Hormones receptors status in Breast Cancer

Text mining and Sentiment Analysis

December 2022

Abstract. The project aims at enforcing classification algorithms in order to predict hormones receptors status in breast cancer. Learning whether a tumor has particular proteins helps doctors determine a person's risk of recurrence and give them the most beneficial treatment. Focusing on two main hormones, estrogen and progesterone, a Multioutput-multiclass classification approach was used and on human epidermal growth factor receptor 2 a multiclass classification was implemented. In the first section of the report I give some medical definitions and I describe the dataset, then I illustrate the pre-processing phase performed to clean the records and tokenize texts. In the last part I explain the implemented algorithms and show the results.
Keywords: hormones receptor, estrogen, progesterone, Natural Language Processing, Multioutput-multiclass classification, supervised learning.

Paola Serra ID 960547
Università degli Studi di Milano
Data Science and Economics

Contents

List of Figures

1 Introduction

The aim of the project is to predict hormones receptors status on a breast cancer dataset. Receptors are proteins in or on cells that can attach to certain substances in the blood. Normal breast cells and some breast cancer cells have receptors that attach to the hormones estrogen and progesterone, and need these hormones for the cells to grow. Keeping the hormones estrogen and progesterone from attaching to the receptors can help keep the cancer from growing and spreading and help doctors decide how to treat it.[1][2][3] Possible results with respective treatments are:

- Hormone receptor-positive: breast cancer cells have either estrogen (ER) or progesterone (PR) receptors or both. These breast cancers can be treated with hormone therapy drugs that lower hormones levels or block receptors.
- Hormone receptor-negative: breast cancers have no estrogen or progesterone receptors. Treatment with hormone therapy drugs is not helpful for these cancers.
- Triple-negative: breast cancer cells don't have estrogen or progesterone receptors and also don't make any or too much of the protein called HER2. Hormone therapy and drugs that target HER2 aren't helpful but Chemotherapy can.
- Triple-positive cancers: are ER-positive, PR-positive, and HER2-positive. These cancers can be treated with hormone drugs as well as drugs that target HER2.

First, classification algorithms were used to predict both ER and PGR levels, secondly we focused on classification of HER2 status knowing that the score will either be 0 to 1+ (HER2 negative), 2+ (borderline) or 3+ (HER2-positive). The fact of dividing a problem in two classification tasks reflects the presence of two tests for breast cancer: Immunohistochemistry (IHC), the most frequent initial test for determining HER2, along with other receptors (such as ER and PR), its accuracy can be markedly affected by technical issues (eg.accentuation by warm and cold tumor ischemia, the duration of tissue fixation in formaldehyde, the tissue-processing technique, and the embedding temperature of the heated paraffin wax); while Fluorescence in situ hybridization (FISH) overcomes IHC weakness. In order to choose the best models that predict this receptors levels, different classification algorithms were performed and their results compared.

2 Research question and methodology

2.1 Dataset

Data were taken from Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, IT). This dataset contains information on 74871 patient's diagnosis. In particular, it contains 16 variables, explained in detail below:

- cod paz: the code of the patient
- id: identification code
- numero caso: number that identifies the cancer's case

- data referto: date of medical report
- reparto: hospital unit
- diagnosi: diagnosis
- pezzo operatorio: specimen
- topografia: topografy of the area used for specimen
- summary: summary of breast cancer position
- gross: details about specimen (eg. dimension)
- er: Estrogen level
- pgr: Progesterone level
- her2: human epidermal growth factor receptor 2 level
- ki67: proliferation marker level
- fish: Fluorescence in situ hybridization is a test that "maps" the genetic material in a person's cells (eg. to see if the cells have extra copies of the HER2 gene) with results 'amplified' or 'not amplified'
- modificato: mostly null values

cod paz	id	numero caso	data referto	reparto	diagnosi	pezzo operatorio	topografia	summary	gross	er	pgr	her2	ki67	fish	modificato
74	2011055893	S2011-006911	12-09-2011	Degenze OCB Senologia	-NON EVIDENZA DI METASTASI A DUE LINFONODI ESA...	linf. sentinella ascella D	T-1963	NaN	Tessuto fibroadiposo da cui si isolano 2 linfo...	NaN	NaN	NaN	NaN	NaN	NaN
74	2011055891	S2011-006911	12-09-2011	Degenze OCB Senologia	CARCINOMA DELLA MAMMELLA DUTTALE IN SITU con a...	quadr. sup. mammella dx	T-1749	npl. mammella dx	Parenchima mammario di cm 5 x 4 x 4 con al tag...	[66,100]	[33,66)	3+	NaN	NaN	NaN
74	2011055892	S2011-006911	12-09-2011	Degenze OCB Senologia	-PARENCHIMA MAMMARIO RIFERIBILE A REGIONE AREO...	dotti capezzolo D	T-1740	NaN	Frammento di cm 1.	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1: Dataset preview

2.2 Data cleaning and Preprocessing

After an overview of the dataset, I decide to retain six variables: 'diagnosi', 'pezzo operatorio', 'summary', 'er', 'pgr', 'her2', since other variables do not add useful information for our research or they are mostly null (eg. fish 99 % of null). All the descriptive variables were concatenated in a final variable called 'text' and then it was performed the classical text classification pipeline: firstly, punctuation, special characters, numbers, multiple spaces and stop words (in Italian) removal; secondly, a transformation to lowercase; thirdly, a tokenization and finally an application of the Snowball stemmer, a stemming algorithm that reduces the word to its word stem that affixes to suffixes and prefixes or to roots of words known as a lemma. A detokenization at the end of our process was applied to obtain our column "text_final". In the code I reported some example that clarify how our text changed based on each transformation. Since the scope of our analysis is the prediction of er and pr levels, all records that contain empty or NaN values were discarded and our dataset was reduced

to 15000 records. The plots below represent the amount of records for each value for estrogen and progesterone as it was transcribed during the diagnosis:

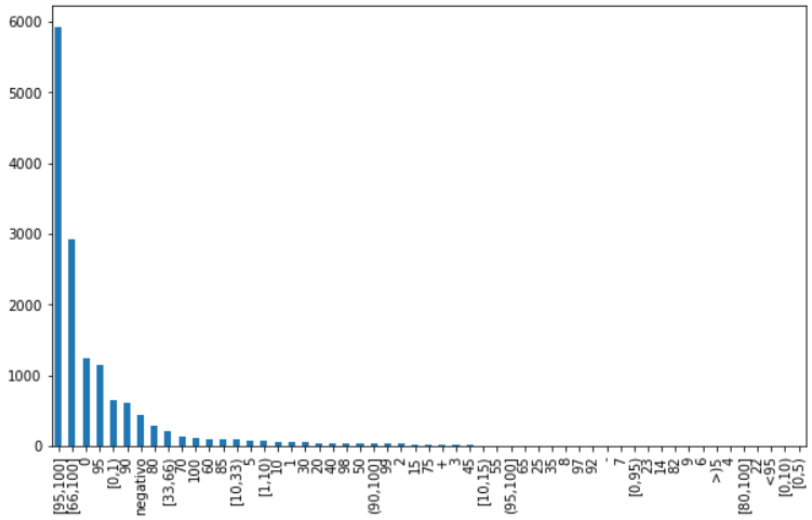


Figure 2: Estrogen levels

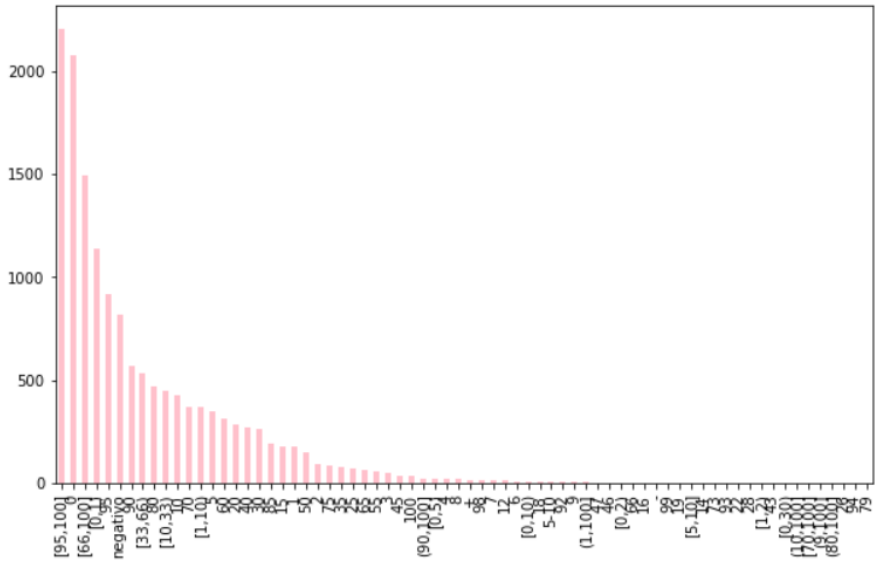


Figure 3: Progesterone levels

As we can see, the labels are unbalanced and they do not follow a standard (sometimes there is an interval while in other case a specific number or even string such as 'negativo'). I decide to divide the levels in three intervals of equal lenght ('1-33','33-66','66-100') and add another label for 0 or negative values. Plotting our data with the new labels we can see that the distribution is slightly improved, however we suffer from the fact that the majority of records have high level of hormones.

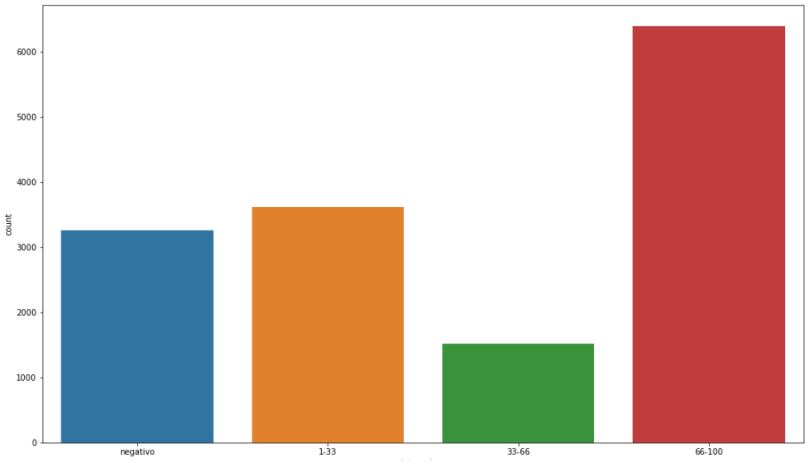


Figure 4: Progesterone distribution in intervals

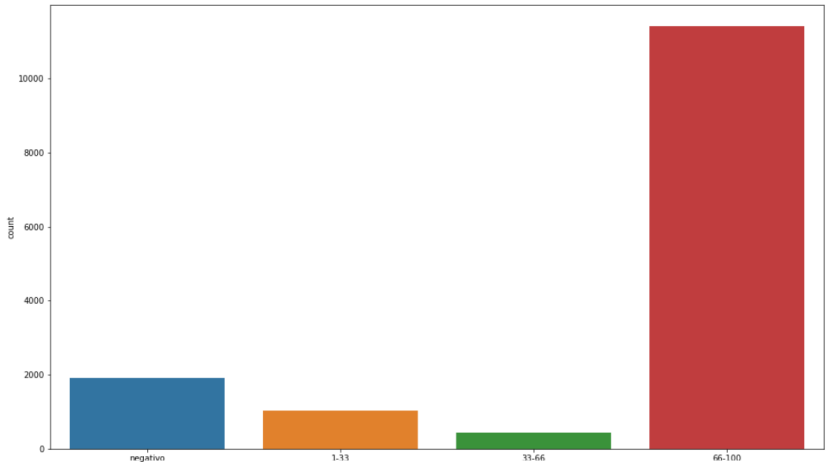


Figure 5: Estrogen distribution in intervals

Furthermore, the column 'her2' was divided in three classes according to the standard results (negative, borderline and positive) and representend in Figure 6. As we ex-

pected from literature, the mayority of case are ER+, PR+ and HER2- (Luminar A ∼65%), followed by triple negative (TNBC ∼20%), HER2 enriched (HER2 ∼10%) and tripLe positive (Luminar B ∼5%).

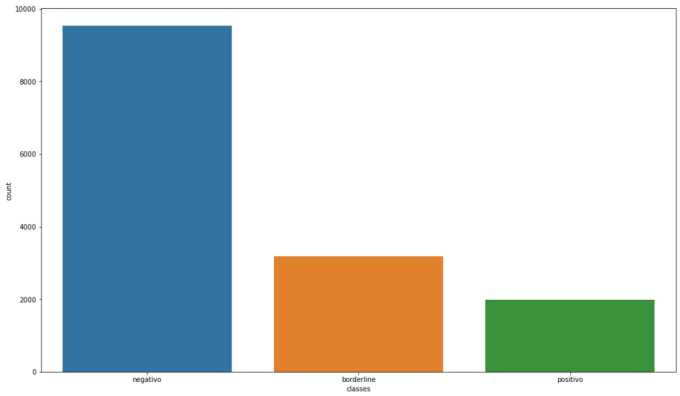


Figure 6: HER2 distribution in intervals

2.3 Learning procedure

Most of the classification or regression problem that one works with involves one target class label (dependent variable) and multiple independent features, but in this case there are two dependent variables (Estrogene and progesterone). Based on the number of dependent variables, we can define different classification problems:

	# of dependent variables	Cardinality of Target variable
Multi-Class Classification	1	> 2
Multi-Label Classification	> 1	2
Multi-Output Classification	> 1	> 2
Multi-Output Regression	> 1	Continuous

Figure 7: Classification type

Most of the classification machine learning algorithms are not able to handle multi-label classification. One needs to use a wrapper around them to train multi-label classification data. Scikit-learn comes up with 2 wrapper implementations:

- MultiOutputClassifier: This strategy fits one binary classifier per target. This wrapper can be used for estimators that do not support multi-target classification such as logistic regression.
- ChainClassifier: This strategy can be used when the classification targets are dependent on each other. In this strategy chain of binary estimators are trained with the independent features along with the prediction of the last estimator.

For each algorithm that I have used I built a general pipeline, defined as follow: first, the text was transformed to a matrix of token counts with CountVectorizer, secondly ,I normalize the result weighting it with TF-ID, thirdly I applied the algorithm passed before in MultiOutputClassifier. Also I randomized the search on hyper parameters in order to find the best ones and on the model based on that parameters I calculate the accuracy. Four algorithms were applied: XGB, Logistic Regression, KNN and SVC. XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In the figure below the evolution from decision trees to XGBoost is shown:

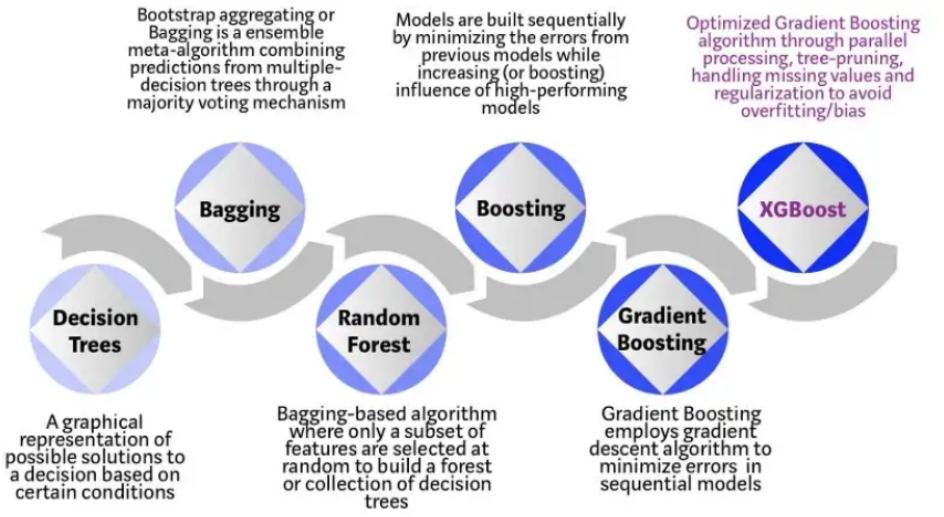


Figure 8: Evolution of XGBoost Algorithm from Decision Trees

Natively, Logistic Regression only supports binary classification,as a matter of fact the general expression for Logistic Regression is:

$$y = \frac{1}{1 + e^{-x}} \tag{1}$$

As you can see, the function range is between 0 and 1. When we fit the model to the logistic function we will change the x value with the coefficients of our data parameters and get a probability of our outcome being closer to 1, or closer to 0. However there are two option to "adapt" this model to multi-class problems: One-vs-Rest, that compare one class against all other classes or Multinomial Logistic Regression. It (basically) works in the same way as binary logistic regression. The analysis breaks the outcome variable down into a series of comparisons between categories and logistic transformation of the odds serves as the depending variable. The C value in Logistic Regression is an user adjustable parameter that controls regularisation. In simple terms, higher

values of C will instruct our model to fit the training set as best as possible, while lower C values will favour a simple models with coefficients closer to zero. K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data.

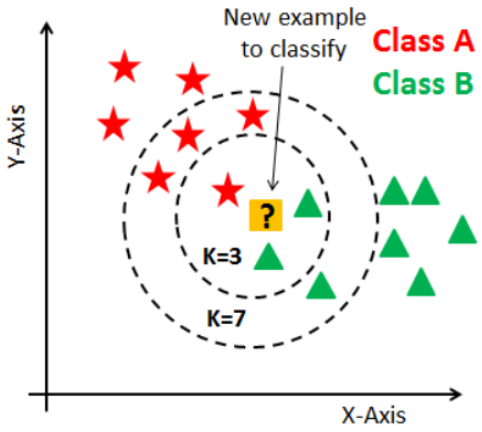


Figure 9: K-nearest neighbours example

Support Vector Machine is responsible for finding the decision boundary to separate different classes and maximize the margin in the hyperplane. Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform this classification. While on the HER2 classification problem I applied Multinomial Naïve Bayes Classifiers, that aims to assign fragments of text to classes by determining the probability that a text belongs to the class of other texts, having the same subject. It is based on the Bayes Theorem:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{2}$$

3 Experimental results

To evaluate model performance comprehensively, since our dataset is very unbalanced we should examine both precision and recall instead of relying on accuracy. The F1 score serves as a helpful metric that considers both of them:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

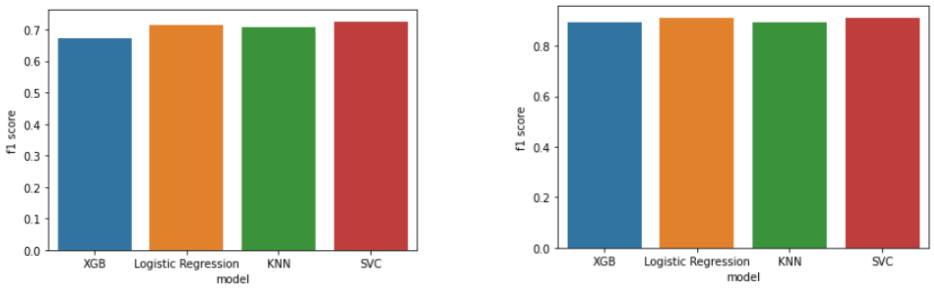
where precision can be seen as the number of predictions that are truly positive with respect to the number of all the positive predictions I made:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

and recall can be seen as the number of predictions that are truly positive with respect to the correctly positive predictions and the negative predictions that are actually positive:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

In the case of multi-class classification, it is common to adopt averaging methods for F1 score calculation, resulting in a set of different average scores (macro, weighted, micro) in the classification report. In particular, the macro-averaged F1 score is computed using the arithmetic mean of all the per-class F1 scores; the weighted-averaged F1 score is calculated by taking the mean of all per-class F1 scores while considering each class's support relative to the sum of all support values and the Micro averaging computes a global average F1 score by counting the sums of the True Positives (TP), False Negatives (FN), and False Positives (FP). Remember that in our Estrogen and progesterone classification problem, classes between 66 and 100 (both of er and pr) contain mostly of records, same for class 'negativo' in our HER2 problem, due to this I decide to use the weighted F1 score.
In the figure below I show the weighted F1 score for each algorithm that have been applied:



(a) Progesterone

(b) Estrogen

Figure 10: F1 score weighted for PR and ER

As we can see and expect, the Estrogen F1 score is quite high (above 85%) from all the algorithm applied, while in Progesterone this metrics is around 70%. Going deeper with a classification matrix, I discover that the metric, both in Progesterone and Estrogen, was strongly influenced by the class '33-66' that has the lowest support overall. The algorithm that performs better in term of f1-score for the classes with lowest support was KNN, with weights distance and 1 neighbor. Overall the best

accuracy was achieved by LinearSVC, secondly by Logistic Regression, thirdly KNN and XGBoost.
For HER2 classification, instead best results were reached by SVC and Logistic regression. Below the confusion matrix computed using LinearSVC:

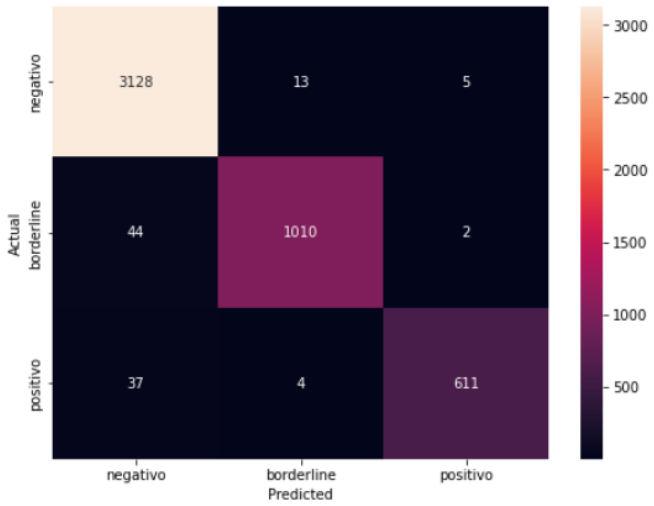


Figure 11: Confusion Matrix in Her2 prediction

4 Concluding remarks

Starting from diagnoses scratch and hormones levels, I tried to build models that are able to generate Progesterone and Estrogen levels and HER2 status. These predictions are very useful for doctors, since they are able to give a probability of recurrency's risk and to determine the right therapy for patients. Two different classification problems were explored: Multioutput-multiclass for ER and PR and Multiclass for HER2. The decision of dividing a problem in two sub-problems reflect the existency of two tests for breast cancer: IHC (Immunohistochemistry),the most frequent initial test and FISH, the most accurate test for HER2. The fact that the data was unbalanced is confirmed by literature and also confirmed by our final results: the mayority of breat cancer has high values in Estrogen, Progesterone and low values in Humanan epidermal growth factor. As expected, in lowest category (by number of records) the models has worst performance but in highest category reach satisfying results. For Estrogen and Progesterone status I recommend the KNN algorithm if we want to predict better more rare types of breast cancer, otherwise while I will choose between LinearSVC and Logistic regression, also for HER2. To improve this work, a first attempt in collecting more diagnoses of the least categories and uniforming the levels in classes can be made. Secondly, a different combination of text can be used, eg. instead of using the concatenation of "diagnosi", "pezzo operatorio" and "summary", one can add

also ”gross” texts (info about dimension of cancer) and discard some other columns. Third, It is possible to try to predict simultaneously all the three hormones levels in one step . Fourth, an exhaustive parameter GridSearch of the algorithms used and/or the application of other algorithms could lead to better results. The code of the project is available on Github.

References

[1] "https://academic.oup.com/clinchem/article/57/7/980/5621098"

[2] "https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html"

[3] "https://www.pennmedicine.org/cancer/types-of-cancer/breast-cancer/types-of-breast-cancer/her2-positive-breast-cancer"

[4] "https://scikit-learn.org/stable/modules/classes.htmlmodulesklearn.multioutput"