MRR project: Breast cancer diagnosis

Fitahiry RAJAOBELINA, Dorart RAMADANI

10 décembre 2023

1 Introduction

The dataset that we will use all along the project is named "Breast cancer Winsconsin (Diagnostic)" and is associated to health and medicine subject. It is a multivariate dataset (569 observations of 30 variables), all the features values are numerical except the feature "diagnosis".

The main information that the data-set is containing is the diagnosis of breast cancer, saying that a person (identified by the column "ID") is positive or not (in the column "diagnosis") to breast cancer according to plenty of features. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

2 Variables description and visualization

As we stated just before, the dataset contain information on cancer diagnosis, which is related to other variables of the breast mass such as radius mean, smoothness mean, concavity worst, etc.

2.1 Target variable : diagnosis

The columns "diagnosis" contain the values M (for malign) and B (for benign). Here's a bar plot which sum up the number of benign and malign in the dataset, nearly 350 person are benign where 200 are malign.

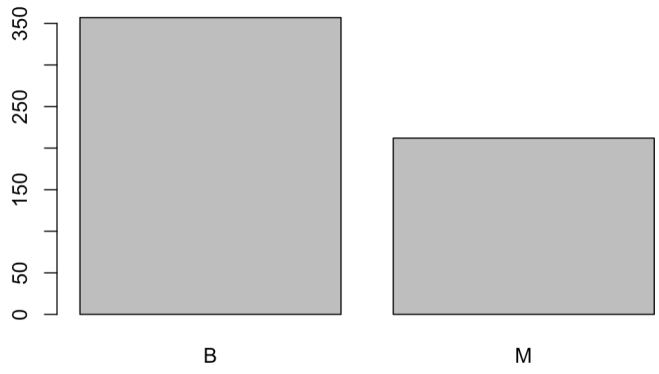

Figure 1 – Bar plot of malign and benign

2.2 Features visualization

The final goal of the project is to predict if a breast mass is benign of malign considering its , those characteristics are all the columns in the dataset except the column "Id" and the column "diagnosis".

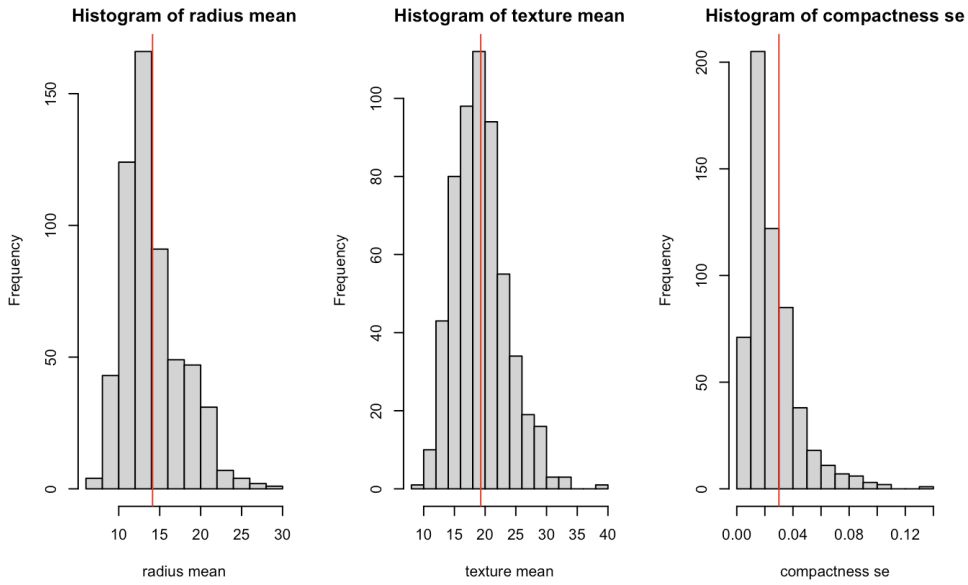

FIGURE 2 – Histogram of some features

We can observe that the majority the variables have their values gathered in a certain interval, some histograms are even approaching a bell curve (texture mean for instance).

2.3 Features relation to target variable

To visualize the relationship between the features and the target variable, scatter plots are created. Each feature is plotted against the target variable, with different marker colors representing benign or malignant tumors. These visualizations help in understanding the distribution of feature values and potential separability between the two classes.

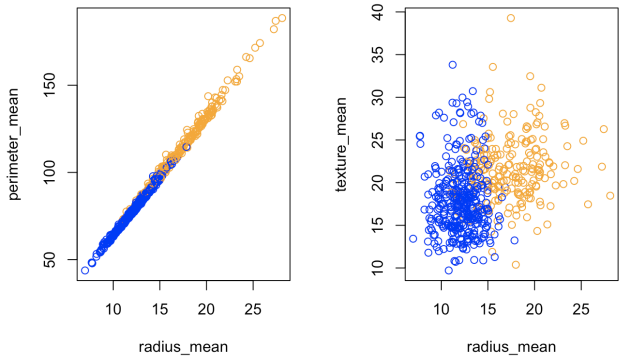

FIGURE 3 –

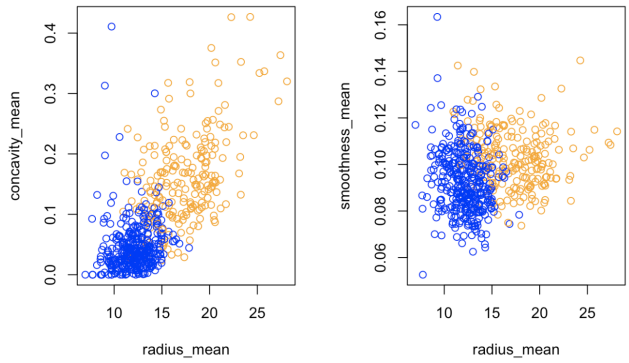

FIGURE 4 –

Those plots shows that a classification is possible to predict the cancer diagnosis knowing the breast mass characteristics.

2.4 Understanding correlation between features

Correlation between variables has a predictive power, it helps to predict how changes in one variable will affect changes in another variable. We can used it as a starting point for investigating causal relationships between variables.

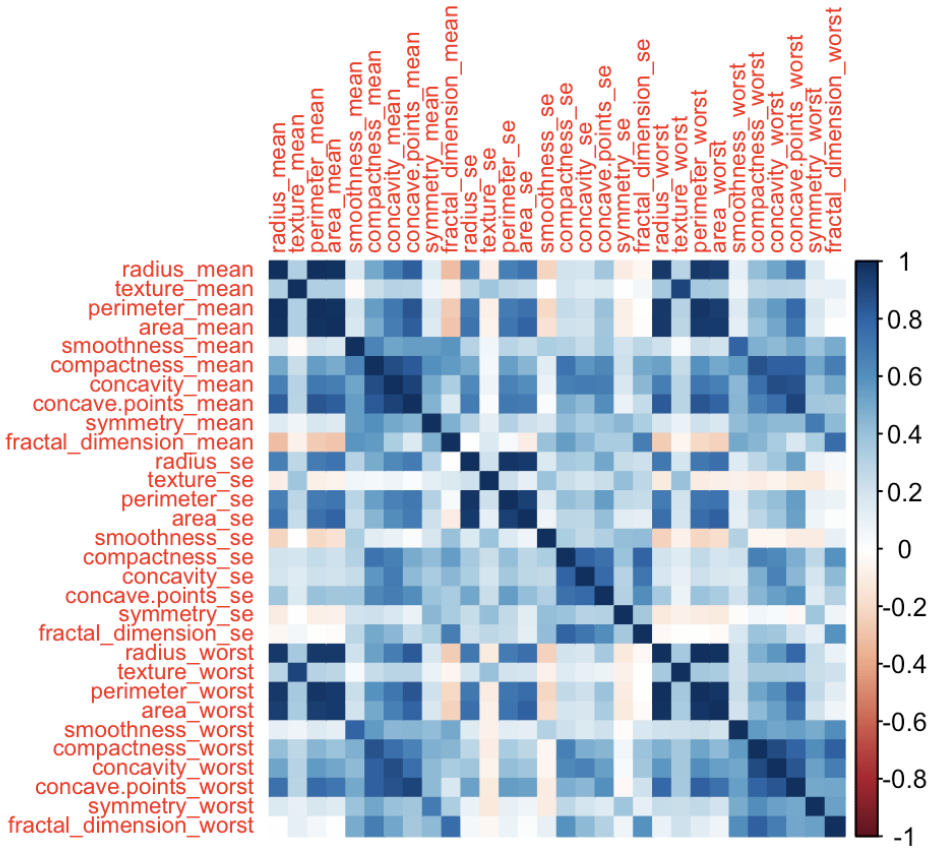

FIGURE 5 – Correlation matrix on the features

The first phenomen we observe is that all the values have their correlation greater than -0.4. The variables which have the lowest correlation ratio between them is the variable "fractal dimension mean" and "radius mean".

Here some variables which have a correlation ratio greater than 0.9 : radius mean/perimeter mean , radius mean/area mean , radius mean/radius worst , radius mean/perimeter worst , radius mean/area worst , texture worst/texture mean , area mean/radius worst , area mean/perimeter worst , area mean/area worst.

The high positive correlation between those variable can be explain geometrically. As those variables are geometric indicators of the breast mass, it is clear that their values are strongly related.

3 Diagnosis prediction (using different methods)

In this section, different model are used to predict the cancer diagnosis given features. Going from training, predicting to evaluation of performances to see which model seems the most efficient.

3.1 Logistic regression model using L2-regularization (Ridge)

Since the target variable is a binary variable (taking only two value : malignant or benign), it is more appropriate to use logistic model regression rather than a linear model regression. Moreover, since the section 2.4 showed that variables in the feature are highly correlated, the use of Ridge regression is coherent to estimate the coefficients.

3.1.1 First model evaluation

After building and training, the evaluation of the model (using cross-validation with 10 folds) gives the following results :

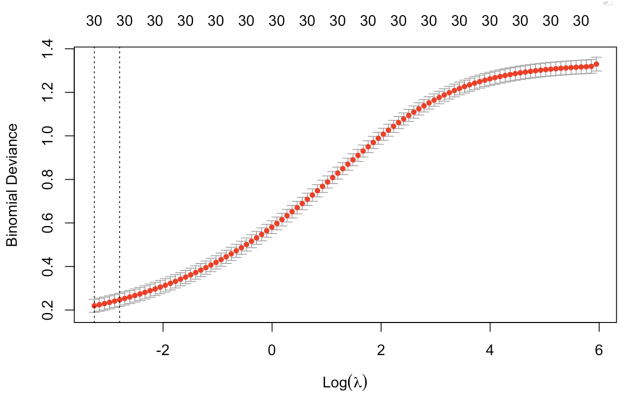

FIGURE 6 – Binomial deviance in function of $log(\lambda)$

```
Call:  cv.glmnet(x = X, y = y, family = "binomial", alpha = 0)

Measure: Binomial Deviance

     Lambda Index Measure      SE Nonzero
min 0.03837   100  0.2195 0.02978      30
1se 0.06109    95  0.2470 0.02994      30
```

FIGURE 7 – Result after cross validation

It shows that the best λ from cross-validation is $\lambda_{Ridge} \approx 0.3836$ and at that value, all variables in the features are included.

3.1.2 Model performance

After using the best lambda to retrain the model, the predictions give the following results :

```
        Predicted
Actual   0   1
     0 356   1
     1  10 202
```

FIGURE 8 – Confusion matrix

```
[1] "Accuracy: 0.98066783831283"
[1] "Precision: 0.995073891625616"
[1] "Recall: 0.952830188679245"
[1] "F1-score: 0.973493975903614"
```

FIGURE 9 – Indicators

Interpretation

The accuracy of 0.98 means that the model correctly predicted approximately 98.07 pourcent of the cases in the dataset.
A precision of approximately 99.51 pourcent means that when the model predicts a tumor as malignant (positive class), it is correct about 99.51 pourcent of the time. In other words, very few instances predicted as malignant were actually benign.
A recall of around 95.28 pourcent indicates that the model correctly identified about 95.28 pourcent of the actual malignant tumors in the dataset. It means that the model missed predicting some malignant cases (false negatives), leading to a lower recall.
With an F1-score of approximately 97.35 pourcent, it signifies a good balance between precision and recall. It indicates that the model performs well in terms of both minimizing false positives (high precision) and capturing actual positives (moderate recall).

Overall interpretation

The model shows an extremely high precision, indicating a very low rate of false positives, which is crucial in cancer diagnosis to avoid misclassifying benign cases as malignant.
The recall, while good, suggests that the model is missing some malignant cases (false negatives). Balancing precision and recall, the F1-score indicates a strong performance of the model in identifying malignant tumors while maintaining a low false positive rate.

3.2 Variable selection (Statistical approach)

3.2.1 Stepwise logistic regression

In order to select which variables in the feature are significant in the logistic model, we tried to use stepwise logistic regression. But after some convergence issues (the following warning appeared : "glt.fit algorithm did not converge"), the script returned the following result :

```
Call:
glm(formula = y ~ diagnosis, family = binomial, data = data_df)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.409e-06  -2.409e-06  -2.409e-06   2.409e-06   2.409e-06

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -26.57   18848.08  -0.001    0.999
diagnosisM     53.13   30878.44   0.002    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7.5144e+02  on 568  degrees of freedom
Residual deviance: 3.3011e-09  on 567  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

FIGURE 10 – Result of stepwise approach

Overall interpretation

In this output, most variables in the feature are not present, suggesting that many predictors are not statistically significant in predicting the diagnosis, which is absurd. In fact, due to high multicolinearity among variables in the feature, there are convergence issues because the model cannot differentiate the effects of highly correlated predictors.

3.2.2 Principal Component Analysis method

To manage the highly correlated variables in the feature, the PCA method help to reduce the dimensionality of the feature.

3.2.3 PCA's implementation

Step 1 : Data Centering PCA begins by centering the data, meaning it subtracts the mean of each feature from the dataset, ensuring that each feature has a mean of zero.

Step 2 : Covariance Matrix PCA calculates the covariance matrix of the centered data. This matrix represents the relationships between different variables, showcasing how they vary together.

Step 3 : Eigenvalue ("valeurs propres") Decomposition The covariance matrix is then decomposed into its eigenvectors and eigenvalues. Eigenvectors represent the directions (principal components) of the maximum variance in the data, while eigenvalues depict the magnitude of variance along those directions.

Step 4 : Selection of Principal Components The eigenvectors are ranked by their corresponding eigenvalues in descending order. The top principal components (eigenvectors) are selected based on the amount of variance they explain in the data.

Step 5 : Dimensionality Reduction Finally, the original data is projected onto the selected principal components, effectively reducing the dimensions while preserving as much of the original variance as possible.

3.2.4 Model performance

```
        Predicted
Actual   0   1
     0 355   2
     1   7 205
```

FIGURE 11 – Confusion matrix

```
Accuracy: 0.9841828
Precision: 0.9903382
Recall: 0.9669811
F1-score: 0.9785203
```

FIGURE 12 – Indicators

Overall interpretation

In this case, it suggests that approximately 98.4 pourcent of the predictions made by the model are correct. These metrics collectively suggest that the model performs quite well in making accurate predictions, especially for a binary classification task.

3.3 K-Nearest-Neighbour (KNN) algorithm

The K-Nearest-Neighbour algorithm is suited for the classification problem. Recall that K-NN classification output is a class membership. Here, the two class correspond to Malignant (M) and Benign (B).

3.3.1 KNN's implementation

The algorithm calculates the distances between the test samples and all training samples. It identifies the k nearest neighbors based on these distances. Finally, it predicts the class of the test sample based on the majority class among its k nearest neighbors.
Our implementation follow the following steps :
- Split the dataset into training and testing sets.
- Define a function for calculating Euclidean distance.
- Implements a KNN prediction function.
- Train and evaluate the model on the test set.

The value of k is $k = 5$, as this value suits to the performance of our computer. Indeed, trying a greater value make the execution time too long.

3.3.2 Model performance

```
predictions   B   M
          B 101   7
          M   6  56
```

FIGURE 13 – Confusion matrix

```
Accuracy: 0.9235294
Precision: 0.9351852 0.9032258
Recall: 0.9439252 0.8888889
F1-score: 0.9395349 0.896
```

FIGURE 14 – Indicators

Interpretation

The model achieved an accuracy of approximately 92.35 pourcent, indicating that it correctly predicted the class of around 92.35 pourcent of the samples in the test dataset.
For the first class (the positive class), precision is approximately 93.52 pourcent , indicating that when the model predicted this class, it was correct about 93.52 pourcent of the time. For the second class, the precision is around 90.32 pourcent.
For the first class, the recall is approximately 94.39 pourcent, indicating that the model identified about 94.39 pourcent of the actual instances of this class. For the second class, the recall is about 88.89 pourcent.
For the first class, the recall is approximately 94.39 pourcent, indicating that the model identified about 94.39 pourcent of the actual instances of this class. For the second class, the recall is about 88.89 pourcent.
For the first class, the F1-score is approximately 93.95 pourcent, balancing precision and recall for that class. For the second class, the F1-score is around 89.6 pourcent, similarly balancing precision and recall.

Overall interpretation

The model demonstrates strong predictive performance, with high accuracy and relatively balanced precision and recall across both classes. These metrics collectively suggest that the model is effective in distinguishing between the two classes in the dataset. However, the accuracy is low compared to PCA and Ridge, maybe because the cross validation was not used this time (due to computer performance limitation).

4 Conclusion

The variety of model used in this project shows that there is not a unique method of predicting a dataset, as each of them produce high performance indicator to predict the diagnosis result. The implementation using Principal Component Analysis has the highest accuracy, this shows that variable selection using statistical approach is very efficient in this project. However, due to the high multicolinearity between variables in the feature (due to the fact that they have strong geometrical relationship), not all regression/prediction methods is appropriate to use. (especially variable selection method such as stepwise logistic regression).