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

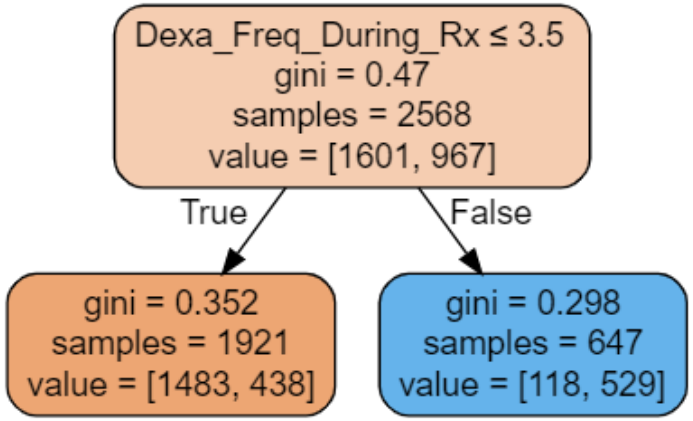
Your Deep Learning Partner

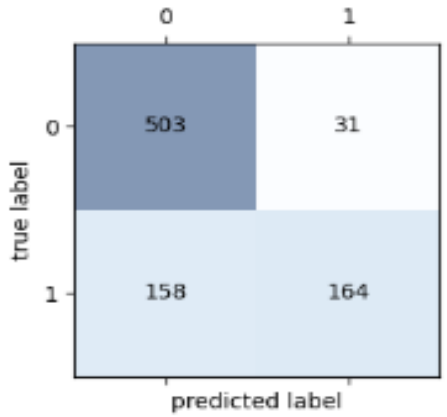
Business Understanding

- 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.

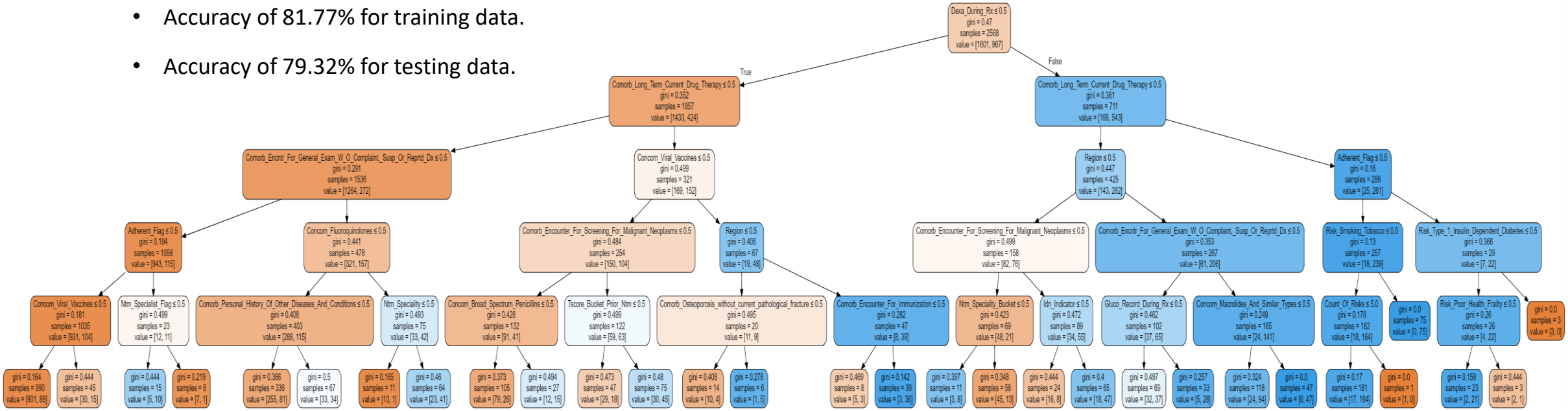




  - "Dexa_Freq_During_Rx" is the variable that has the most value in terms of predictive power. It could be interesting to see if we can get similar results without it. Considering "Dexa_Freq_During_Rx" is kind of another type of treatment. It would be interesting to be able to predict persistence without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

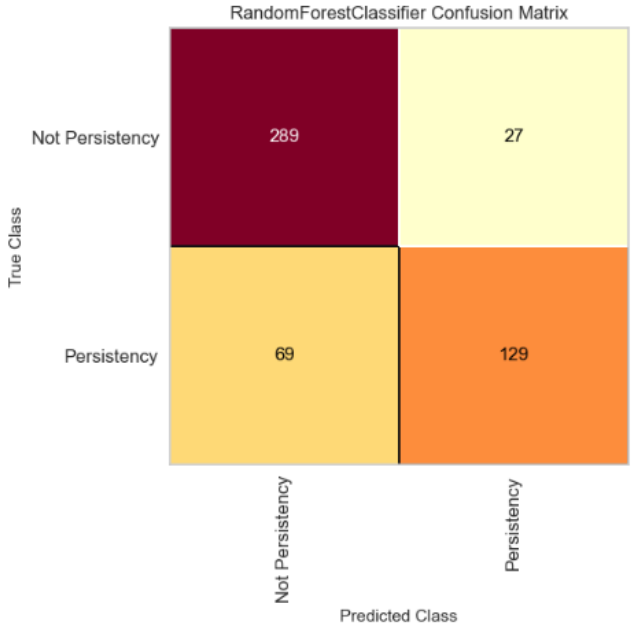
Besides “Dexa_Freq_During_Rx” and “Dexa_During_RX”, in order to predictive power of persistence in taking the drug is Comorbidity factors, Region and viral vaccines.

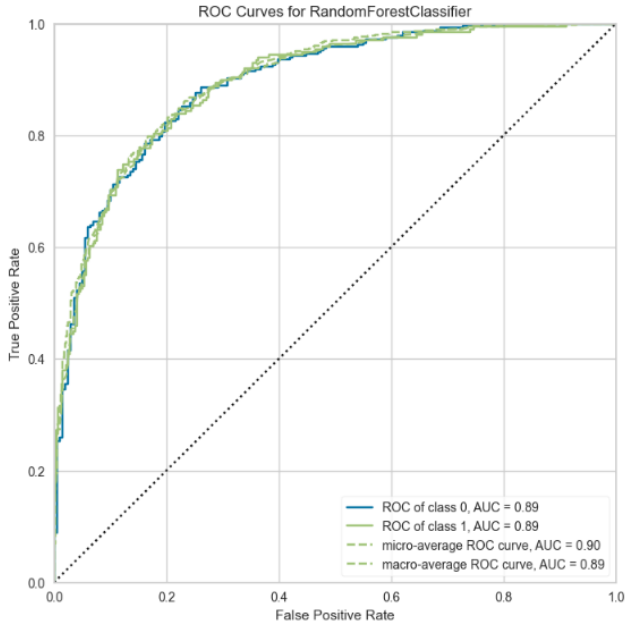
	0	1
Dexa During Rx	Dexa During Rx	0.564448
Comorb Long Term Current Drug Therapy	Comorb Long Term Current Drug Therapy	0.129976
Comorb Encntr For General Exam W O Complaint, Susp Or Reprtd Dx	Comorb Encntr For General Exam W O Complaint, ...	0.072297
Region	Region	0.039838
Concom Viral Vaccines	Concom 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.

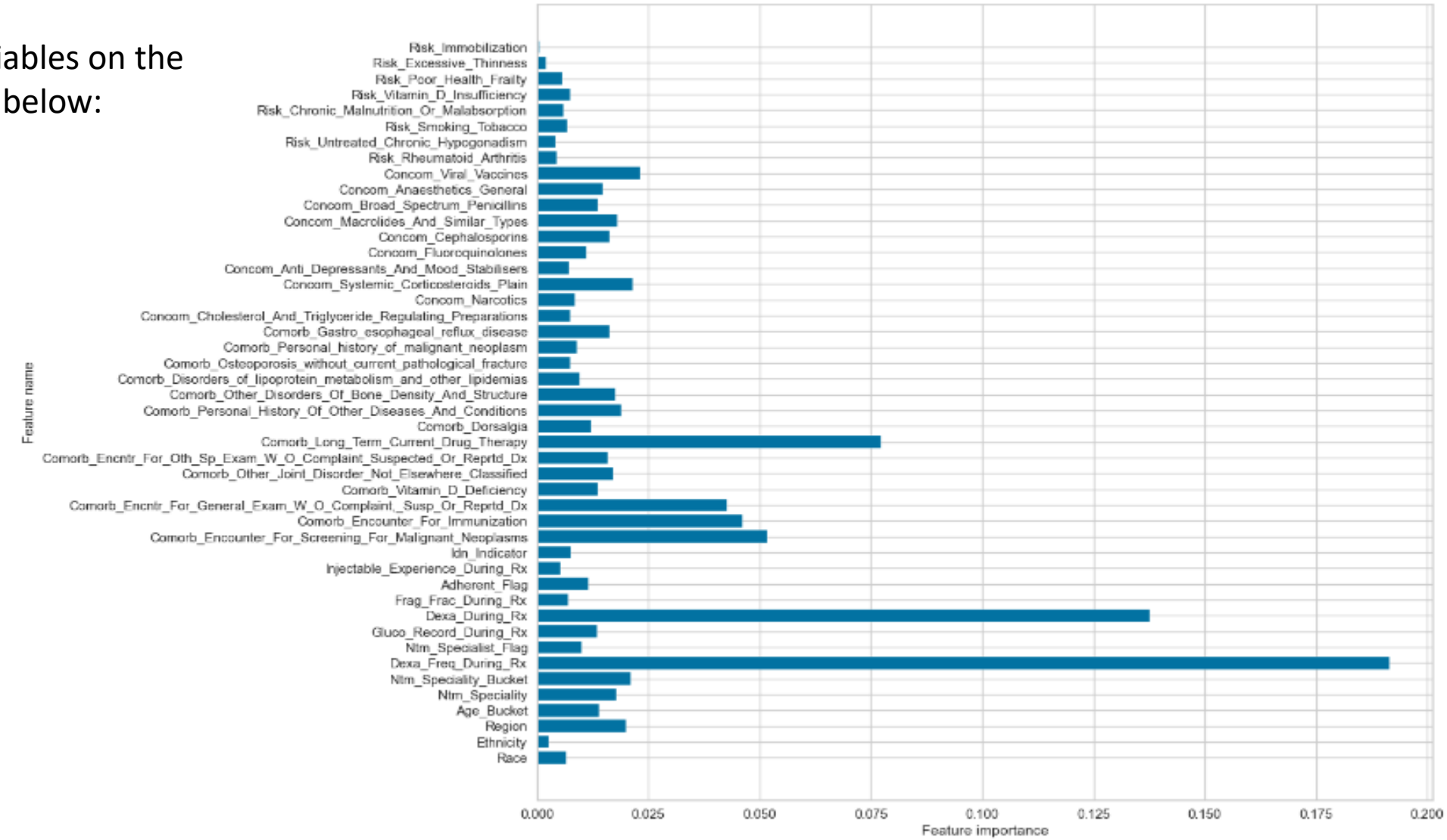
- Accuracy of 88.52% for training data.
- Accuracy of 80.74% for testing data.





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.

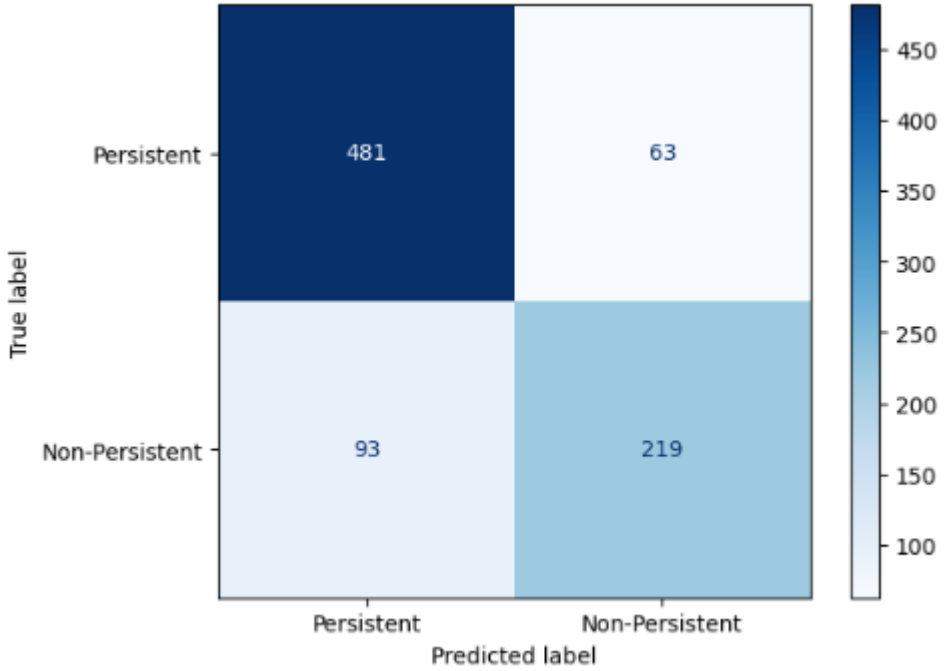
The bottom screenshot displays the different combinations of set hyperparameters and the resulting accuracy when using GridSearchCV. Next slide the best parameters and score will be shown.

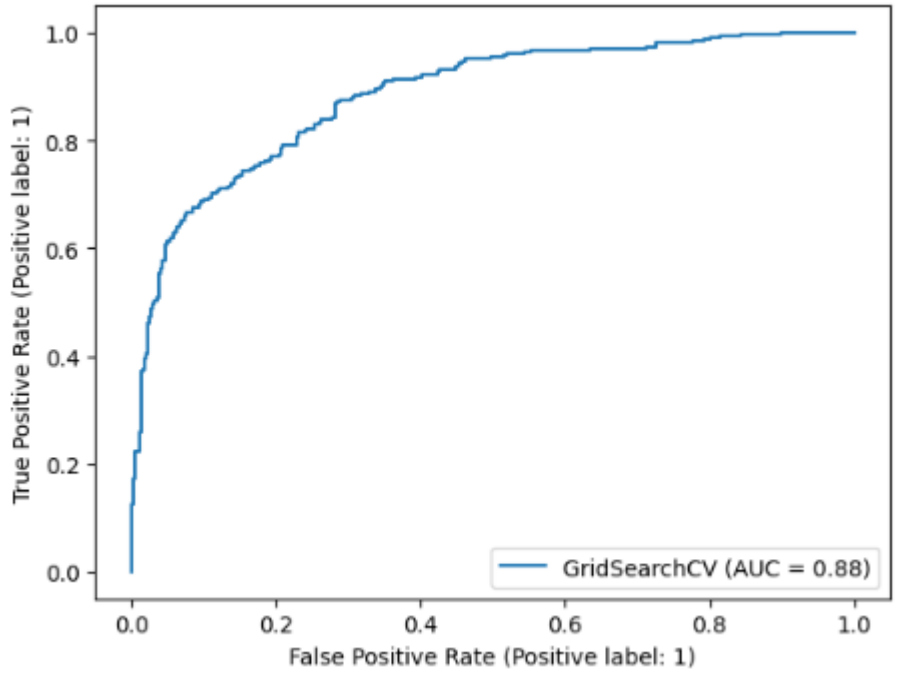
	C	penalty	solver	Accuracy
0	0.001	l1	newton-cg	NaN
1	0.001	l1	lbfgs	NaN
2	0.001	l1	liblinear	0.619549
3	0.001	l1	sag	NaN
4	0.001	l1	saga	0.619549
...	...	...	...	...
65	1000.000	l2	newton-cg	0.807243
66	1000.000	l2	lbfgs	0.809581
67	1000.000	l2	liblinear	0.807632
68	1000.000	l2	sag	0.808800
69	1000.000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

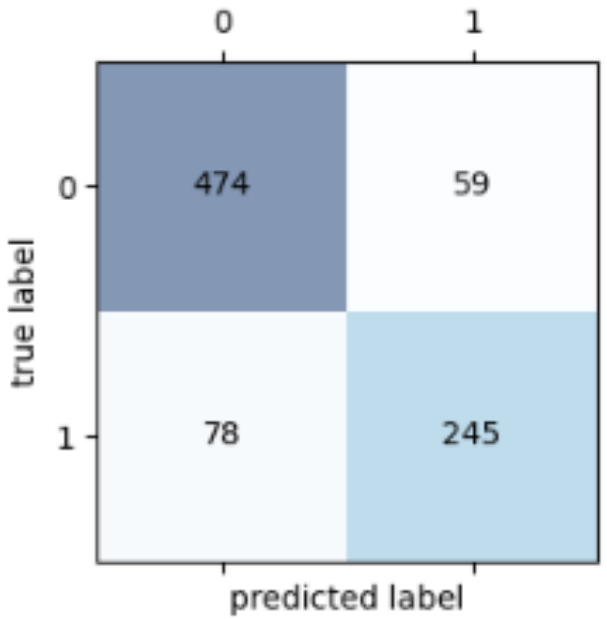




Support Vector Machine

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

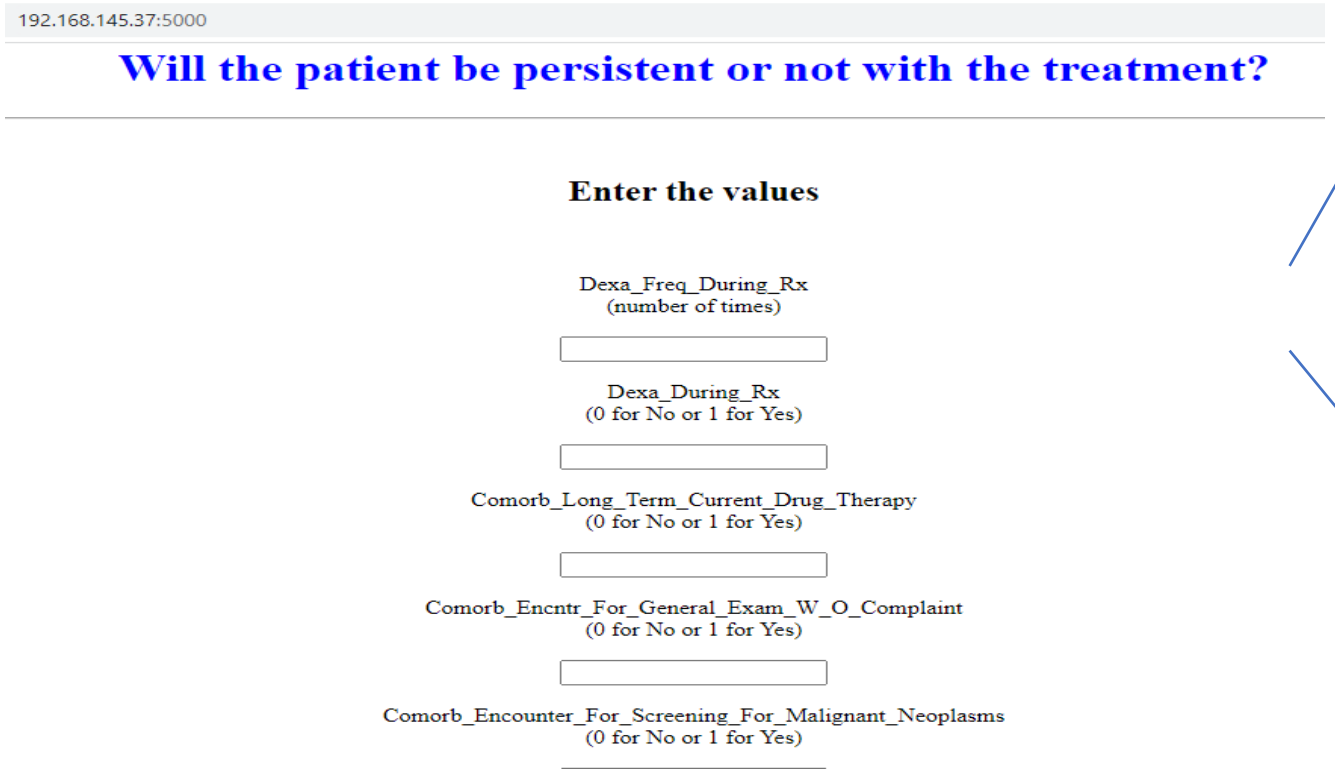


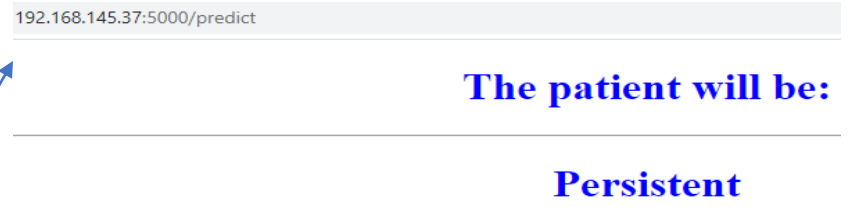
Conclusion

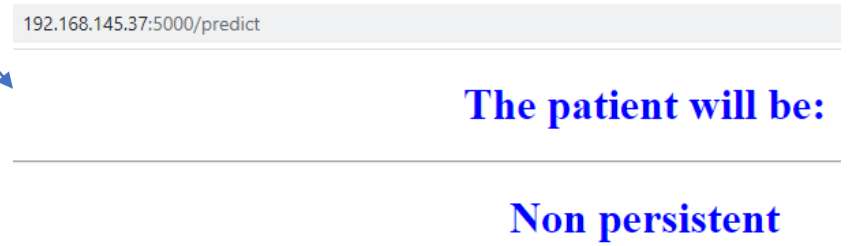
- The “Dexa_freq_during_rx”, is the variable with greater power to predict persistent and non-persistent results. This variable if followed in impontance by “Dexa_During_RX” and “Comorb_Long_Term_Current_Drug_Therapy”.
- After those the following chart summarize the importance of other variables to predict 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.







Thank You

Data Glacier

Your Deep Learning Partner

Github Repo https://github.com/naharift/DataGlacier/tree/main/Week11