Text minig for Biomolecular text

Applying a multilevel and multioutput model & Unsupervised Machine learning solution

Aim of the project

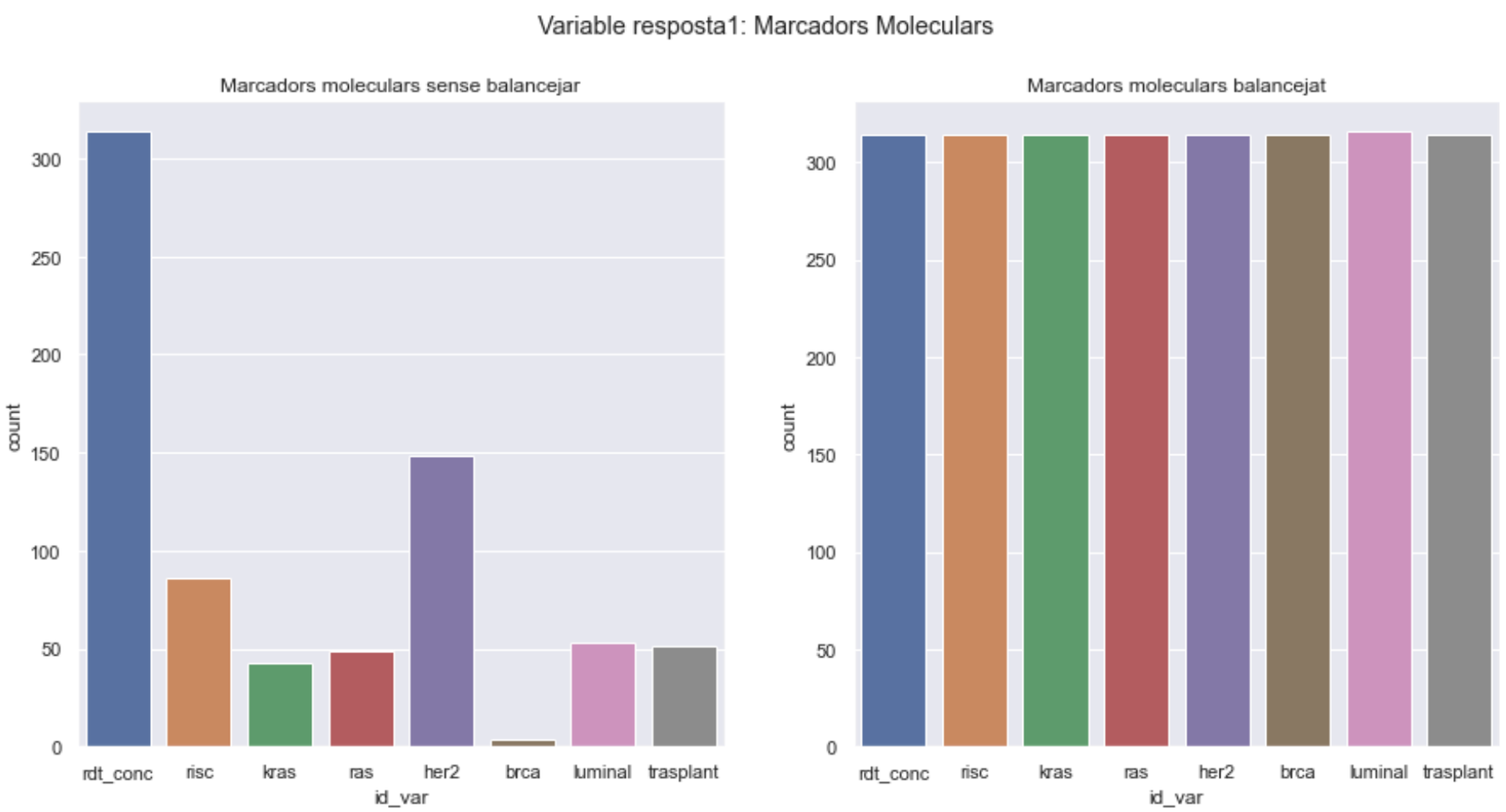
- **Problem**: The text consists of a series of classifications and protocols that are in the ESPOQ data source when a chemotherapy treatment is administered to the patient.

- **Aim of the project:** To **extract biomarcators** form biomolecular text.

Descriptive data

We have realised that we have unbalanced data. So we need to treat this data in order to avoid the overestimation in our prediction.

In the left position, we have unbalanced data and the balanced data in the right.



About this model

We develop a multilevel and multioutput model with crossvalidation because of unbalanced data.

We also tested diferent models as **MultinomialNB** and **Random Forest** but finally we realised that SGDClassifier technique was the best. With a good precision of around **90%.**

```
##predictions

from sklearn.model_selection import cross_val_score, cross_val_predict
from sklearn import metrics

predictions=cross_val_predict(loaded_model, loaded_vectorizer.transform(xtest.ravel()),ytest, cv=4)


for i in range(len(y[1])):
    print('Score of',text[i],':')
    print(metrics.classification_report(ytest[:,i],predictions[:,i]))
```

	precision	recall	f1-score	support
al.lògenic	0.96	1.00	0.98	53
alt	0.96	0.92	0.94	24
alt - mig	1.00	0.43	0.60	7
autòleg	0.92	1.00	0.96	11
baix	1.00	0.62	0.76	13
baix - mig	0.80	0.67	0.73	6
independent	1.00	1.00	1.00	6
mig	1.00	0.80	0.89	5
mutat	0.97	0.97	0.97	119
negatiu	0.78	0.88	0.83	33
no	0.85	0.69	0.76	81
no sobreexpresio	1.00	1.00	1.00	1
positiu	0.78	1.00	0.88	32
si	0.91	0.86	0.89	50
wild type	0.81	1.00	0.89	62
accuracy			0.89	503