Data Glacier

Your Deep Learning Partner

Healthcare Persistency of a drug

Exploratory Data Analysis

Jan 30th, 2023.

Submitted by Nahari Terena
LISUM15 – Data Science Specialization

Outline

- Business Understanding
- Decision Tree
- Random Forest
- Logistic Regression
- SVM
- Conclusion
- App concept

Data Glacier

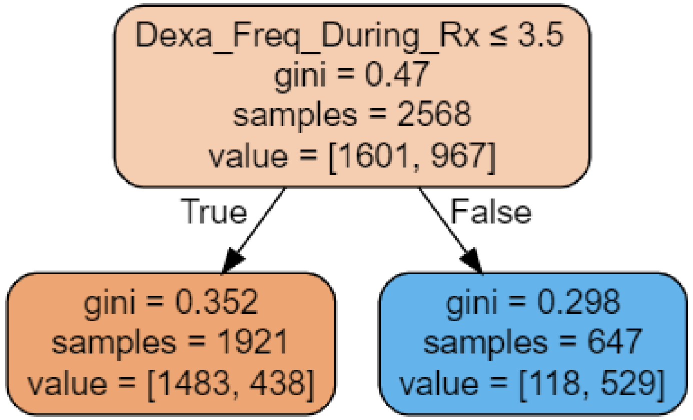
Your Deep Learning Partner

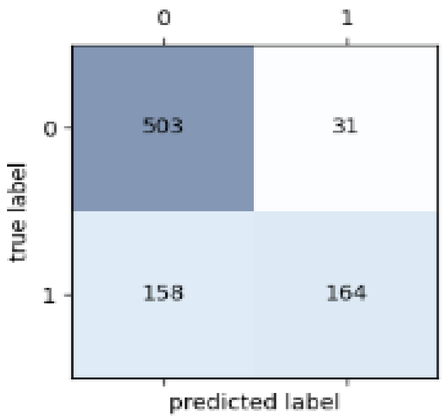
Business Understanding

- ABC Pharma contacted us to analyze the patients' data to have a better understanding of the factors that significantly impact the persistence of their drug. The aim is to know if a patient, based on private information, will follow the prescription and continue taking the medication for all the treatment time or not.
  - We aim to develop a web app to predict if a patient will get or not a drug schedule.

Decision Tree

- In the first scenario, the best parameter was a tree of depth 1. Therefore, the results were:
  - Accuracy of 78.35% for training data.
  - Accuracy of 77.92% for testing data.



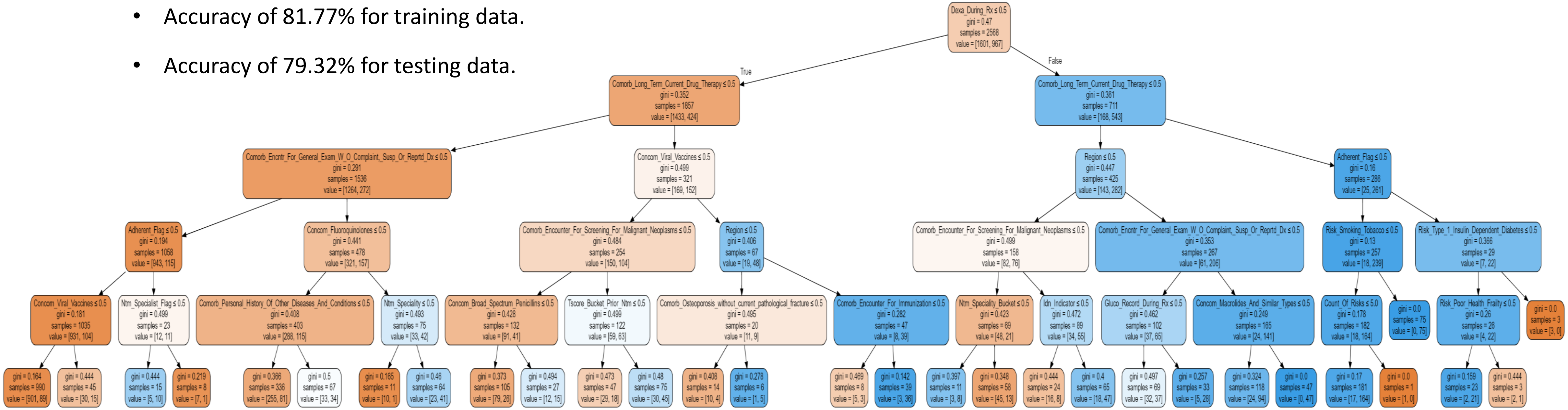


	0	1
0	503	31
1	158	164

- “Dexa_Freq_During_Rx” is the variable that has the most value in terms of predictive power. It could be interesting to see if we can get similar results without it. Considering “Dexa_Freq_During_Rx” is kind of another type of treatment. It would be interesting to be able to predict persistence without it.

Decision Tree

- In the second scenario, the best parameter was a tree of depth of 5. Therefore, the results were:
  - Accuracy of 81.77% for training data.
  - Accuracy of 79.32% for testing data.

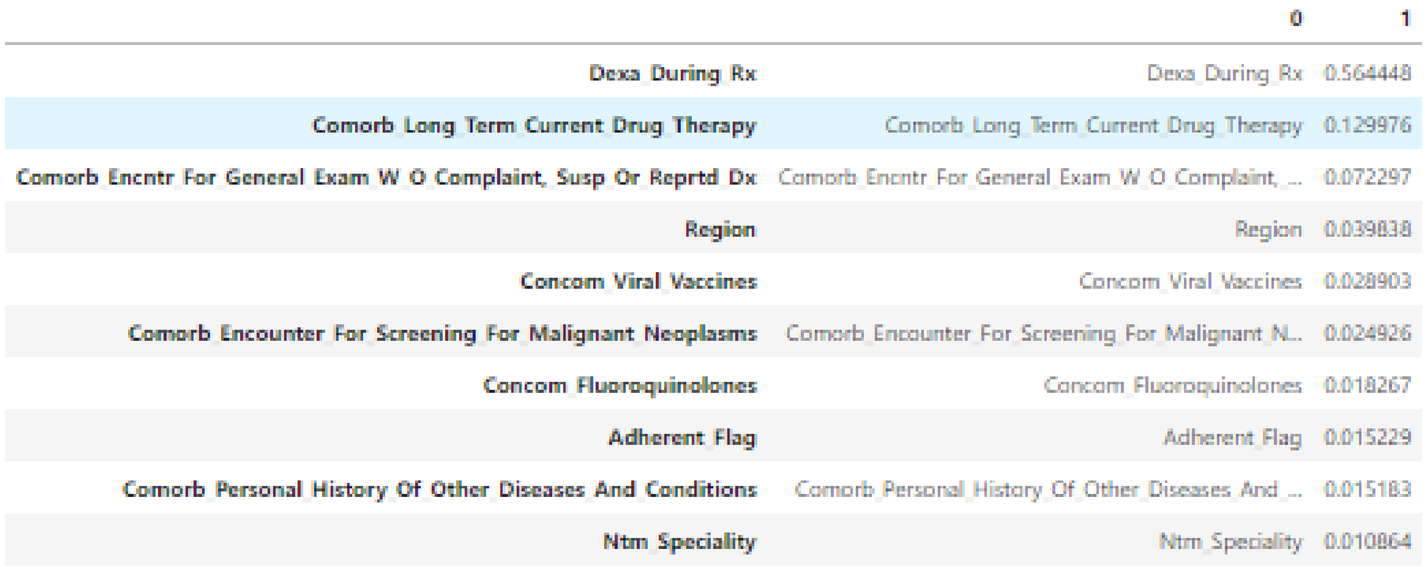


Dexa_During_Rx ≤ 0.5 gini = 0.47 samples = 2568 value = [1601, 967]
True
False
Comorb_Long_Term_Current_Drug_Therapy ≤ 0.5 gini = 0.352 samples = 1857 value = [1433, 424]
Comorb_Long_Term_Current_Drug_Therapy ≤ 0.5 gini = 0.361 samples = 711 value = [168, 543]
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx ≤ 0.5 gini = 0.291 samples = 1536 value = [1264, 272]
Concom_Viral_Vaccines ≤ 0.5 gini = 0.499 samples = 321 value = [169, 152]
Region ≤ 0.5 gini = 0.447 samples = 425 value = [143, 282]
Adherent_Flag ≤ 0.5 gini = 0.16 samples = 286 value = [25, 261]
Adherent_Flag ≤ 0.5 gini = 0.194 samples = 1058 value = [943, 115]
Concom_Fluoroquinolones ≤ 0.5 gini = 0.441 samples = 478 value = [321, 157]
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms ≤ 0.5 gini = 0.484 samples = 254 value = [150, 104]
Region ≤ 0.5 gini = 0.406 samples = 67 value = [19, 48]
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms ≤ 0.5 gini = 0.499 samples = 158 value = [82, 76]
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx ≤ 0.5 gini = 0.353 samples = 267 value = [61, 206]
Risk_Smoking_Tobacco ≤ 0.5 gini = 0.13 samples = 257 value = [18, 239]
Risk_Type_1_Insulin_Dependent_Diabetes ≤ 0.5 gini = 0.366 samples = 29 value = [7, 22]
Concom_Viral_Vaccines ≤ 0.5 gini = 0.181 samples = 1035 value = [931, 104]
Ntm_Specialist_Flag ≤ 0.5 gini = 0.499 samples = 23 value = [12, 11]
Comorb_Personal_History_Of_Other_Diseases_And_Conditions ≤ 0.5 gini = 0.408 samples = 403 value = [288, 115]
Ntm_Speciality ≤ 0.5 gini = 0.493 samples = 75 value = [33, 42]
Concom_Broad_Spectrum_Penicillins ≤ 0.5 gini = 0.428 samples = 132 value = [91, 41]
Tscore_Bucket_Prior_Ntm ≤ 0.5 gini = 0.499 samples = 122 value = [59, 63]
Comorb_Osteoporosis_without_current_pathological_fracture ≤ 0.5 gini = 0.495 samples = 20 value = [11, 9]
Comorb_Encounter_For_Immunization ≤ 0.5 gini = 0.282 samples = 47 value = [8, 39]
Ntm_Speciality_Bucket ≤ 0.5 gini = 0.423 samples = 69 value = [48, 21]
Idn_Indicator ≤ 0.5 gini = 0.472 samples = 89 value = [34, 55]
Gluco_Record_During_Rx ≤ 0.5 gini = 0.462 samples = 102 value = [37, 65]
Concom_Macrolides_And_Similar_Types ≤ 0.5 gini = 0.249 samples = 165 value = [24, 141]
Count_Of_Risks ≤ 5.0 gini = 0.178 samples = 182 value = [18, 164]
gini = 0.0 samples = 75 value = [0, 75]
Risk_Poor_Health_Frailty ≤ 0.5 gini = 0.26 samples = 26 value = [4, 22]
gini = 0.0 samples = 3 value = [3, 0]
gini = 0.164 samples = 990 value = [901, 89]
gini = 0.444 samples = 45 value = [30, 15]
gini = 0.444 samples = 15 value = [5, 10]
gini = 0.219 samples = 8 value = [7, 1]
gini = 0.366 samples = 336 value = [255, 81]
gini = 0.5 samples = 67 value = [33, 34]
gini = 0.165 samples = 11 value = [10, 1]
gini = 0.46 samples = 64 value = [23, 41]
gini = 0.373 samples = 105 value = [79, 26]
gini = 0.494 samples = 27 value = [12, 15]
gini = 0.473 samples = 47 value = [29, 18]
gini = 0.48 samples = 75 value = [30, 45]
gini = 0.408 samples = 14 value = [10, 4]
gini = 0.278 samples = 6 value = [1, 5]
gini = 0.469 samples = 8 value = [5, 3]
gini = 0.142 samples = 39 value = [3, 36]
gini = 0.397 samples = 11 value = [3, 8]
gini = 0.348 samples = 58 value = [45, 13]
gini = 0.444 samples = 24 value = [16, 8]
gini = 0.4 samples = 65 value = [18, 47]
gini = 0.497 samples = 69 value = [32, 37]
gini = 0.257 samples = 33 value = [5, 28]
gini = 0.324 samples = 118 value = [24, 94]
gini = 0.0 samples = 47 value = [0, 47]
gini = 0.17 samples = 181 value = [17, 164]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.159 samples = 23 value = [2, 21]
gini = 0.444 samples = 3 value = [2, 1]

- “Dexa_Freq_During_Rx” is the variable that has the most value in terms of predictive power. It could be interesting to see if we can get similar results without it. Considering “Dexa_Freq_During_Rx” is kind of another type of treatment. It would be interesting to be able to predict persistence without it.

Decision Tree

Besides “Dexa_Freq_During_Rx” and “Dexa_During_RX”, in order to predictive power of persistence in taking the drug is Comorbidity factors, Region and viral vaccines.

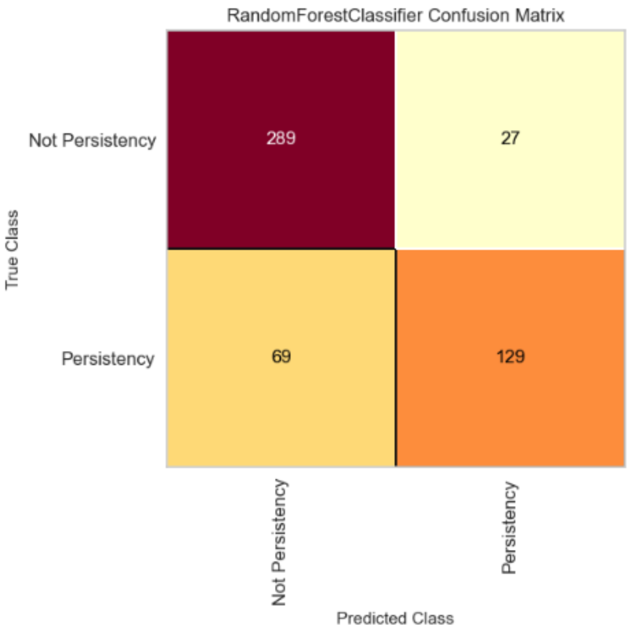


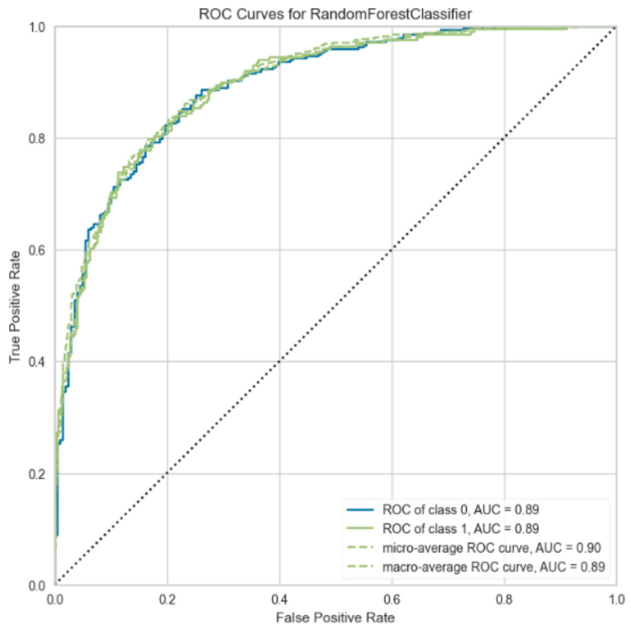
	0	1
Dexa_During_Rx	Dexa_During_Rx	0.564448
Comorb_Long_Term_Current_Drug_Therapy	Comorb_Long_Term_Current_Drug_Therapy	0.129976
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	Comorb_Encntr_For_General_Exam_W_O_Complaint,_...	0.072297
Region	Region	0.039838
Concom_Viral_Vaccines	Concom_Viral_Vaccines	0.028903
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	Comorb_Encounter_For_Screening_For_Malignant_N...	0.024926
Concom_Fluoroquinolones	Concom_Fluoroquinolones	0.018267
Adherent_Flag	Adherent_Flag	0.015229
Comorb_Personal_History_Of_Other_Diseases_And_Conditions	Comorb_Personal_History_Of_Other_Diseases_And_...	0.015183
Ntm_Speciality	Ntm_Speciality	0.010864

Random Forest

Using the chi-square statistic, some variables were eliminated.

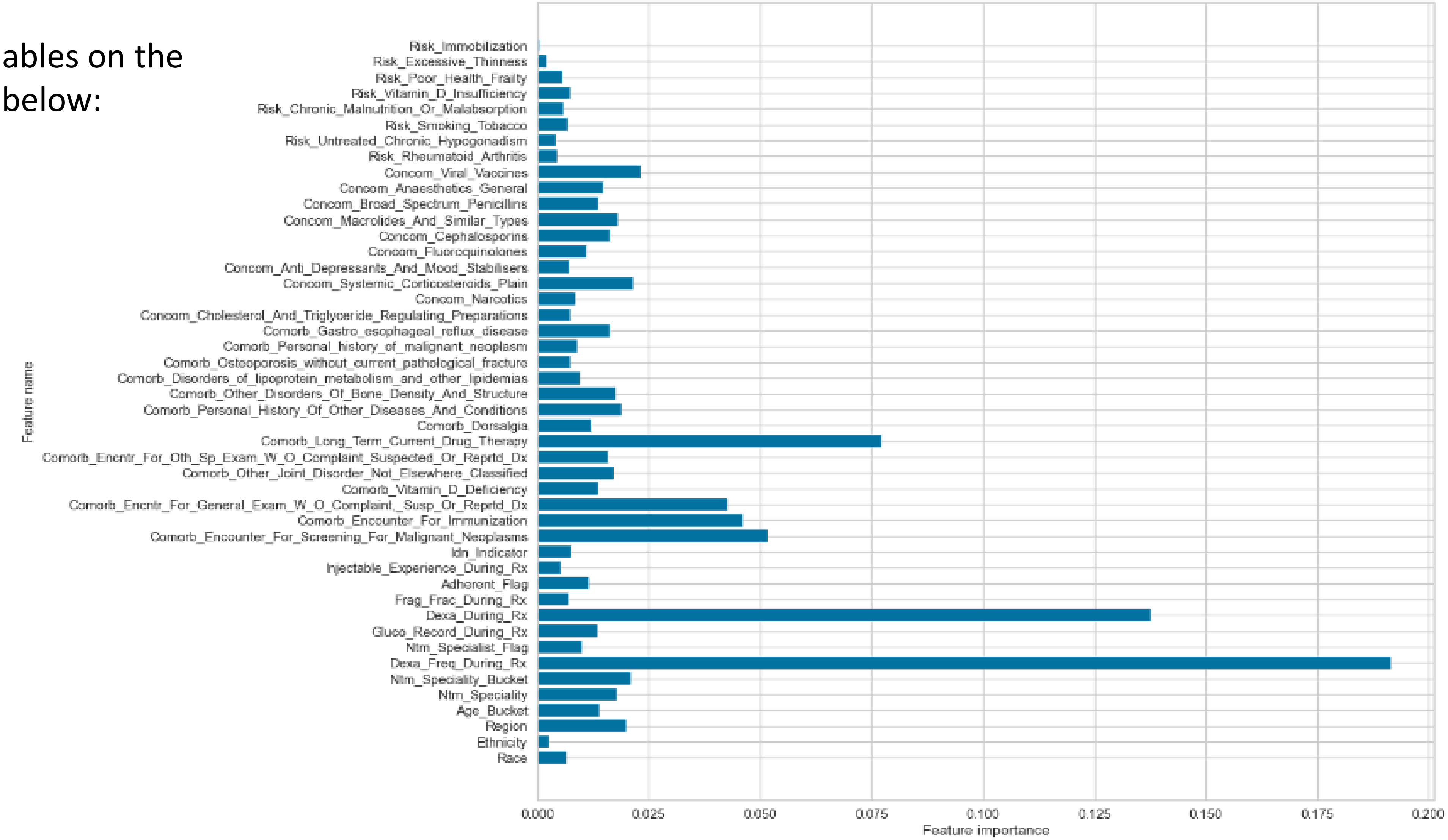
- Accuracy of 88.52% for training data.
- Accuracy of 80.74% for testing data.





Random Forest

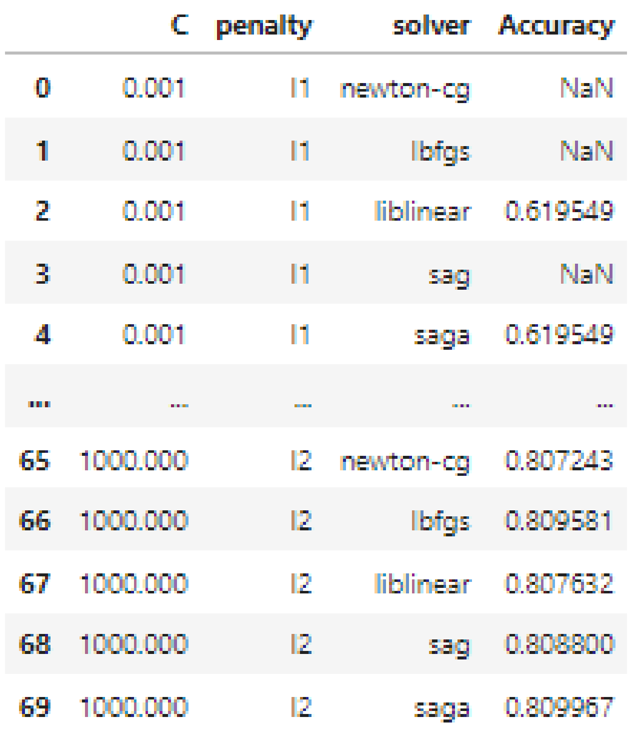
The influence of the variables on the target variable is shown below:



Logistic Regression

Differents parameters and their values when using exhaustive search (GridSearchCV) for tuning the hyperparameters.

The bottom screenshot displays the different combinations of set hyperparameters and the resulting accuracy when using GridSearchCV. Next slide the best parameters and score will be shown.

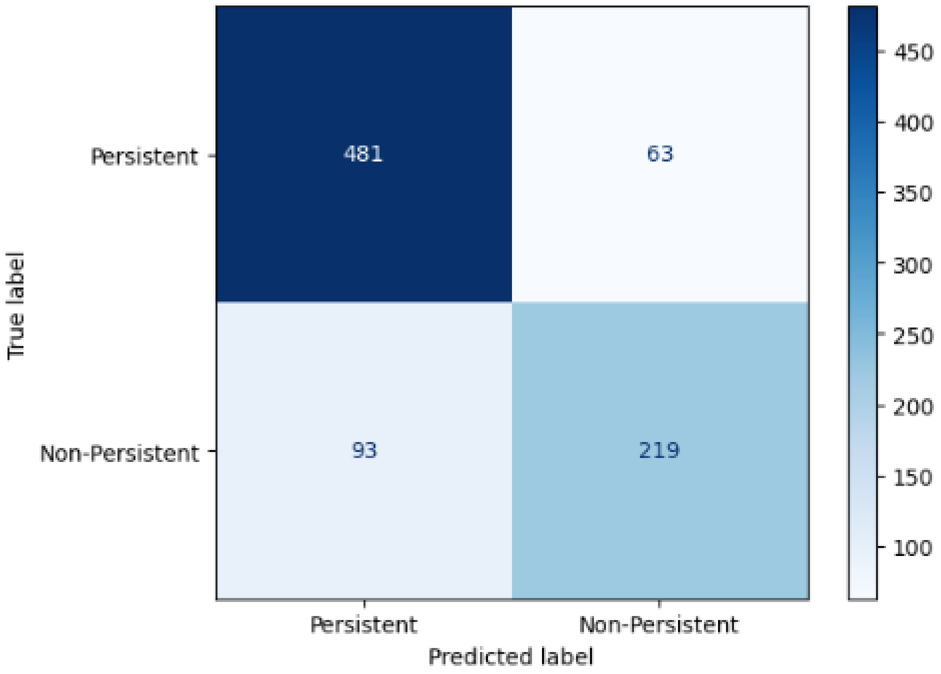


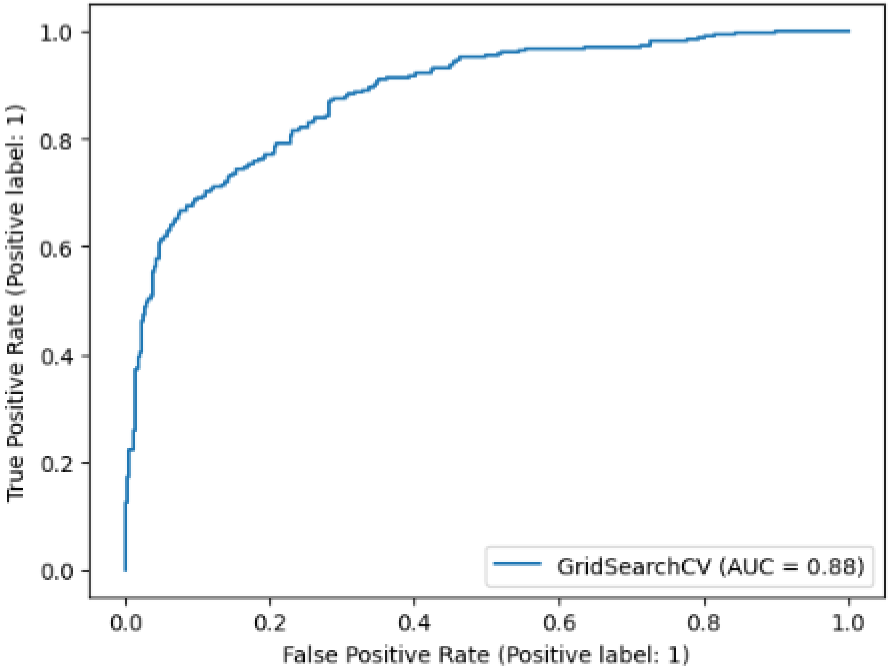
	C	penalty	solver	Accuracy
0	0.001	l1	newton-cg	NaN
1	0.001	l1	lbfgs	NaN
2	0.001	l1	liblinear	0.619549
3	0.001	l1	sag	NaN
4	0.001	l1	saga	0.619549
...	...	...	...	...
65	1000.000	l2	newton-cg	0.807243
66	1000.000	l2	lbfgs	0.809581
67	1000.000	l2	liblinear	0.807632
68	1000.000	l2	sag	0.808800
69	1000.000	l2	saga	0.809967

Logistic Regression

F1 score is used to evaluate the logistic regression model. Along with the confusion matrix and ROC curve.

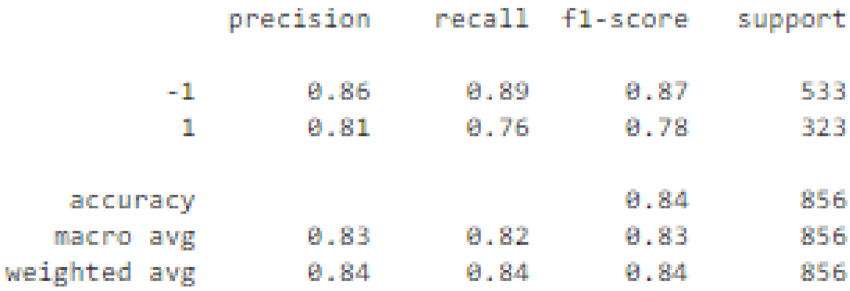
F1-score: 0.8156





Support Vector Machine

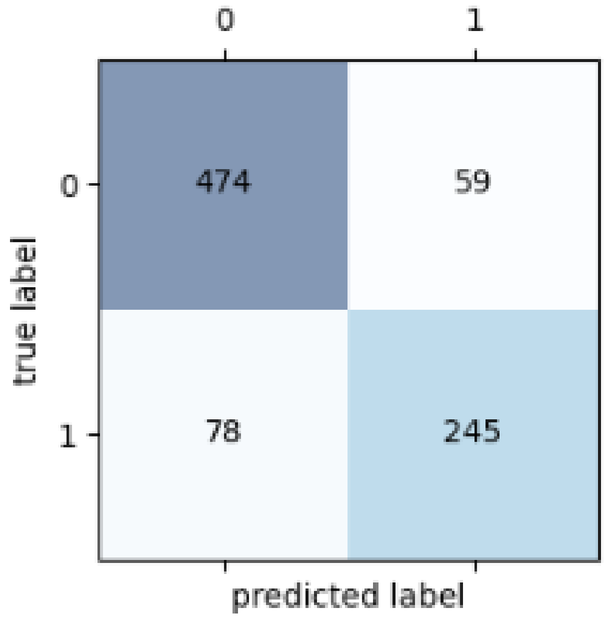
Support Vector Machines algorithm to classify the persistence of patients (1 for positives and -1 for negatives). A linear kernel has been used, obtaining an accuracy of 84.0% over testing data (25 % out of the whole dataset).



```
              precision    recall  f1-score   support

          -1       0.86      0.89      0.87       533
           1       0.81      0.76      0.78       323

    accuracy                           0.84       856
   macro avg       0.83      0.82      0.83       856
weighted avg       0.84      0.84      0.84       856
```

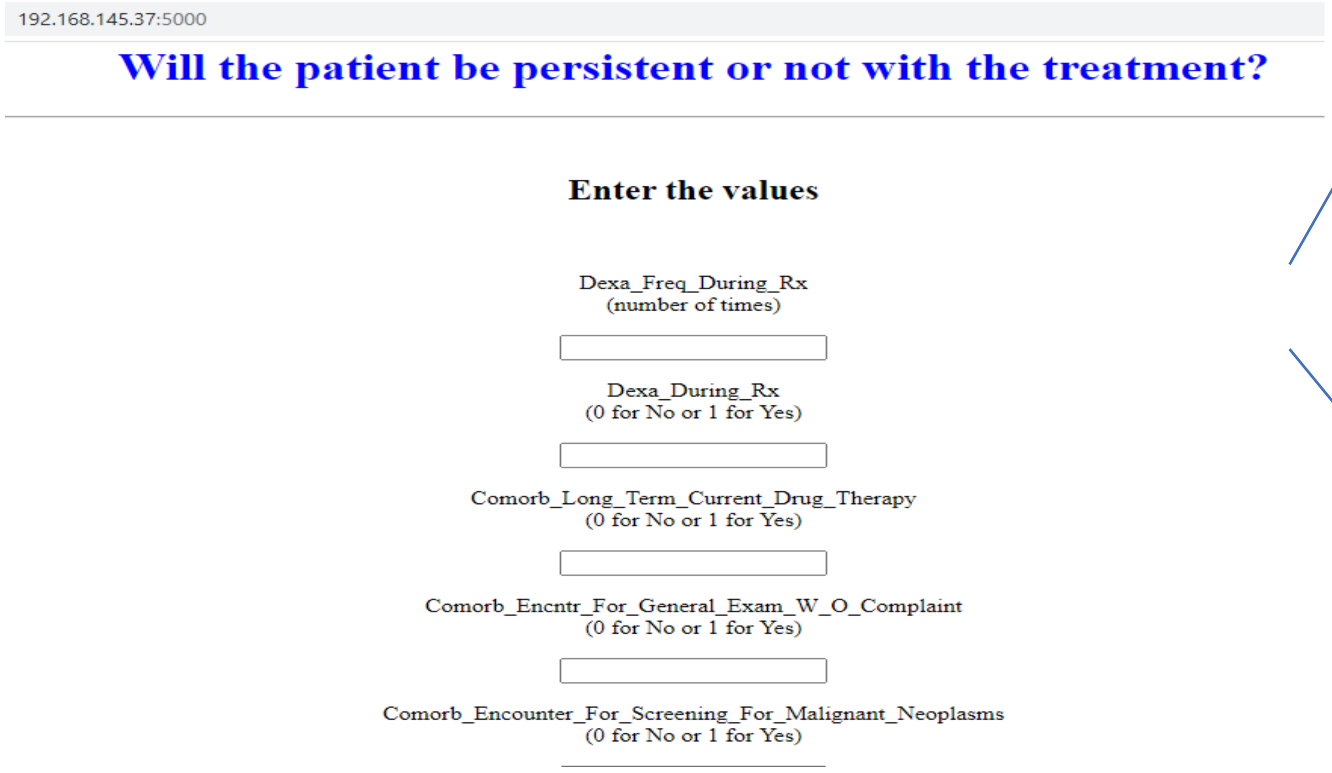


Conclusion

- The “Dexa_freq_during_rx”, is the variable with greater power to predict persistent and non-persistent results. This variable if followed in importance by “Dexa_During_RX” and “Comorb_Long_Term_Current_Drug_Therapy”.
- After those the following chart summarize the importance of other variables to predict the target variable.
- The best model to be used to make predictions is the SVM model with 83.5 % of accuracy.

App Concept

We may check a version of the App with the 5 most important predictors at an 78.27% of accuracy in the prediction.



192.168.145.37:5000

Will the patient be persistent or not with the treatment?

Enter the values

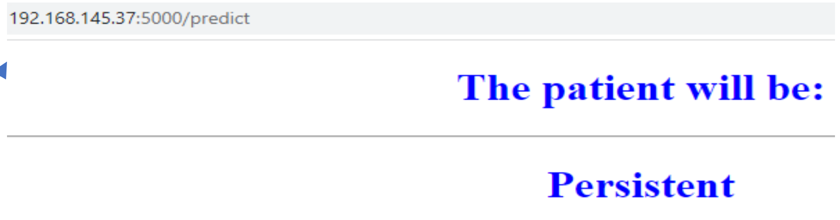
Dexa_Freq_During_Rx
(number of times)

Dexa_During_Rx
(0 for No or 1 for Yes)

Comorb_Long_Term_Current_Drug_Therapy
(0 for No or 1 for Yes)

Comorb_Encntr_For_General_Exam_W_O_Complaint
(0 for No or 1 for Yes)

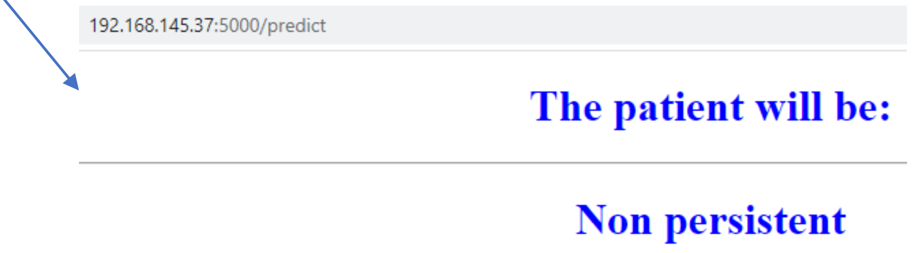
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
(0 for No or 1 for Yes)



192.168.145.37:5000/predict

The patient will be:

Persistent



192.168.145.37:5000/predict

The patient will be:

Non persistent

Data Glacier

Your Deep Learning Partner

Thank You

Github Repo https://github.com/naharift/DataGlacier/tree/main/Week13