concavity\_mean** | mean of severity of concave portions of the contour | شدة التقعر لاجزاء الكونتور |
| **concave points\_mean** | mean for number of concave portions of the contour | عدد الاجزاء المقعرة من الكونتور |
| **symmetry\_mean** |  | متوسط التشابة |
| **fractal\_dimension\_mean** | mean for "coastline approximation" - 1 | تتعلق بالشكل الخارجي حيث يتم تقسيم المحيط الخارجي الى عدة خطوط بحيث تشمل اكبر قدر ممكن الورم |
| **radius\_se** | standard error for the mean of distances from center to points on the perimeter | الخطأ المعياري لمتوسط المسافات من المركز إلى النقاط على المحيط |
| **texture\_se** | standard error for standard deviation of gray-scale values | الخطأ المعياري للانحراف المعياري لقيم التدرج الرمادي |
| **perimeter\_se** | standard error for **perimeter** | الخطأ المعياري في المحيط |
| **area\_se** | standard error for **area** | الخطأ المعياري في المساحة |
| **smoothness\_se** | standard error for local variation in radius lengths | الخطأ المعياري للتغير المحلي في أطوال نصف القطر |
| **compactness\_se** | standard error for perimeter^2 / area - 1.0 | الخطأ المعياري في المحيط تربيع / المساحة-1 |
| **concavity\_se** | standard error for severity of concave portions of the contour | الخطأ المعياري ل شدة التقعر لاجزاء الكونتور |
| **concave points\_se** | standard error for number of concave portions of the contour | الخطا المعياري ل عدد الاجزاء المقعرة من الكونتور |
| **symmetry\_se** |  | الخطا المعياري في متوسط التشابة |
| **fractal\_dimension\_se** | standard error for "coastline approximation" - 1 | الخطأ المعياري لـ "تقريب الخط الساحلي" - 1 |
| **radius\_worst** | "worst" or largest mean value for mean of distances from center to points on the perimeter | القيمة "الأسوأ" أو الأكبر لمتوسط المسافات من المركز إلى النقاط على المحيط |
| **texture\_worst** | "worst" or largest mean value for standard deviation of gray-scale values | القيمة "الأسوأ" أو الأكبر للانحراف المعياري لقيم التدرج الرمادي |
| **perimeter\_worst** |  | القيمة "الأسوأ" أو الأكبر للانحراف المعياري لمتوسط حجم الورم |
| **smoothness\_worst** | "worst" or largest mean value for local variation in radius lengths | "أسوأ" أو أكبر قيمة متوسطة للتغير المحلي في أطوال نصف القطر |
| **compactness\_worst** | "worst" or largest mean value for perimeter^2 / area - 1.0 | القيمة "الأسوأ" أو الأكبر للمحيط ^ 2 / المنطقة - 1.0 |
| **concavity\_worst** | "worst" or largest mean value for severity of concave portions of the contour | القيمة "الأسوأ" أو الأكبر لشدة الأجزاء المقعرة من الكونتور |
| **concave points\_worst** | "worst" or largest mean value for number of concave portions of the contour | "أسوأ" أو أكبر قيمة متوسطة لعدد الأجزاء المقعرة من الكونتور |
| **fractal\_dimension\_worst** | "worst" or largest mean value for "coastline approximation" - 1 | "أسوأ" أو أكبر قيمة متوسطة لـ "تقريب الخط الساحلي" - 1 |

# Steps for writing the code of Breast cancer machine learning model

1. import for the required libraries
2. read the dataset
3. determine X, y
4. Label encoder for y
5. split the data for train and test
6. Feature scaling
7. Apply Logistic regression model or KNN model
8. Confusion matrix to check the accuracy of the two models

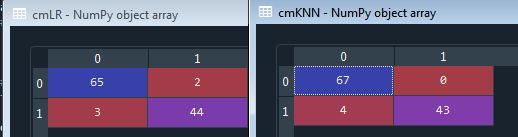


Figure 1: Confusion Matrices for LR and KNN Models