Mammogram Mass Prediction

NOVEMBER 23, 2020

FRANCESKA DORVAL

Table of Contents

Introduction

Breast cancer is the most common cancer diagnosed in women in several countries (Ref American Cancer Society). It is a cancer that forms in the cells of the breasts. Mammography screening for breast cancer is at present the most effective method in reducing breast cancer mortality. A lot of unnecessary surgery arises from false positives arising from mammogram results. This capstone aim to apply several supervised machine learning techniques to identify if a mass lesion is malign or benign. It will help any stakeholder (radiologist, doctor, patient) build a better way to interpret the mammogram results and improve a lot of lives.

Data

The UCI repository made public a dataset called mammographic masses. (source: https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass) It contains 961 instances of masses detected in mammograms, with the following attributes:

1. BI-RADS assessment of the breast density: 1 to 5 (ordinal).This attribute show how confident the severity classification is; it is not a "predictive" attribute
2. Age: patient's age in years (integer)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6. Severity: benign=0 or malignant=1 (binominal) Severity is the classification that our ML will attempt to predict based on those attributes above.

Methodology

This section define and present all steps that help us build our machine learning model. Our methodology contains four big steps: Data Preparation and Data Understanding, Data Modeling,Data Evaluation.

1.1 Data Preparation

This phase consist in multiple activities to construct the final dataset, then we select the attributes (columns) that will be used to train our machine learning model.

Those activities include the steps below:

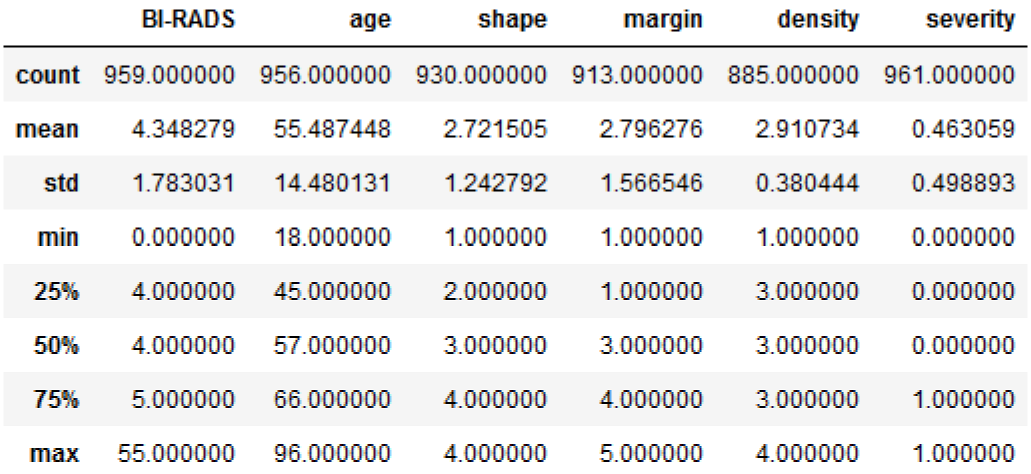
- Cleaning Data
- Add the appropriate column names

- Convert missing data (indicated by a ?)
- Drop every row that is missing data
- Convert the age column into discrete bucket value.

1.2 Data Analysis

After clearing the data, several useful information have been extract to help the user having a better understanding .Meaningful data visualizations have been presented to show patterns ,summarize the data to give better guidance to the user.

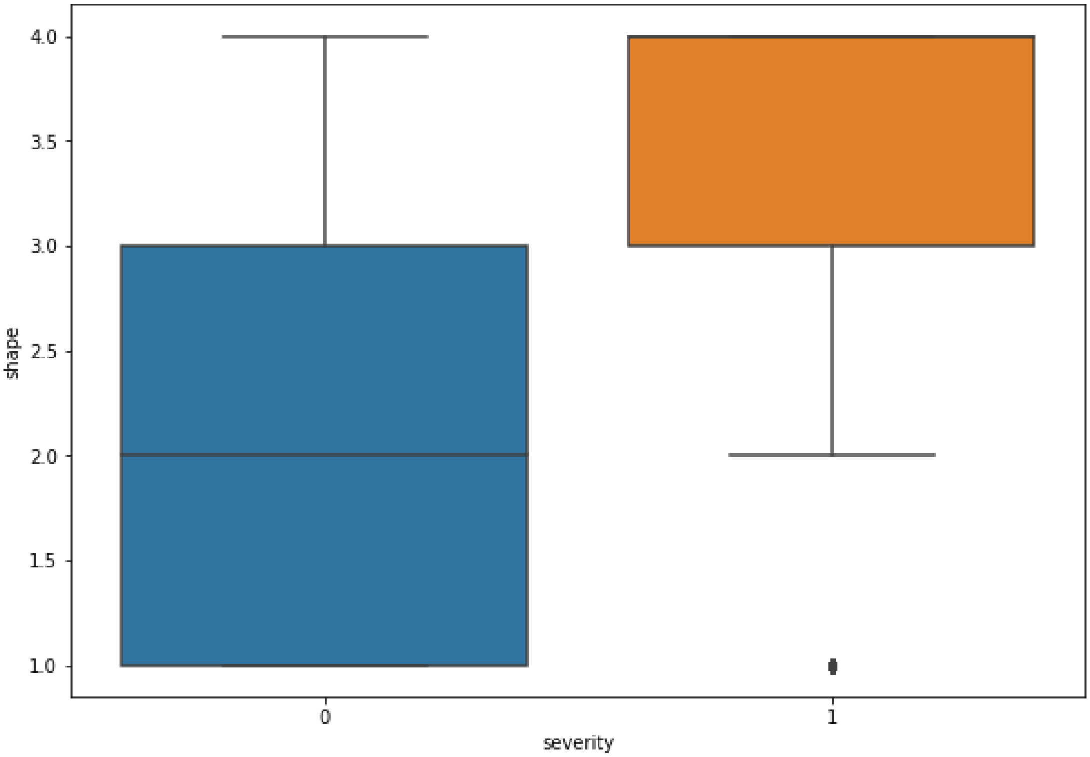
1) Summarization of data



	BI-RADS	age	shape	margin	density	severity
count	959.000000	956.000000	930.000000	913.000000	885.000000	961.000000
mean	4.348279	55.487448	2.721505	2.796276	2.910734	0.463059
std	1.783031	14.480131	1.242792	1.566546	0.380444	0.498893
min	0.000000	18.000000	1.000000	1.000000	1.000000	0.000000
25%	4.000000	45.000000	2.000000	1.000000	3.000000	0.000000
50%	4.000000	57.000000	3.000000	3.000000	3.000000	0.000000
75%	5.000000	66.000000	4.000000	4.000000	3.000000	1.000000
max	55.000000	96.000000	4.000000	5.000000	4.000000	1.000000

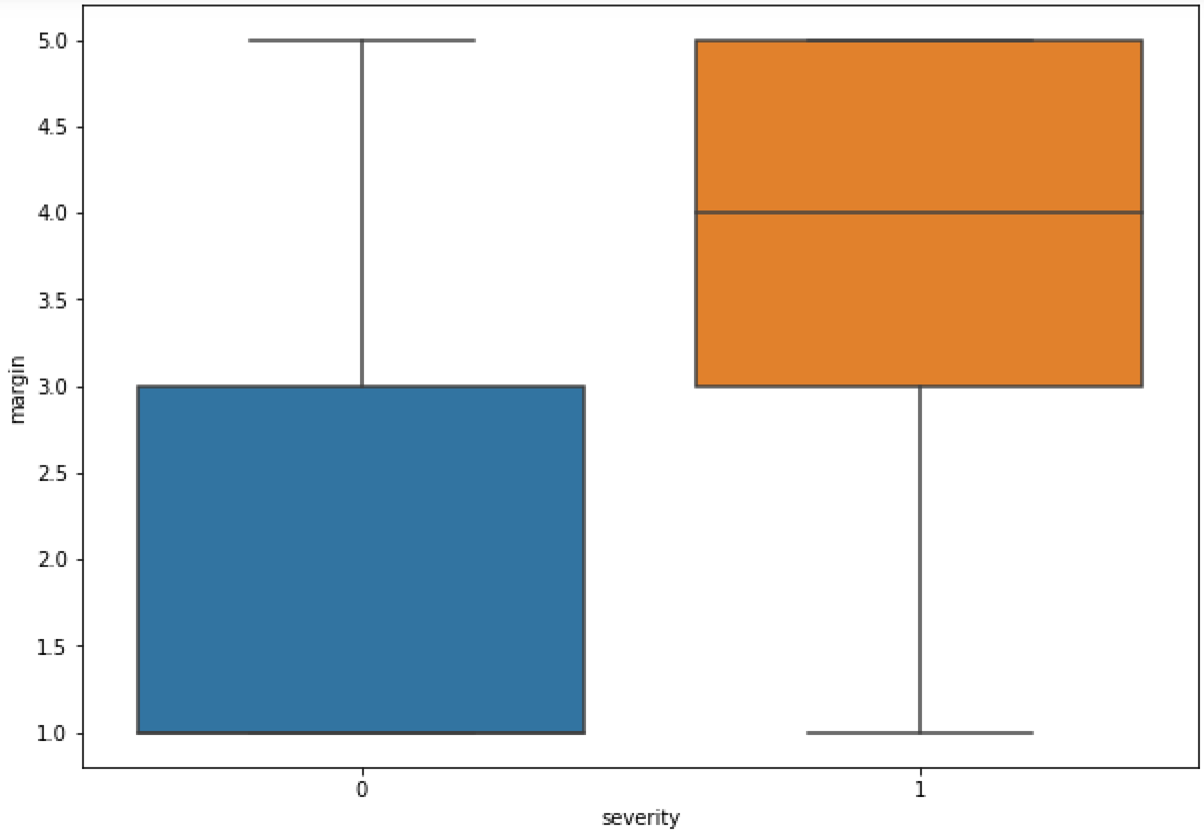
1.2.1 Graphical Representation

1)Mean severity by shape



The mean of the severity by shape : `3.503722`

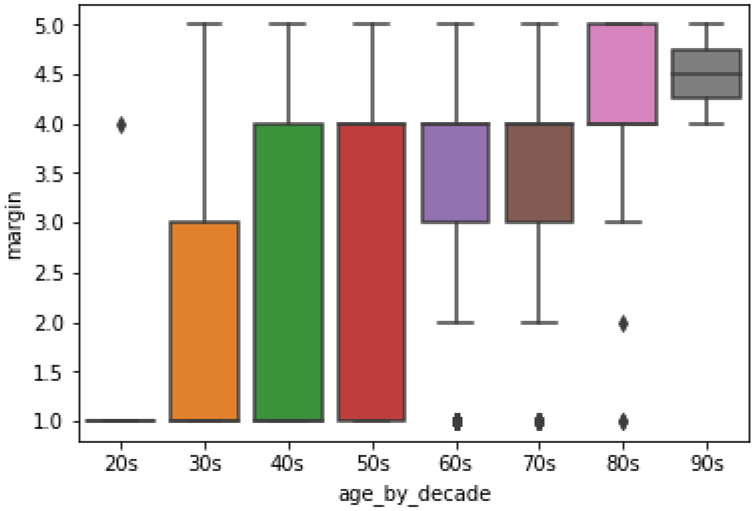
2）Mean severity by margin



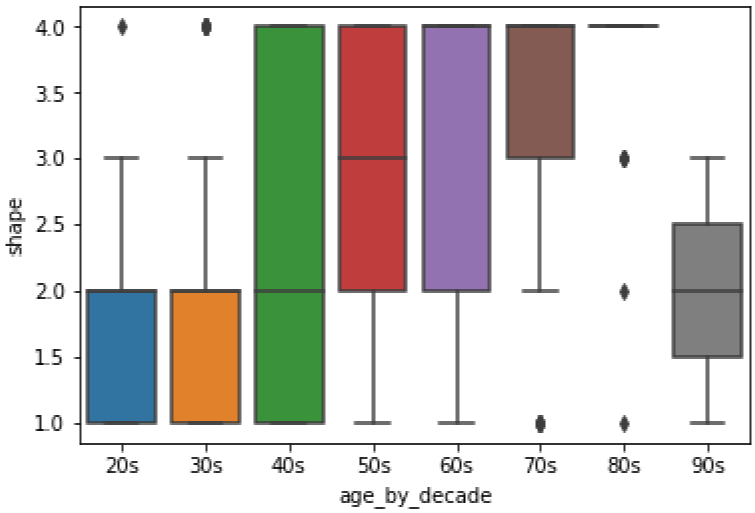
The mean of the severity by margin: 3.739454

3) The severity by margin by group of age

The data of age field have been have been converted in group by decade and try to see which sign of symptoms we can reveal:



4)The severity by shape by age



5) Details explanation

The analysis of the graphic above and the interpretation of the data show the age bracket where breast cancer severity start increasing is between 40 and 60. Most women who has severity of breast cancer are in their age bracket [60, 80[.

It shows as well some outliers: 1 Woman in age bracket of 20 who has breast cancer of severity 4.

3 women in age bracket of 90 who has breast cancer of severity 4.

It shows women in [70, 80] who don't present any sign of breast cancer.

Severity=1	**Total**
[20,40[	**14**
[40,60[	**164**
[60,80[	**226**
[80,100[	**36**
Grand Total	**440**

Total by severity

Severity	Age	Total
0	[20,100[	516
1	[20,100[	440
1	Age is null	5

Outlier

Age	Total with severity
28	1
]30,40[	13
[90,99[	3

Data modeling

Various algorithms and methods have be selected and applied to build the model .Before applying those algorithms, we have performed several steps to prepare the data in order to build the model. The models have been evaluated thoroughly to ensure that the business problem objectives are achieved.

The steps for modeling the data are the following

- Convert the Pandas dataframes into numpy arrays that can be used by scikit_learn.
- Create an array that extracts only the feature.age, shape, margin, and density
- Create an array that contains the classes (severity)
- Normalize the attribute data with preprocessing.StandardScaler().
- Train split test set aside 75% for training, and 25% for testing
- Use K-Fold CV to split our training set into K number of subsets, called folds.
- Several ML classification algorithm have been used to predict if the cancer is malign or benign.

The table below shows all classification algorithms to predict and method used to estimate the accuracy.

Algorithm	**Evaluation**
• Logistic Regression •	• cross_val_score
• KNN	• cross_val_score
• Naïve Baises	• cross_val_score
• DecisionTrees	• K-Fold cross validation •
• SVM with kernel (Linear,poly,rbf)	• cross_val_score
• RandomForestClassifier	• cross_val_score

Results

The table below summarize the result of each classification algorithm used to identify if a mass lesion is malign or benign.

For KNN algorithm we write a for loop with K values ranging from 1 to 100. The best performance we could get out of KNN start with K starting to 60.

SVM algorithm we used several hyperparameter .We try the rbf, sigmoid, and poly kernels and the rbf give the best-performing kernel.

For DecisionTree we use K-Fold cross validation to get a better measure of our model's accuracy.

Algoritm	**Accuracy**			
Logistic Regression	0.80735			
KNN	0.8036 for k= 60 until 100			
Naïve Baises	0.7844			
DecisionTrees	0.7373 for k=10			
SVM with kernel (Linear,poly,rbf)	Linear=0.7964	RBF=0.80**6**2	Sigmoid=0.7351	Poly=0.7927
RandomForestClassifier	0.7528 for n_Estimator=10			

Discussion

Aside all those algorithm DecisionTree and SVM with sigmoid kernel give the worst performance. LogisticRegression and SVM with rbf kernel can be used to predict breast cancer.

Conclusion

This study help to answer to the business problem. It give to stakeholder a different way to view the data, they can easily visualize the relationship between the attributes.

The relationship among between different attribute provided in the dataset have been analyzed. The feature BI-RAID have been excluded and the feature **age,shape,margin,density** have been selected among important feature to predict the breast cancer. After applying several ML classification algorithm, we can conclude the accuracy of the result can help any stakeholder (radiologist, doctor, patient) to take decision quickly and build a better way to interpret the mammogram results and improve a lot of lives. However other studies to predict breast cancer must be done using the breast images.