# Data Glacier Internship – Week 13

## Group Name: Persistency of my own

| Name | E-mail | Country | College | Specialization |
|---|---|---|---|---|
| Nahari Terena | naharifterena@gmail.com | Brazil | La Sapienza University of Rome | Data Science |

# Healthcare - Persistency of a drug

## Problem Description

One of the challenges for all pharmaceutical companies is to understand the persistence of drugs as per the physician's prescription. To solve this problem, ABC pharma company approached an analytics company to automate this process of identification.

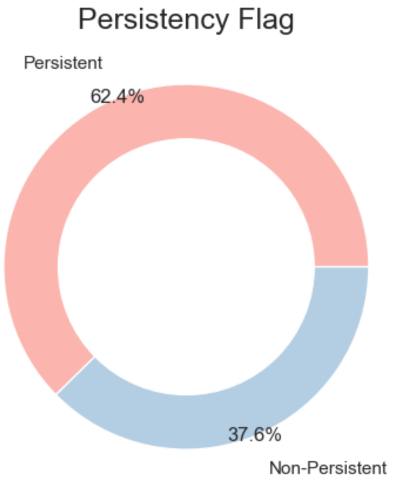
## Business Understanding

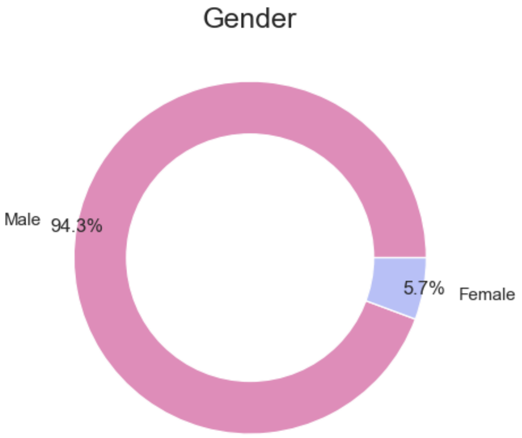
We aim to develop a web app for a pharmaceutical company to predict if a patient will get or not a drug schedule. Since we want to split the patients into categories considering their characteristics, we will develop a classification model.

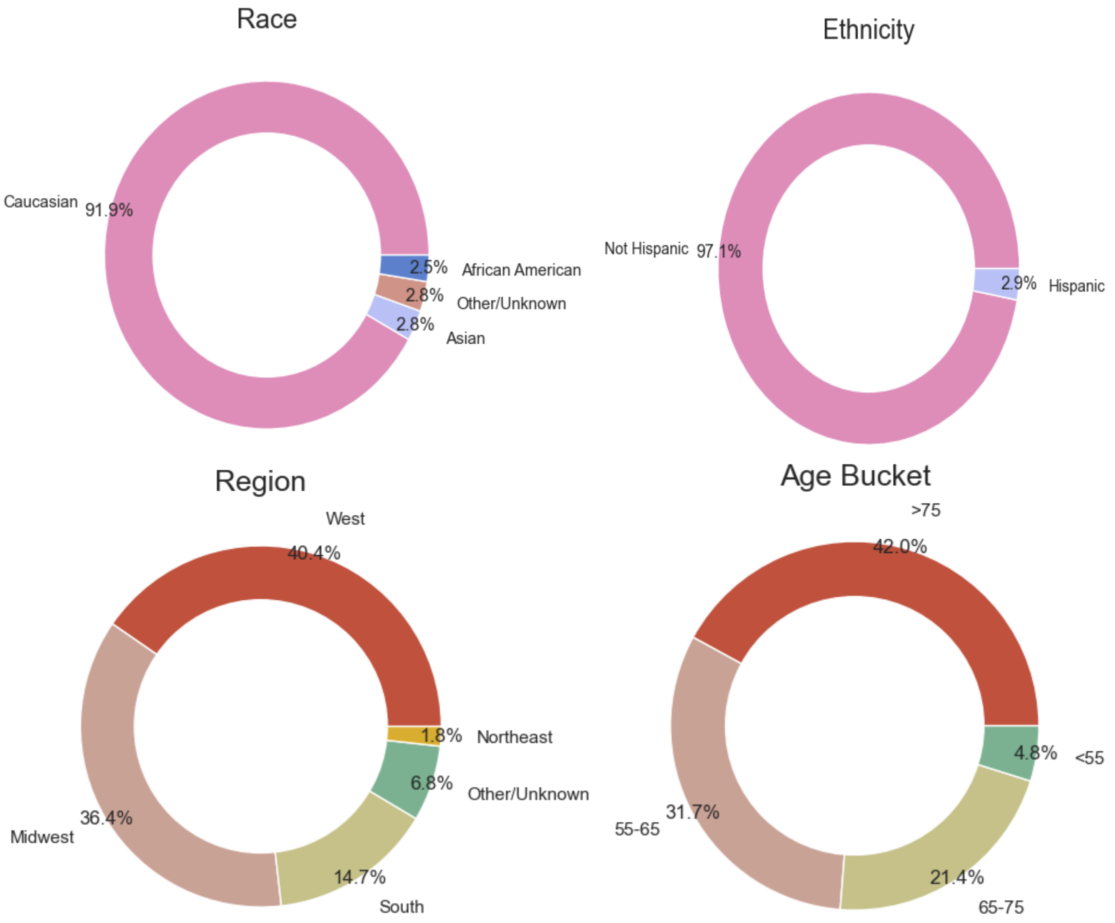
## Exploratory Data Analysis

### Demographics Unidimensional Analysis

Over 60% of patients are persistency to a drug. Of all patients, 94.3% are male. Concerning race, Caucasian is the most common (91%), followed by Asian and Other/Unkwnown. Not Hispanic is 97% of the patients. Around 40% are from the West region. Only 4.8% is less than 55 years old.

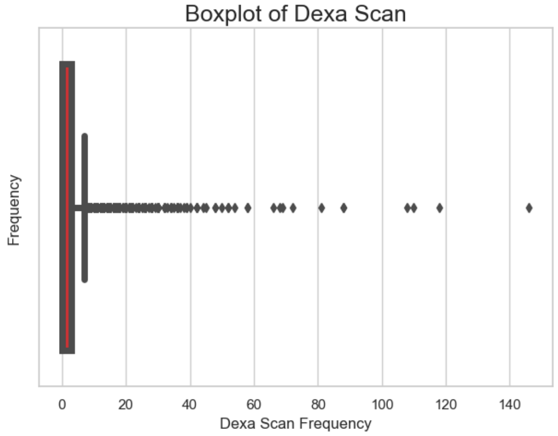


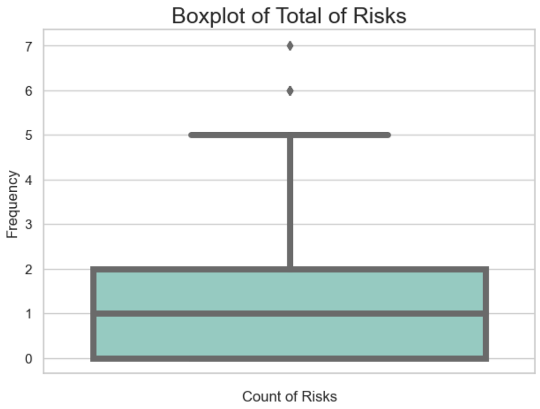




## Outliers Analysis

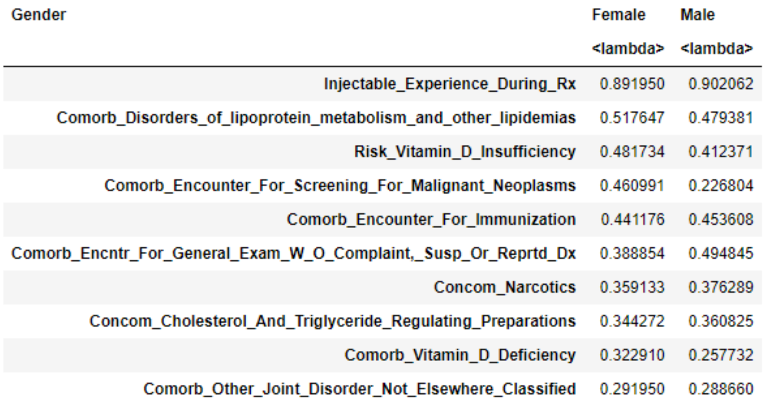
We decided to maintain the outliers so we may probably identify these patients as a group itself.





## Main Questions

- How do risk factors relate to demographics?

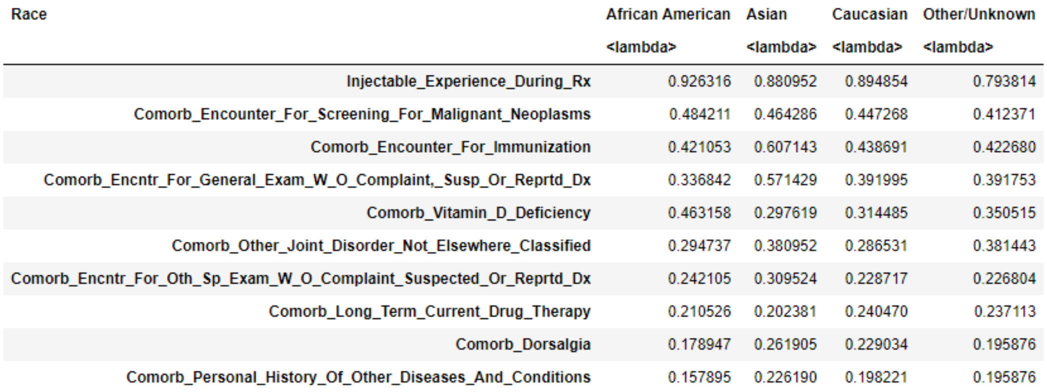


| Gender | Female | Male |
|---|---|---|
| | <lambda> | <lambda> |
| Injectable_Experience_During_Rx | 0.891950 | 0.902062 |
| Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias | 0.517647 | 0.479381 |
| Risk_Vitamin_D_Insufficiency | 0.481734 | 0.412371 |
| Comorb_Encounter_For_Screening_For_Malignant_Neoplasms | 0.460991 | 0.226804 |
| Comorb_Encounter_For_Immunization | 0.441176 | 0.453608 |
| Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx | 0.388854 | 0.494845 |
| Concom_Narcotics | 0.359133 | 0.376289 |
| Concom_Cholesterol_And_Triglyceride_Regulating_Preparations | 0.344272 | 0.360825 |
| Comorb_Vitamin_D_Deficiency | 0.322910 | 0.257732 |
| Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified | 0.291950 | 0.288660 |

There are some significant differences between genders:

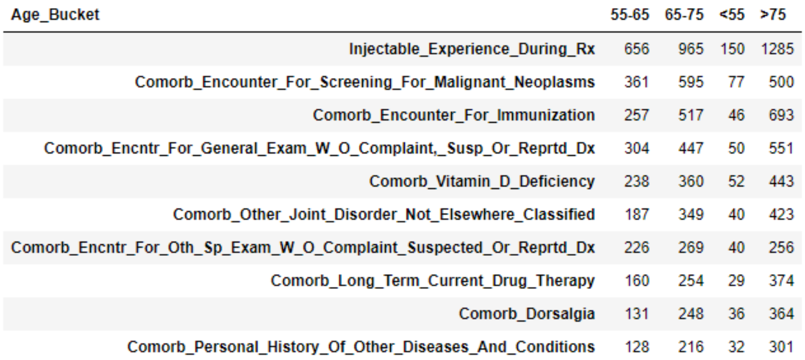
- Women seem to be more affected by vitamin D deficiencies.
- More than twice as many women as men have passed as screening for malignant neoplasms.
- Four times as many men as women suffer from Hypogonadism (untreated).

There are some risks and other factors that seem to be significantly higher in South and West regions. It might be interesting to find out about socioeconomic factors aside.



| Race | African American | Asian | Caucasian | Other/Unknown |
|---|---|---|---|---|
| | <lambda> | <lambda> | <lambda> | <lambda> |
| Injectable_Experience_During_Rx | 0.926316 | 0.880952 | 0.894854 | 0.793814 |
| Comorb_Encounter_For_Screening_For_Malignant_Neoplasms | 0.484211 | 0.464286 | 0.447268 | 0.412371 |
| Comorb_Encounter_For_Immunization | 0.421053 | 0.607143 | 0.438691 | 0.422680 |
| Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx | 0.336842 | 0.571429 | 0.391995 | 0.391753 |
| Comorb_Vitamin_D_Deficiency | 0.463158 | 0.297619 | 0.314485 | 0.350515 |
| Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified | 0.294737 | 0.380952 | 0.286531 | 0.381443 |
| Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx | 0.242105 | 0.309524 | 0.228717 | 0.226804 |
| Comorb_Long_Term_Current_Drug_Therapy | 0.210526 | 0.202381 | 0.240470 | 0.237113 |
| Comorb_Dorsalgia | 0.178947 | 0.261905 | 0.229034 | 0.195876 |
| Comorb_Personal_History_Of_Other_Diseases_And_Conditions | 0.157895 | 0.226190 | 0.198221 | 0.195876 |

There seems to be some remarkable differences between Asian and other races. They are probably due to cultural factors and other behaviors.

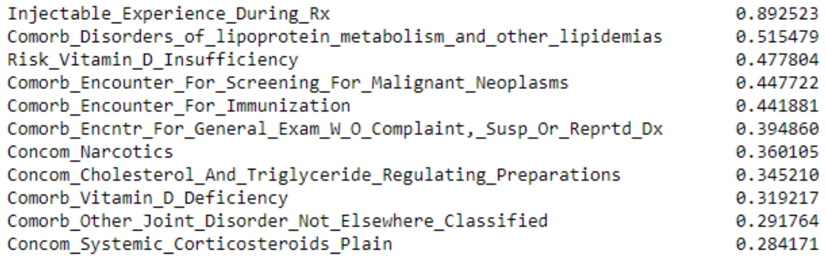


| Age_Bucket | 55-65 | 65-75 | <55 | >75 |
|---|---|---|---|---|
| Injectable_Experience_During_Rx | 656 | 965 | 150 | 1285 |
| Comorb_Encounter_For_Screening_For_Malignant_Neoplasms | 361 | 595 | 77 | 500 |
| Comorb_Encounter_For_Immunization | 257 | 517 | 46 | 693 |
| Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx | 304 | 447 | 50 | 551 |
| Comorb_Vitamin_D_Deficiency | 238 | 360 | 52 | 443 |
| Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified | 187 | 349 | 40 | 423 |
| Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx | 226 | 269 | 40 | 256 |
| Comorb_Long_Term_Current_Drug_Therapy | 160 | 254 | 29 | 374 |
| Comorb_Dorsalgia | 131 | 248 | 36 | 364 |
| Comorb_Personal_History_Of_Other_Diseases_And_Conditions | 128 | 216 | 32 | 301 |

Patients older than 65 years old are affect by the risks in higher proportion.

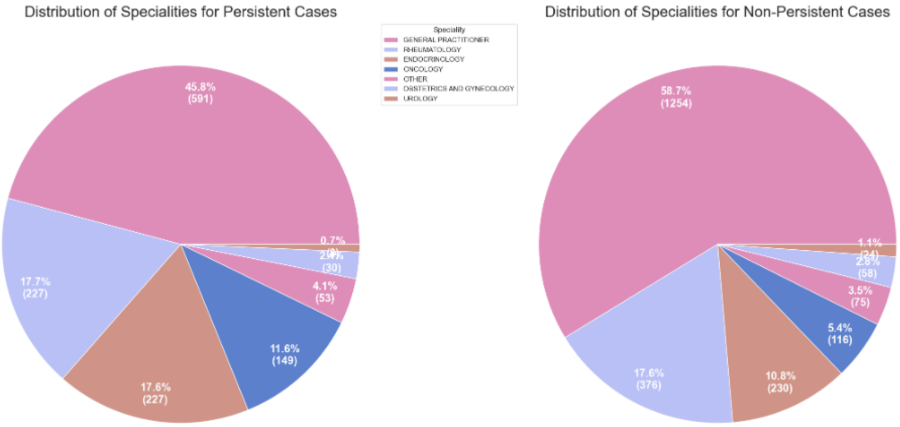
- What is the percentage of patients holding at least one factor?
  About 99% holds at least one risk, comorbid or concomitant factor.

- What are the most common risk factors?
  It’s easy to see that most of the patients already hold comorbidity factors, while holding risk factors is less common.
  The main comorbidity factor is related to lipoproteins and metabolism (cholesterol). On the other hand, the main risk factor is deficiency in vitamin D. More than one third has been found to have taken narcotics.



```
Injectable_Experience_During_Rx                                       0.892523
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias       0.515479
Risk_Vitamin_D_Insufficiency                                          0.477804
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms                0.447722
Comorb_Encounter_For_Immunization                                     0.441881
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx       0.394860
Concom_Narcotics                                                      0.360105
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations           0.345210
Comorb_Vitamin_D_Deficiency                                           0.319217
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified                  0.291764
Concom_Systemic_Corticosteroids_Plain                                 0.284171
```

- Does “Ntm_Speciality” variables seems to have useful information for the classification task?
  The distributions of frequency for the target variable by speciality are pretty similar. Thus, we may rule out the possibility that one of the factors is who prescribed the drug in the first place.



## Final Recommendation

Demographics variables seems to be relevant because it brings the context in which the patient is inserted. As well as comorbid and risks factors.

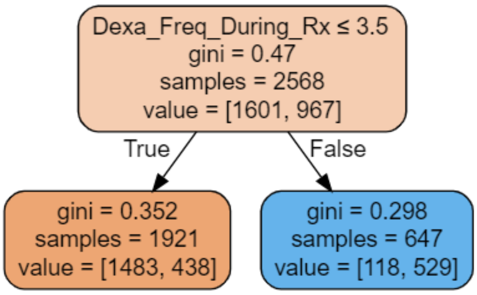
Regarding the prediction model, we may consider interpretable model such as decision tree to support the predictions of more complex models.

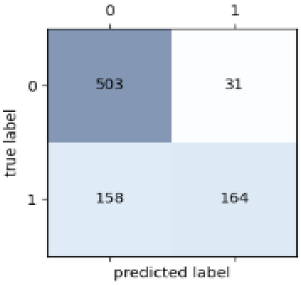
# Modeling

1. **Decision Tree**

In the first scenario, the best parameter was a tree of depth 1. Therefore, the results were:

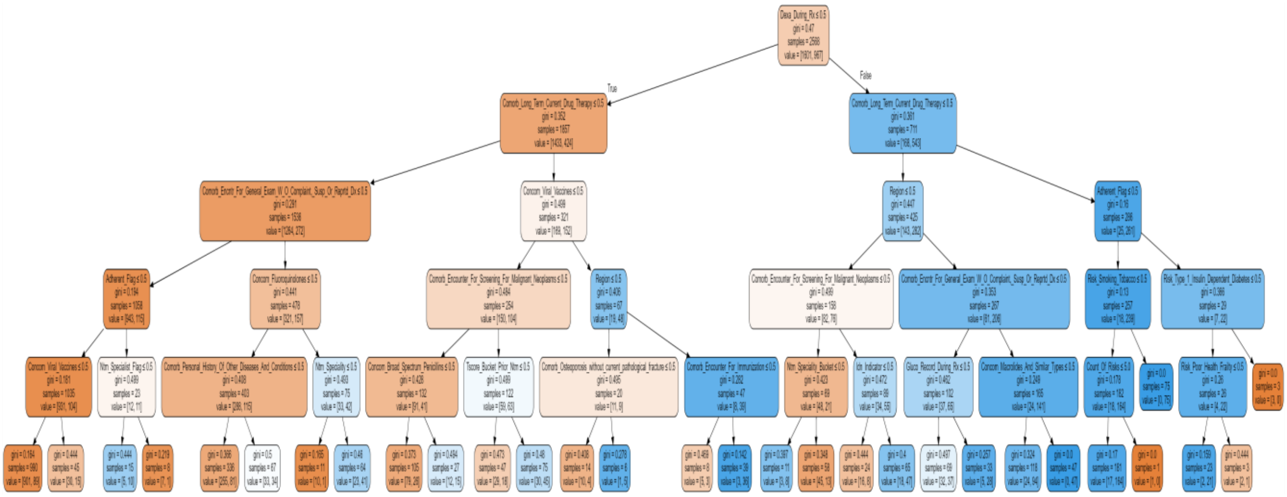
- Accuracy of 78.35% for training data.
- Accuracy of 77.92% for testing data.
- “Dexa_Freq_During_Rx” is the variable that has the most value in terms of predictive power. It could be interesting to see if we can get similar results without it. Considering “Dexa_Freq_During_Rx” is kind of another type of treatment. It would be interesting to be able to predict persistence without it.





In the second scenario, the best parameter was a tree of depth of 5. Therefore, the results were:

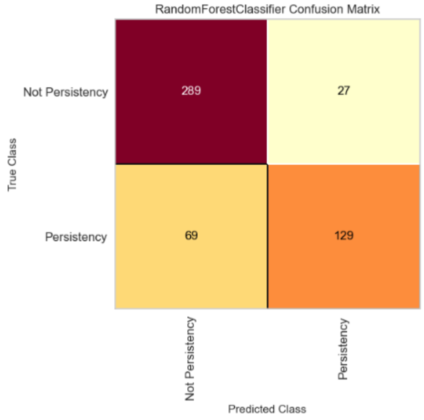
- Accuracy of 81.77% for training data.
- Accuracy of 79.32% for testing data.
- “Dexa_Freq_During_Rx” is the variable that has the most value in terms of predictive power. It could be interesting to see if we can get similar results without it. Considering “Dexa_Freq_During_Rx” is kind of another type of treatment. It would be interesting to be able to predict persistence without it.

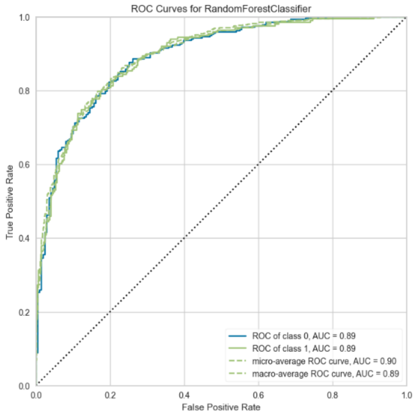

Dexa_During_Rx ≤ 0.5 gini = 0.47 samples = 2568 value = [1601, 967]
True
False
Comorb_Long_Term_Current_Drug_Therapy ≤ 0.5 gini = 0.352 samples = 1857 value = [1433, 424]
Comorb_Long_Term_Current_Drug_Therapy ≤ 0.5 gini = 0.361 samples = 711 value = [168, 543]
Comorb_Encntr_For_General_Exam_W_O_Complaint_Susp_Or_Reptd_Dx ≤ 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_Reptd_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 = 0.184 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 = 0.48 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]

## 2. 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.

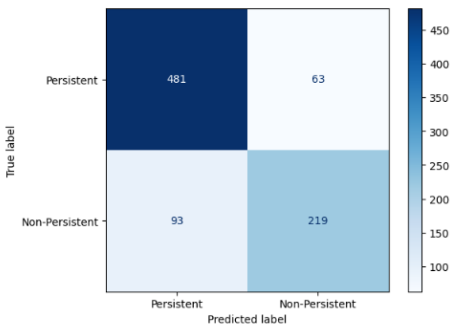


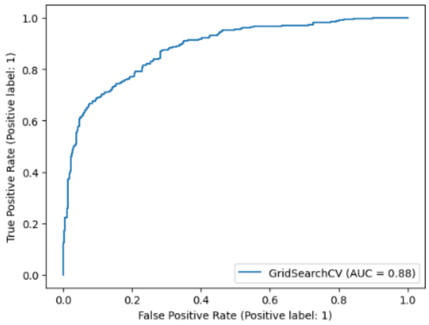


## 3. Logistic Regression

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

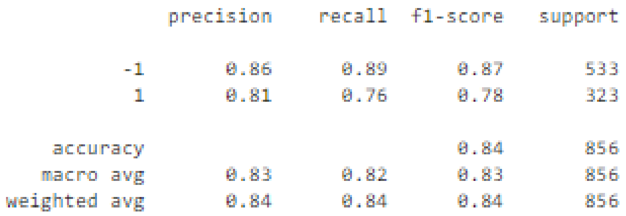
F1-score: 0.8156





## 4. Support Vector Machine

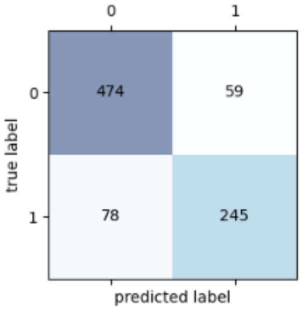
Support Vector Machines algorithm to classify the persistence of patients (1 for positives and -1 for negatives). A linear kernel has been used, obtaining an accuracy of 84.0% over testing data (25 % out of the whole dataset).



```
              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
```



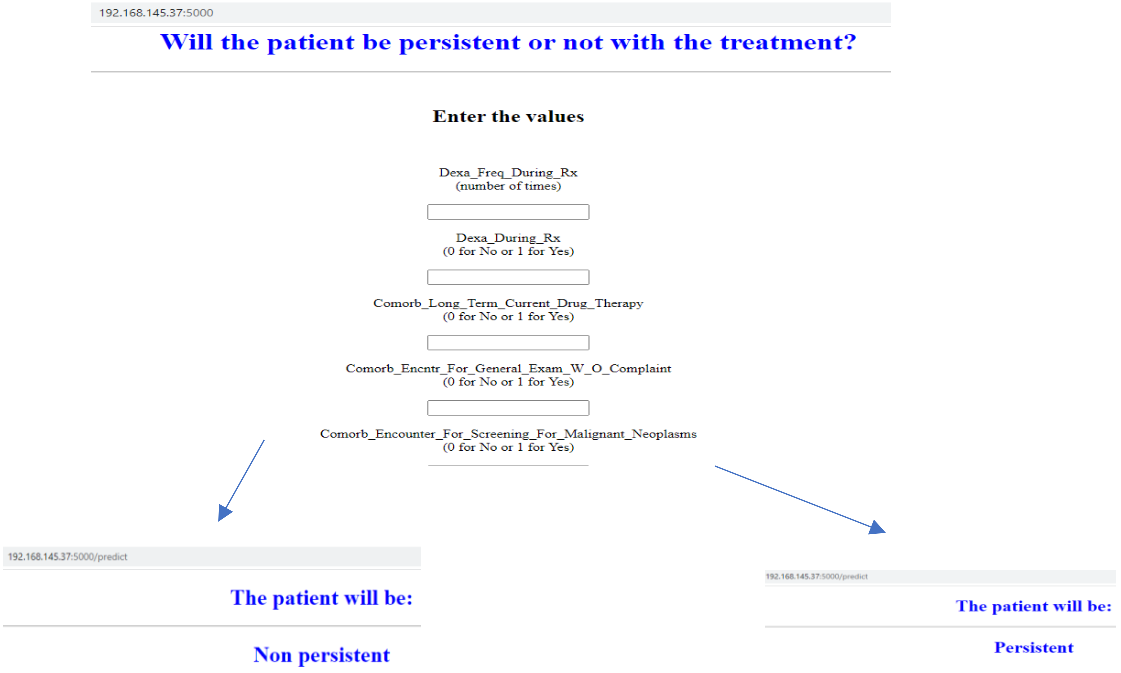
The “Dexa_freq_during_rx”, is the variable with greater power to predict persistent and non-persistent results. This variable if followed in importance by “Dexa_During_RX” and “Comorb_Long_Term_Current_Drug_Therapy”.

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 Deployment

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:**

**Non persistent**

192.168.145.37:5000/predict

**The patient will be:**

**Persistent**

## Github Repo [https://github.com/naharift/DataGlacier/tree/main/Week13](https://github.com/naharift/DataGlacier/tree/main/Week13)