Healthcare Persistency of a Drug Classification

Name: Igor Azevedo de Queiroz
E-mail: igor_queiroz17@yahoo.com.br
Country: Ireland
Specialization: Data Science

Indices

1. Context
2. Challenge
3. Solution Development
4. Conclusion & Demonstration
5. Next Steps

1. Context

- One of the challenges for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription.

- To solve this problem a pharma company, located in the USA, wishes to automate this process of identification.

2. Challenge

Problem.

- Gather insights on the factors that are impacting the persistency of a drug during the treatment.

Solution

- Create a dashboard with several hypotheses and insights to help the company CEO with future decisions.
- Use Machine Learning to classify futures patients, informing if they will use the drugs during the entire treatment or if they won't.
- Visualization if they will use the drugs during the role treatment or won't can be done in a webapp.

3. Solution Development

Solution adopted to generate business insights and create a ML classification model to solve the proposed problem. This solution includes:

- Data Description;
  - ➢ Data Dimensions
  - ➢ Descriptive Statistics
- Mind Map Hypothesis;
- Hypothesis Creation;
- Exploratory Data Analysis (EDA);
  - ➢ Target Variable analysis
  - ➢ Numerical Variable analysis
  - ➢ Hypothesis Validation
  - ➢ Summary of Hypothesis
  - ➢ Multivariate Analysis – Numerical Variables
- Data Preparation;
- Machine Learning Models Performance;
  - ➢ Extra Trees Classifier
  - ➢ LGBM Classifier
  - ➢ Stochastic Gradient Descent
  - ➢ Random Forest
  - ➢ Support Vector Machine.
- Business Performance;
  - ➢ Accuracy
  - ➢ Confusion Matrix & Classification Report
- Persistency of a Drug Classification ML WebApp.

3.1. Data Description

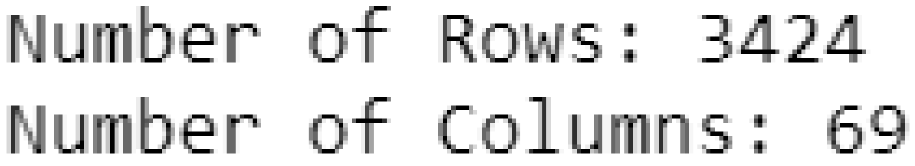
Here it is checked:

- Amount of data;

- General summary of the data;

- Descriptive statistics: Used to have a notion of the magnitude of the data; notion of minimum and maximum limits of the variables; what is the range; what is the median and others.

3.1.1. Data Dimensions



```
Number of Rows: 3424
Number of Columns: 69
```

We are going to work with 3424 rows, informing the quantity of patients that are persistent of drug and patients that are not persistent of a drug.

The number of features is 69.

This dataset, in terms of number of values, is not big, with a large number of features, which means that will not be so easy to train the classifier. Also we have to exclude some features, because this large number is prejudice for the training of the ML classifier.

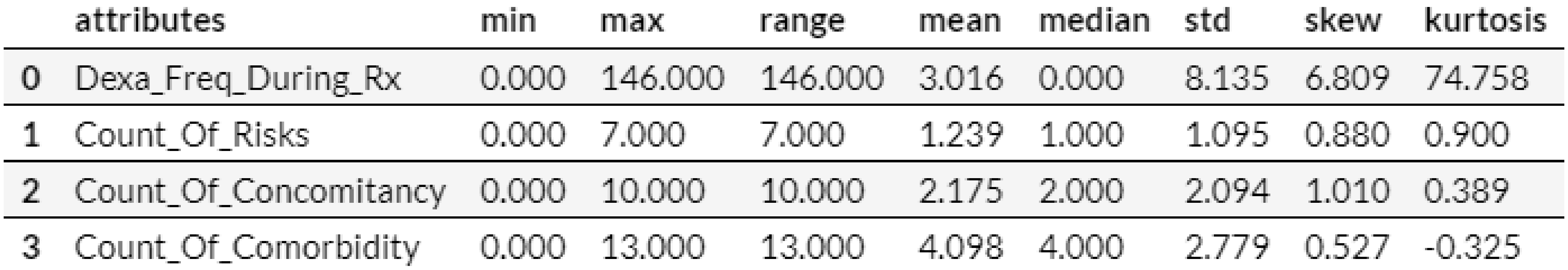
3.1.2. Descriptive Statistical

Used to gain business knowledge and to be able to detect errors. There are two major metrics within descriptive statistics:

- Central Tendency Metrics: summarize data into a single number. Ex: mean and median.

- Dispersion metrics: These tell you whether the data is too concentrated near the mean or too dispersed from it. Ex: Variance, standard deviation, minimum and maximum, range, Skew and Kurtosis.

- For example, the attribute Dexa_Freq_During_RX have a very high range and kurtosis, which means that the feature has a large number of values concentrated at a same value, for example, there are several values equal 0. Also, its range is very different of the others attributes, which means that it can cause a problem at the model training, and has to be fixed.



	attributes	min	max	range	mean	median	std	skew	kurtosis
0	Dexa_Freq_During_Rx	0.000	146.000	146.000	3.016	0.000	8.135	6.809	74.758
1	Count_Of_Risks	0.000	7.000	7.000	1.239	1.000	1.095	0.880	0.900
2	Count_Of_Concomitancy	0.000	10.000	10.000	2.175	2.000	2.094	1.010	0.389
3	Count_Of_Comorbidity	0.000	13.000	13.000	4.098	4.000	2.779	0.527	-0.325

3.2. Mind Map Hypothesis

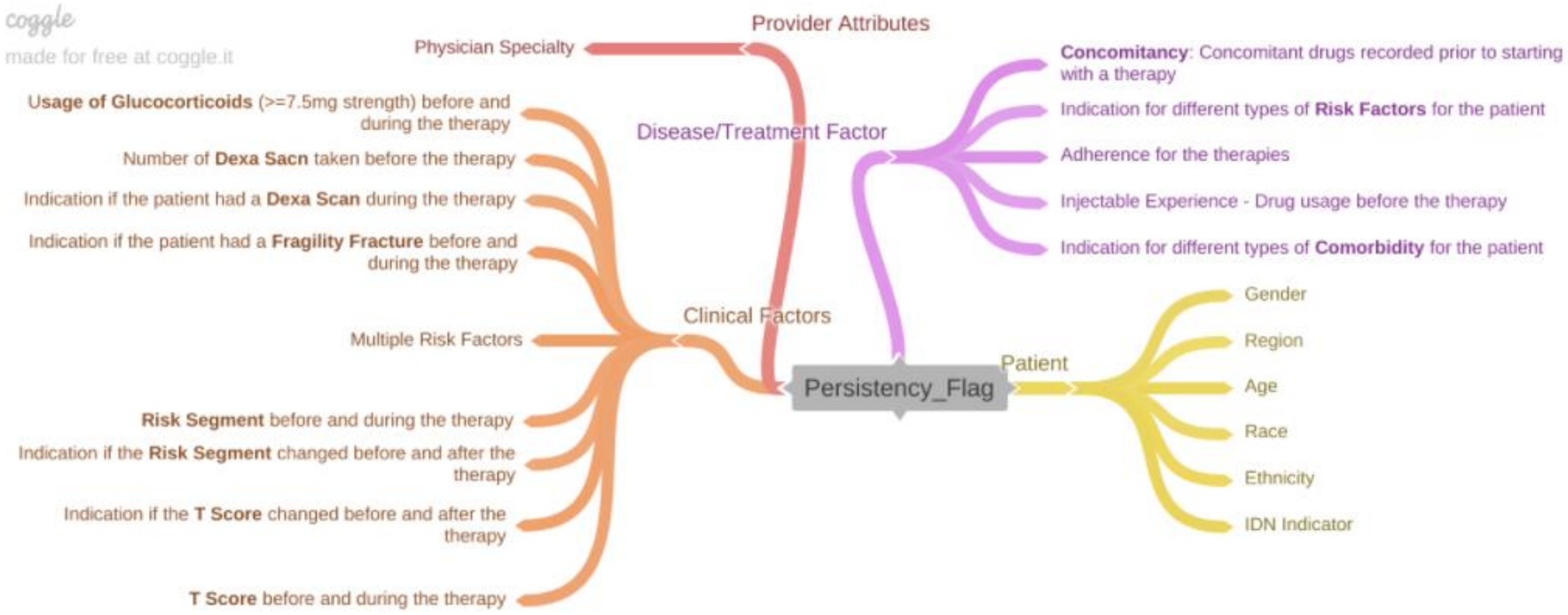
- This mind map is a guide for exploratory data analysis.

- It describes in the centre, the phenomenon that should be modelled, in this case the Persistency Flag attribute.

- The first level are all the entities that impact this phenomenon, making the persistence to be yes or no.

- In the last level there are the attributes that describe the entities by themselves.

- This allows us to create hypotheses about them and guide us during the exploratory data analysis.



coggle

made for free at coggle.it

Persistency_Flag

Provider Attributes

Physician Specialty

Disease/Treatment Factor

Concomitancy: Concomitant drugs recorded prior to starting with a therapy

Indication for different types of **Risk Factors** for the patient

Adherence for the therapies

Injectable Experience - Drug usage before the therapy

Indication for different types of **Comorbidity** for the patient

Patient

Gender

Region

Age

Race

Ethnicity

IDN Indicator

Clinical Factors

Usage of Glucocorticoids (>=7.5mg strength) before and during the therapy

Number of **Dexa Sacn** taken before the therapy

Indication if the patient had a **Dexa Scan** during the therapy

Indication if the patient had a **Fragility Fracture** before and during the therapy

Multiple Risk Factors

Risk Segment before and during the therapy

Indication if the **Risk Segment** changed before and after the therapy

Indication if the **T Score** changed before and after the therapy

T Score before and during the therapy

3.3. EDA Hypothesis

3.3.1. Patient Hypothesis

1. Female patients are more persistent of a drug than male.

2. Patients from Northeast are more persistent of a drug than patients from South.

3. Patients over 65 years of age are more persistent of a drug than patients 65 years of age or younger.

4. Caucasian patients, not Hispanic are more persistent of a drug than patients with different race and ethnicity.

5. Patients mapped to IDN are more persistent of a drug than patient not mapped.

3.3.2. Provider Attributes Hypothesis

6. Patients that received the drug prescription from General Practitioner Specialty are less persistent of a drug than patients that received the drug prescription from others Specialty.

3.3.3. Clinical Factors Hypothesis

7. Patients that used Glucocorticoid and had a Fragility Fracture, before and during the therapy, are more persistent of a drug than patients that not used Glucocorticoid neither had a Fragility Fracture, in any situation.

8. Patients that had Dexa Scan more than 0 times are more persistent of a drug than patients that had Dexa Scan less than 0 times.

9. Patients that presents VLR_LR as Risk Segment during the therapy and presents Unknown value in Risk Segment after the therapy are more persistent of a drug than patients that presents other types of Risk Segment and Change in Risk Segment.

10. Patients that presents T Socre of >-2.5 before the therapy and presents No Change T Score status after the therapy are less persistent of a drug than patients that presents other types of T Score and T Score status.

3.3.4. Disease/Treatment Factor Hypothesis

11. Patients that presents the amount of risk factor higher than 1 are more persistent of a drug than patients that presents the amount of risk factor lower than 1.

12. Patients adherent for therapies and that used drugs before the therapy are more persistent of a drug than patients in different satiations related to both attributes.

13. Patients that used more than 2 concomitancy drugs before the therapy are less persistent of a drug than patients that used more than 2 concomitancy drugs before the therapy.

14. Patients that presents more than 3 different types of comorbidity are more persistent of a drug than patients that presents less than 3 different types of comorbidity.

3.4. Exploratory Data Analysis (EDA)

EDA is a step in a data science process where we dive into the data and start to understand some behaviours that visually or through data manipulation, become clearer.

With EDA it is possible to gain business experience, to understand how the business works, and to also understand the behaviour that the business has through the data.

Here also, we validate or refute the hypotheses.

This information provided by the hypotheses can be very useful and assists in the decision making process.

Each validated or discarded hypothesis is called insight.

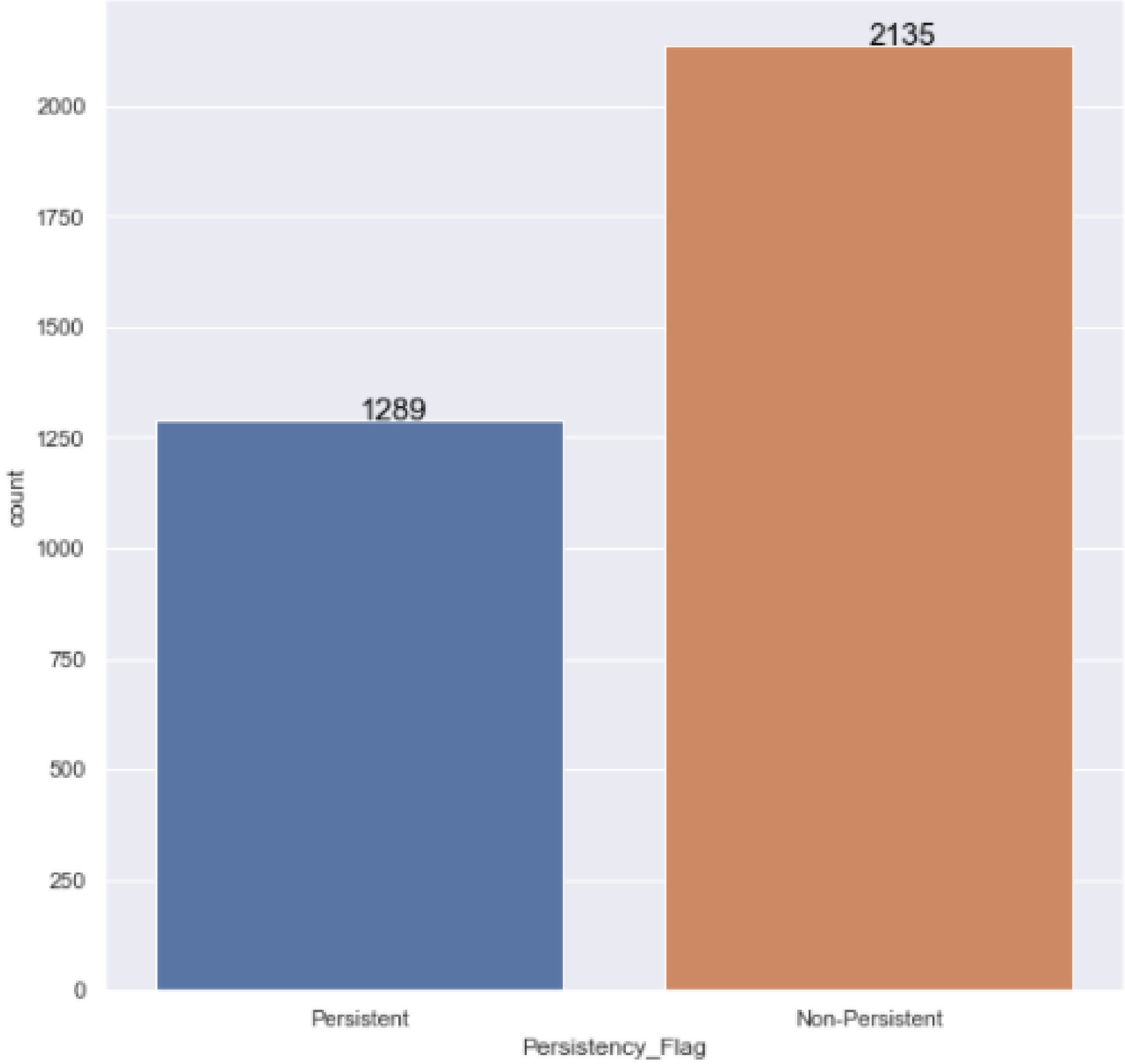
The creation of insights is generated through surprise and through the counterpoint of the CEO's belief. It surprises and/or alters the CEO's belief about the business.

3.4.1. Target Variable

On this bar graph, the blue bar informs the quantity of patients that persist in the use of the drug during the whole treatment.

The orange bar informs the quantity of patients that do not persist in the use of the drug, i.e., give up the use of the drug before finishing the treatment.

As it is possible to analyse, the quantity of patients that do not persist in the use of the drug during the treatment is bigger than the quantity of patients that persist.



3.4.2. Numerical Variables

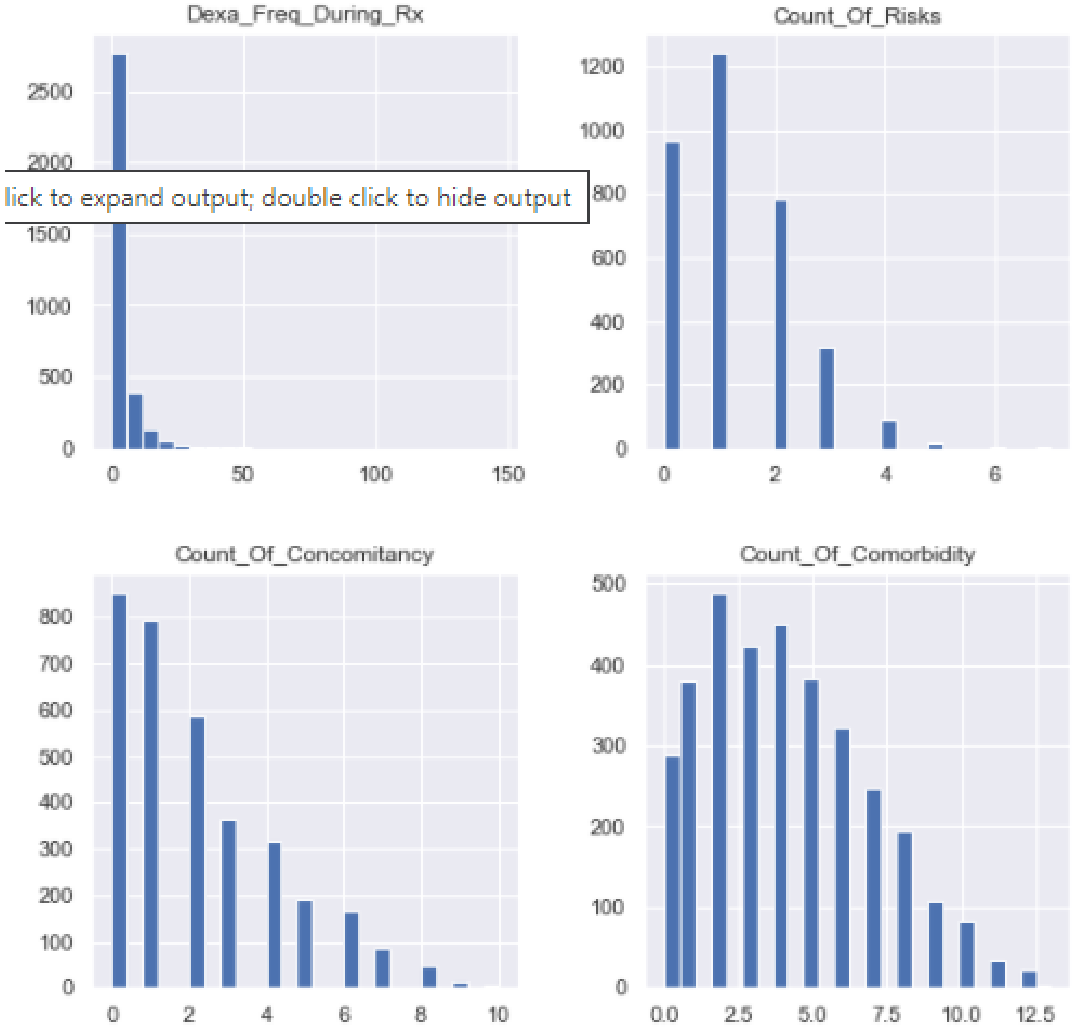
All numerical attributes show a well varied behaviour, not stagnated, presenting relevant information for the learning of the ML model. In other words, this information can help to obtain more accuracy in the evaluation of the ML model.

All variables show a positive skew, tending to the left.

The variable 'Dexa_Freq_During_Rx' presents a very high kurtosis, because its values are extremely concentrated on the value 0.

The other variables report that the majority of patients demonstrate an amount of risk, concomitancy and comorbidity between 0 and 2.5.

And patients who show more risk, more concomitancy and more comorbidity, are in smaller numbers.

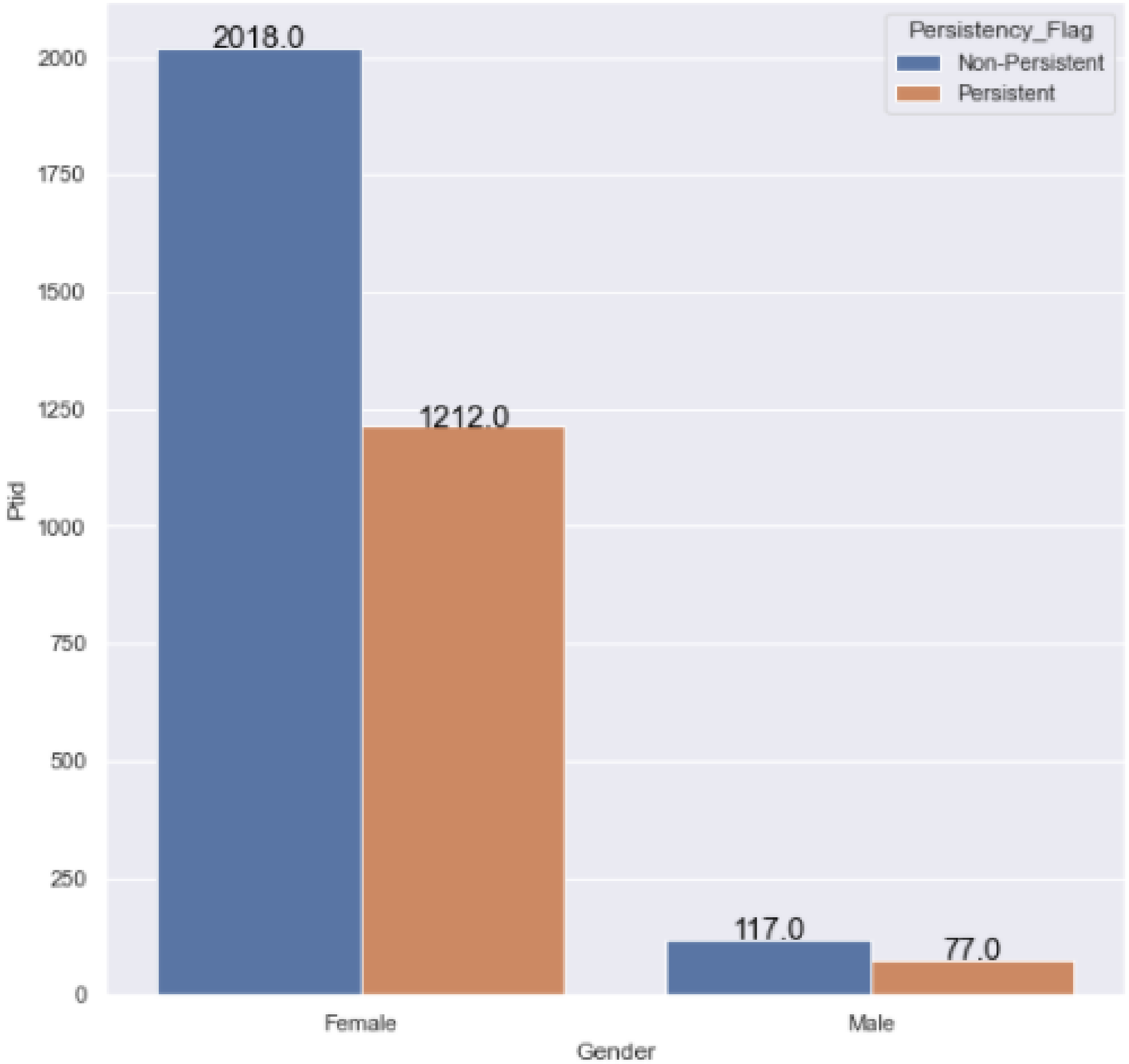


3.4.3. Hypothesis Validation

With the assistance of graphs it will be verified if the hypotheses are true or not.

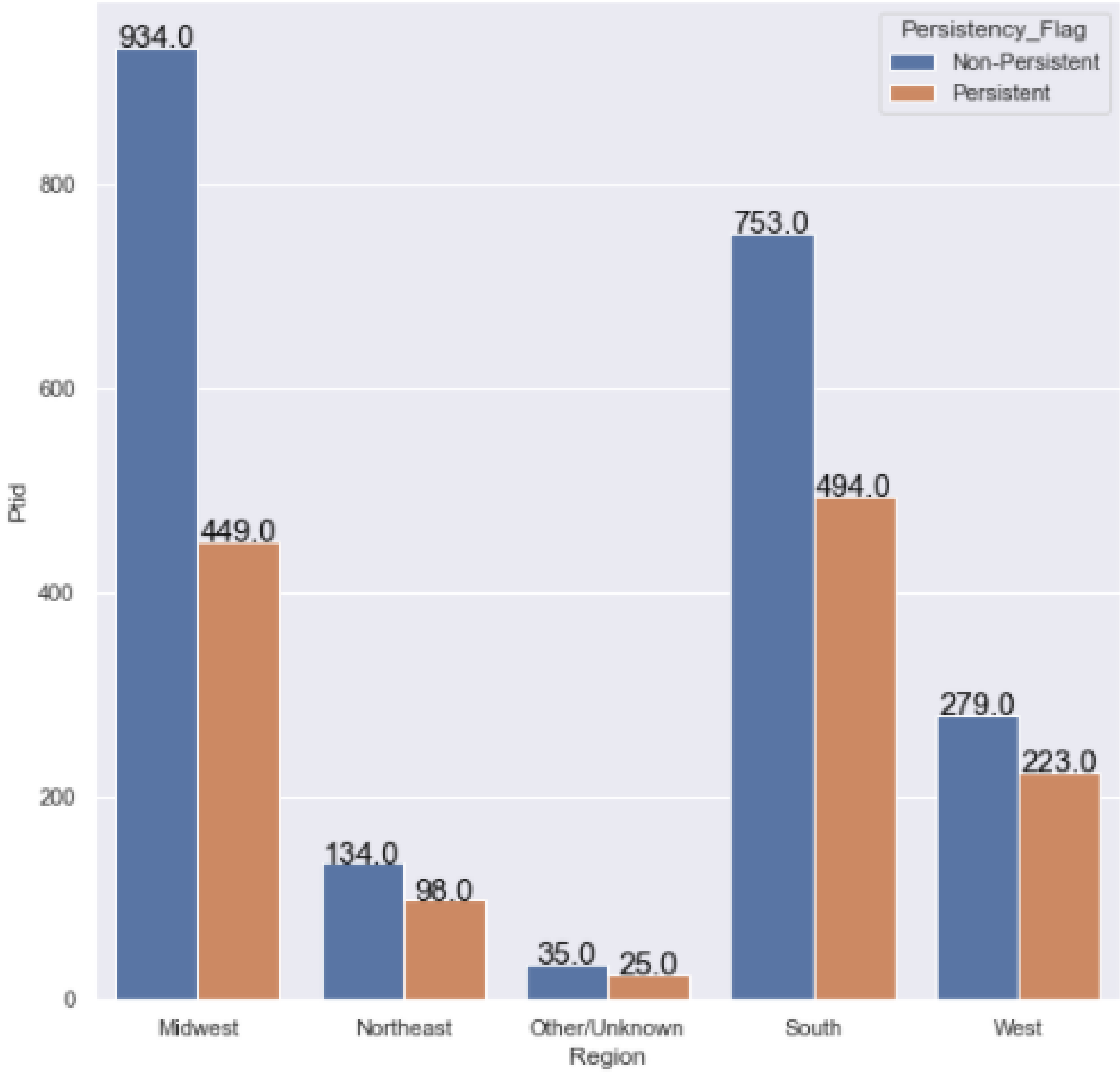
Hypotheses 1: Female patients are more persistent of a drug than male.

TRUE There are 1135 more persistent female patients than male patients.



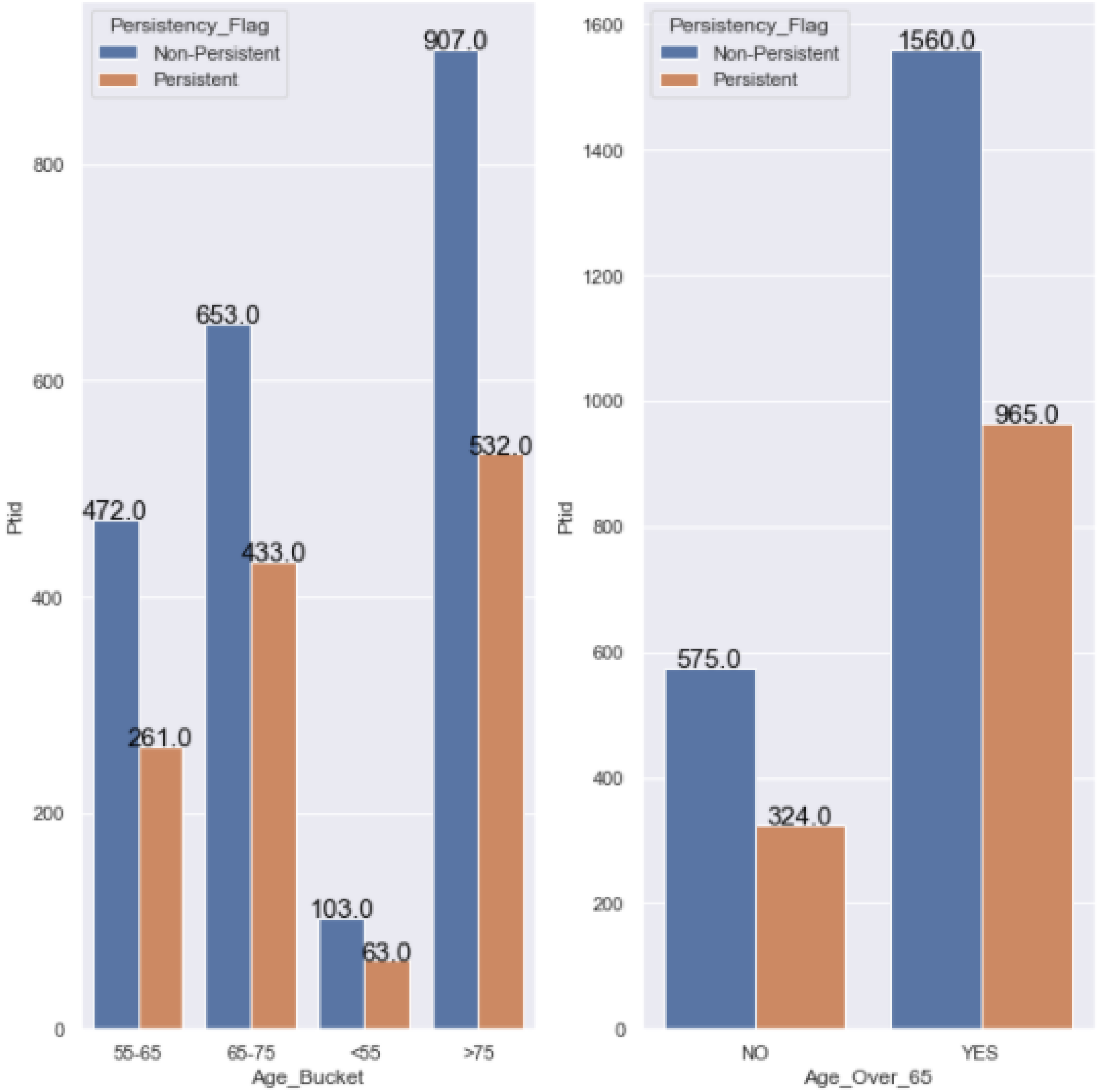
Hypotheses 2: Patients from Northeast are more persistent of a drug than patients from South.

FALSE Comparing patients from the Northeast and South, there are 396 more persistent patients from the South than from the Northeast.



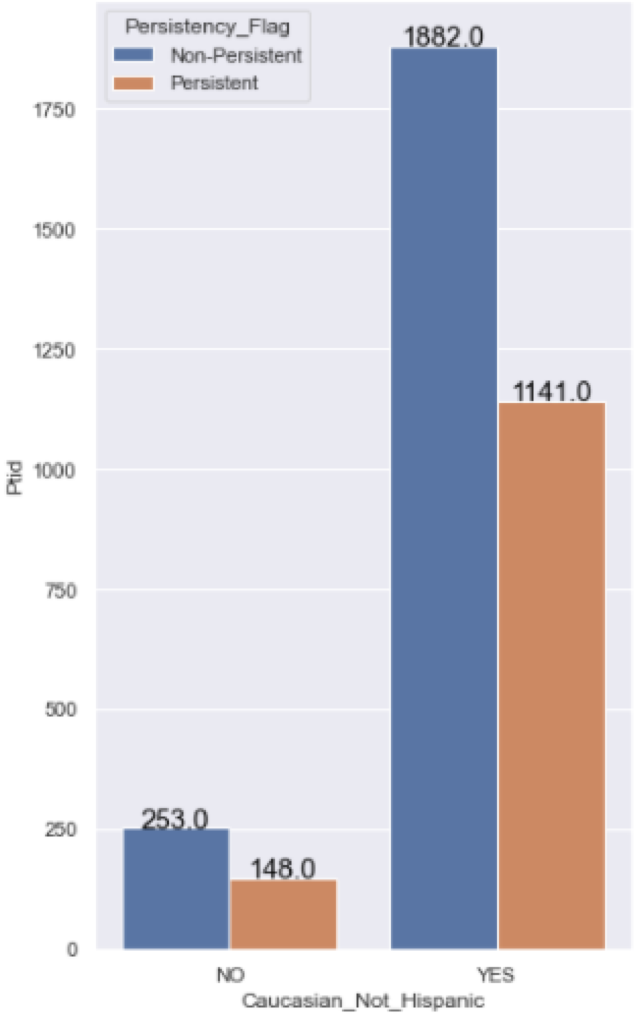
Hypotheses 3: Patients over 65 years of age are more persistent of a drug than patients 65 years of age or younger.

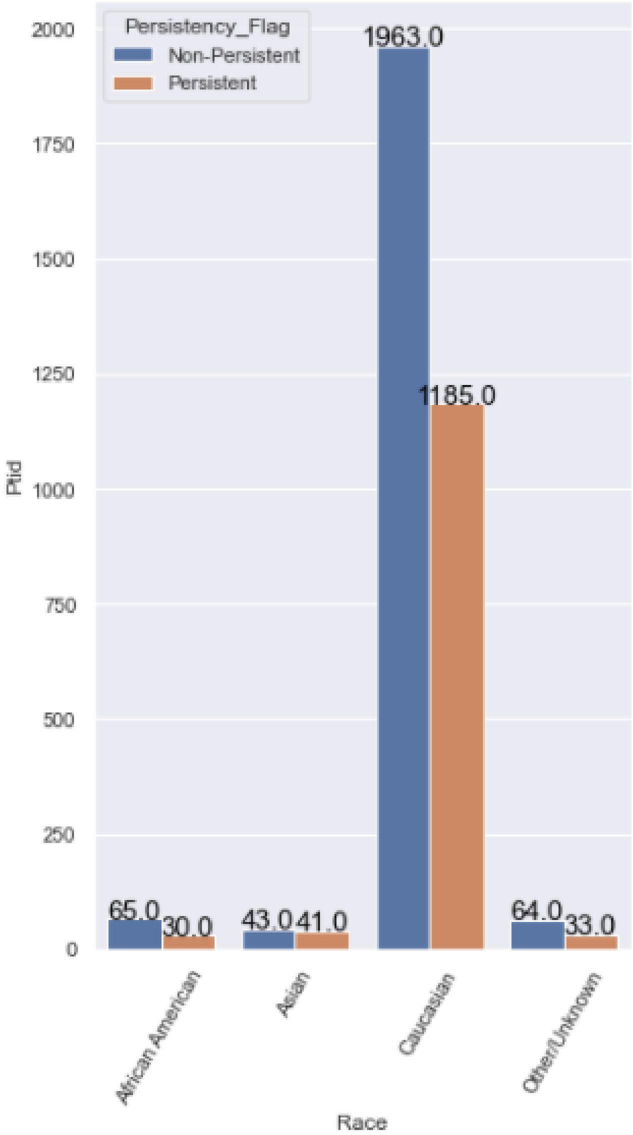
TRUE There are 641 more persistent patients over 65 years of age than patients younger than 65 years of age.

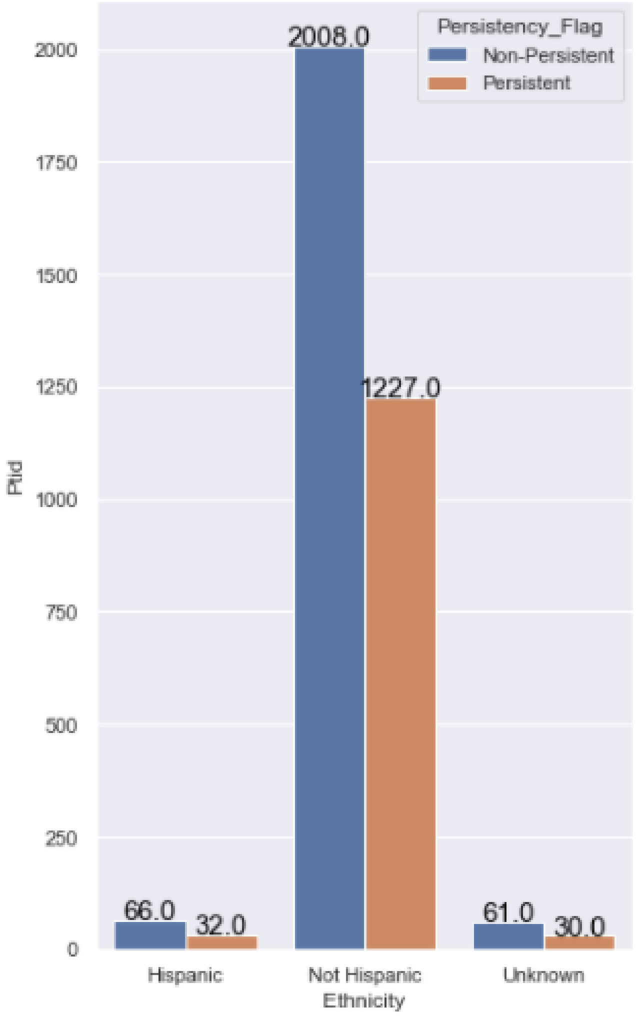


Hypotheses 4: Caucasian patients, not Hispanic are more persistent of a drug than patients with different race and ethnicity.

TRUE There are 993 more Caucasian & not Hispanic persistent patients than patients from different race and ethnicity.

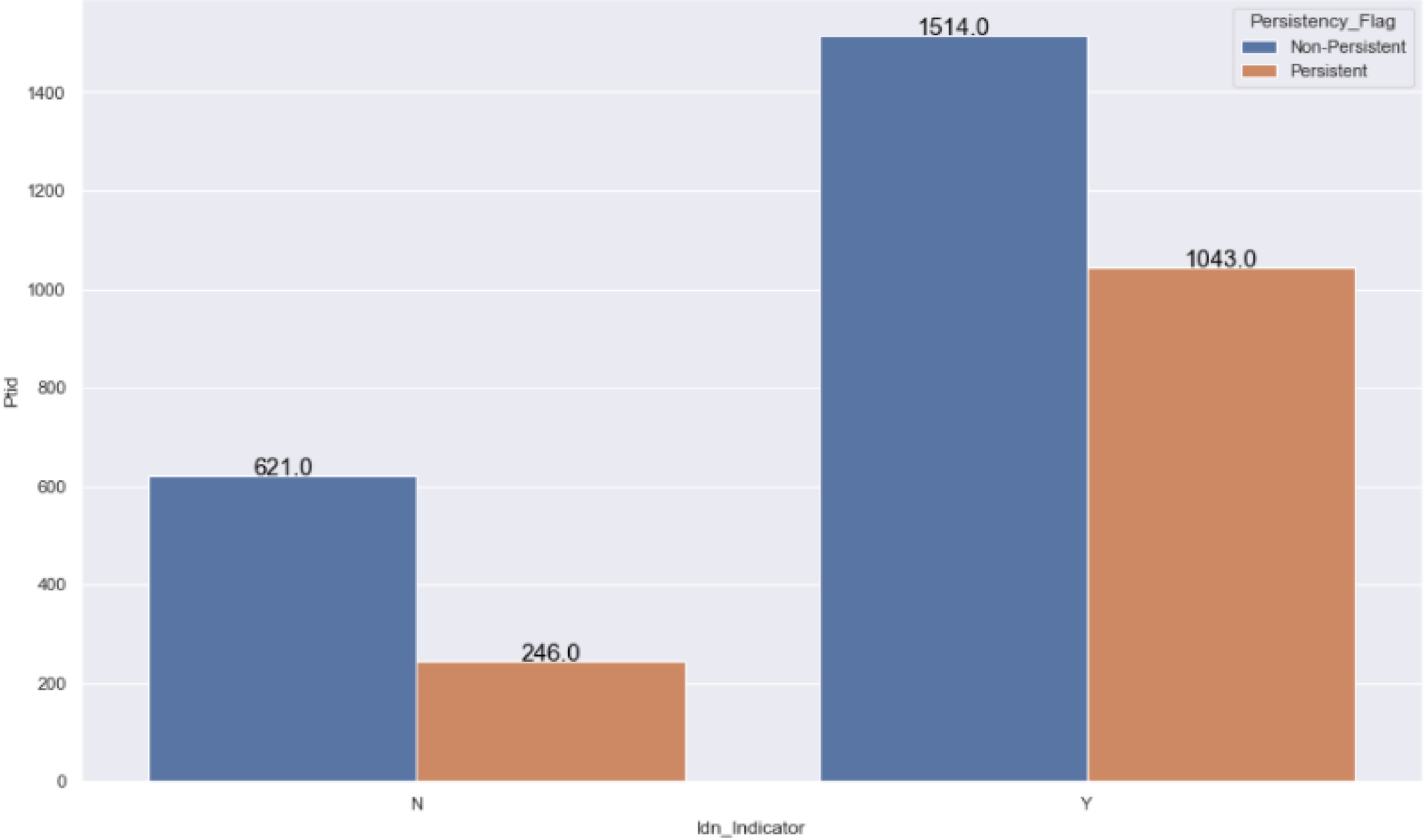






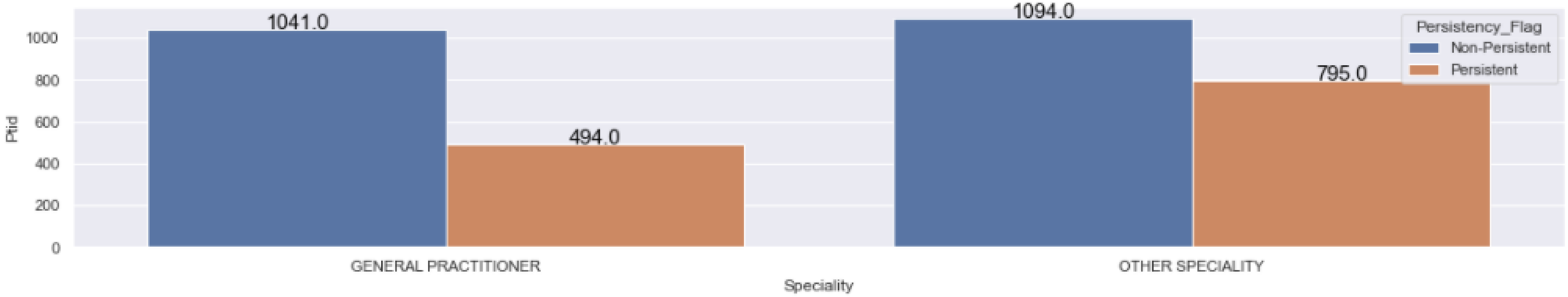
Hypotheses 5: Patients mapped to IDN are more persistent of a drug than patient not mapped.

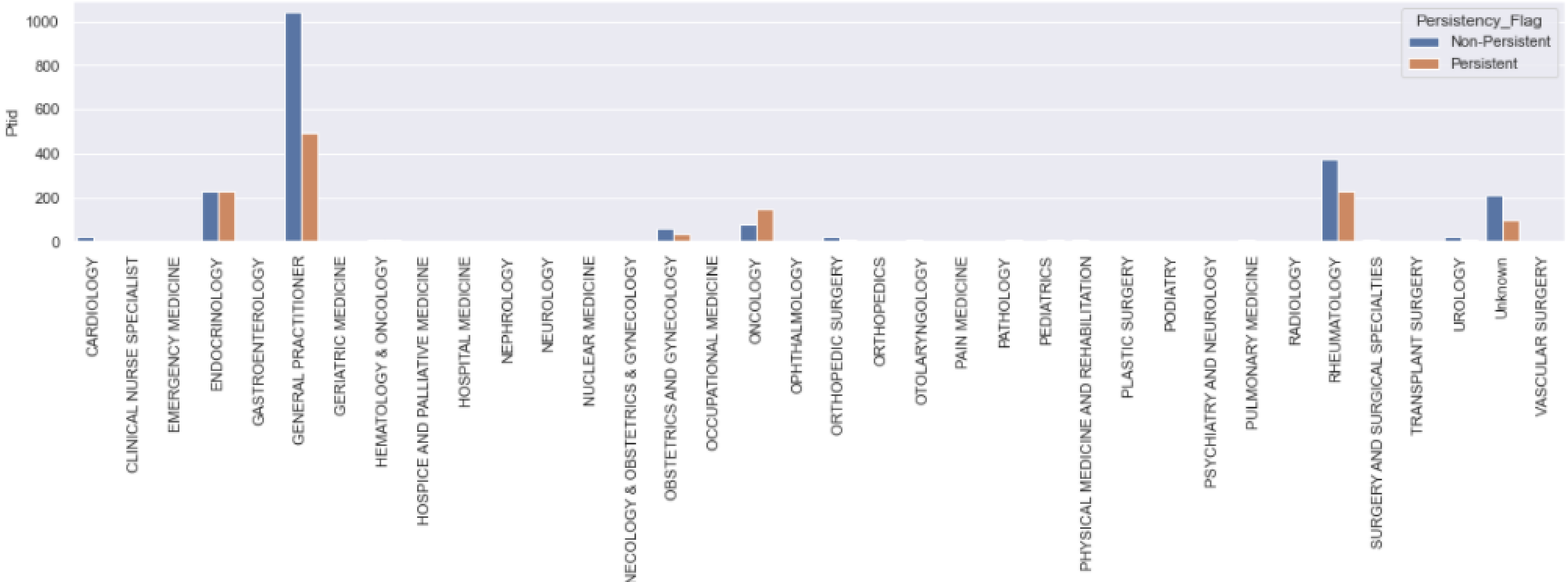
TRUE There are 797 more persistent patients mapped to IDN than persistent patients not mapped to IDN.



Hypotheses 6: Patients that received the drug prescription from General Practitioner Specialty are less persistent of a drug than patients that received the drug prescription from others Specialty.

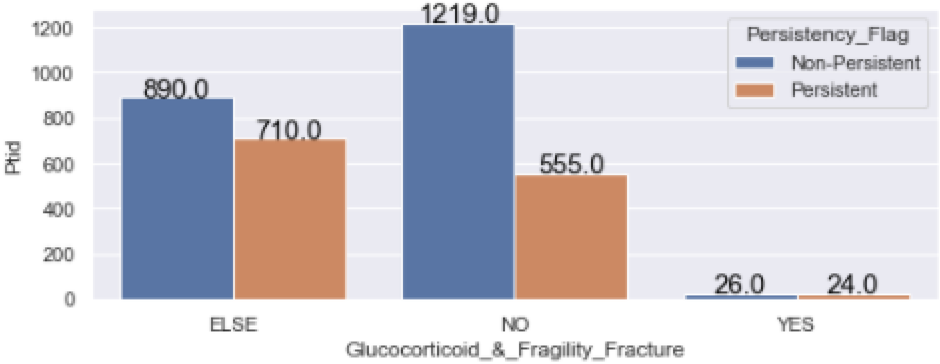
FALSE There are 301 more persistent patients that received the drug prescription from others Specialty than persistent patients that received the drug prescription from General Practitioner Specialty.

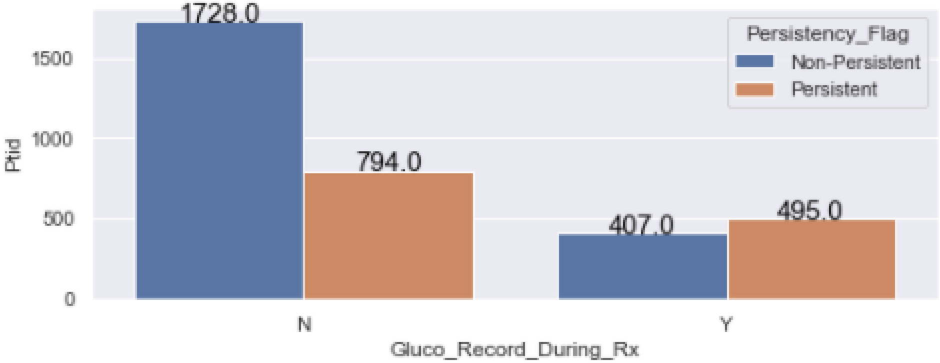


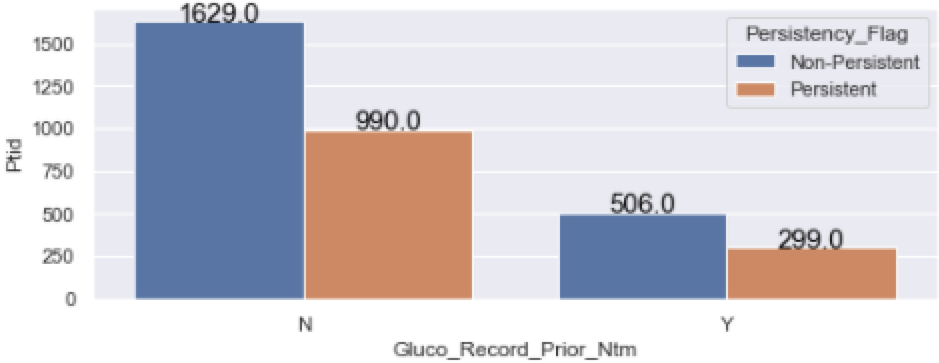


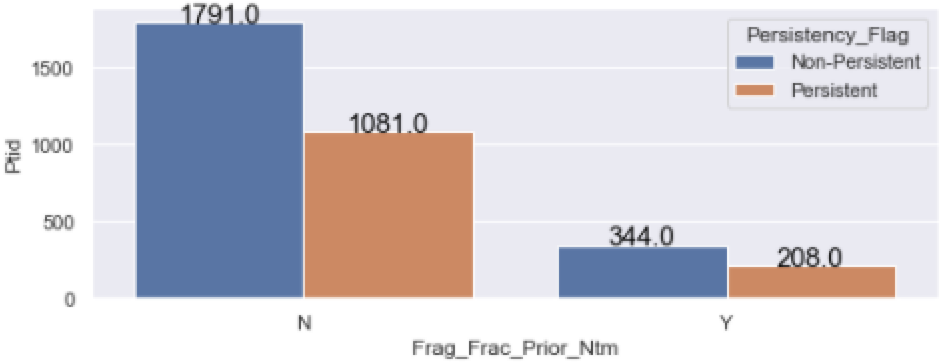
Hypotheses 7: Patients that used Glucocorticoid and had a Fragility Fracture, before and during the therapy, are more persistent of a drug than patients that not used Glucocorticoid either had a Fragility Fracture, in any situation.

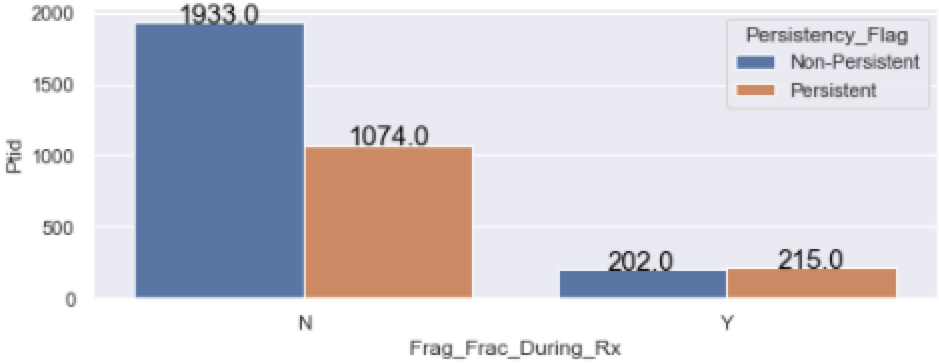
FALSE There are 531 more persistent patients that did not use Glucocorticoid neither had a Fragility Fracture than persistent patients that used Glucocorticoid and had a Fragility Fracture.





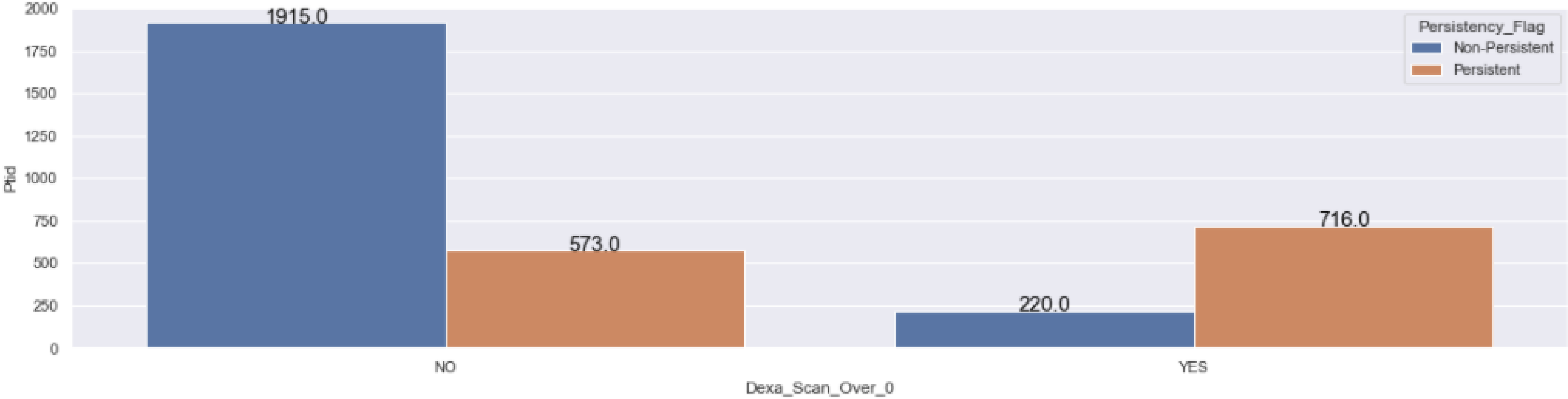


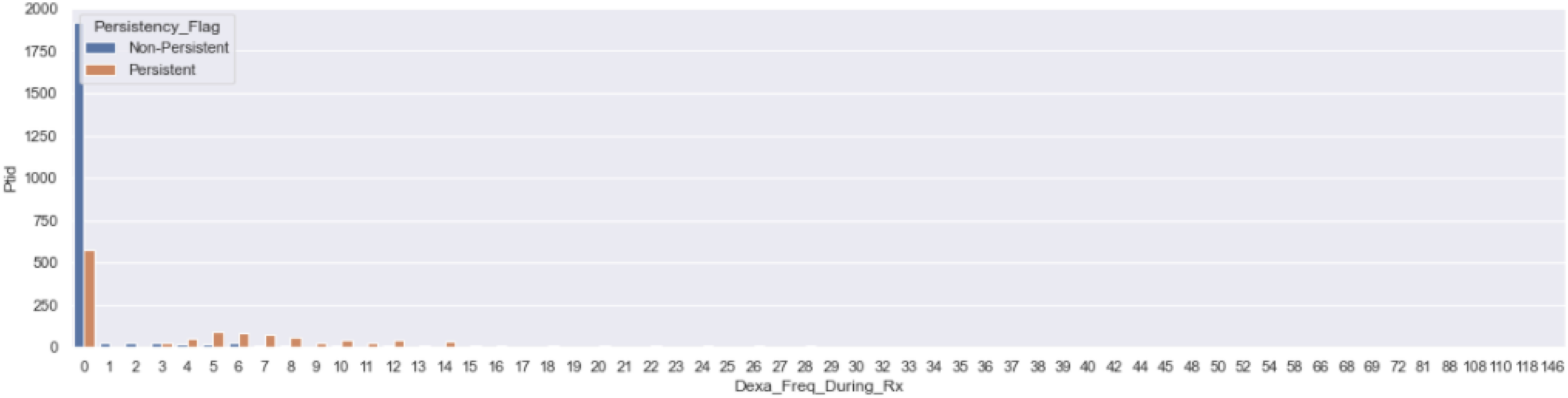




Hypotheses 8: Patients that had Dexa Scan more than 0 times are more persistent of a drug than patients that had Dexa Scan 0 times.

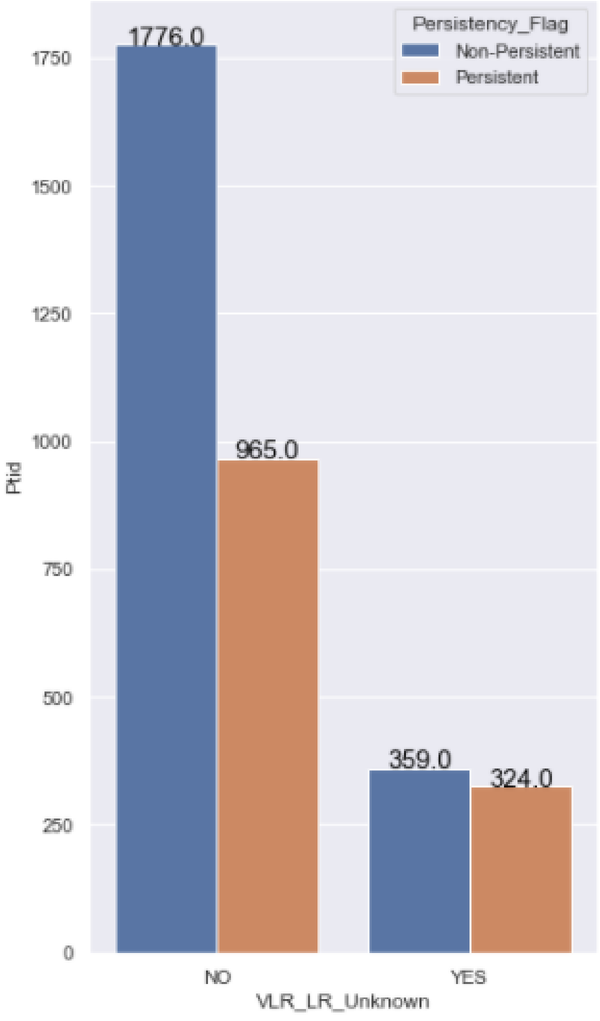
FALSE There are 185 more persistent patients that had Dexa Scan more than 0 times than persistent patients that that had Dexa Scan 0 times.

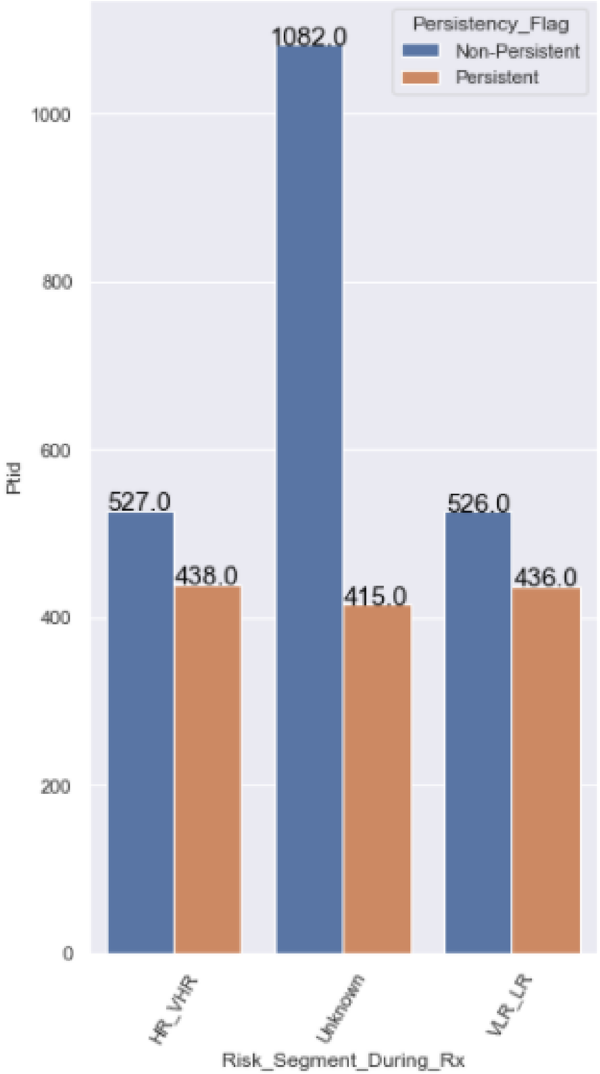


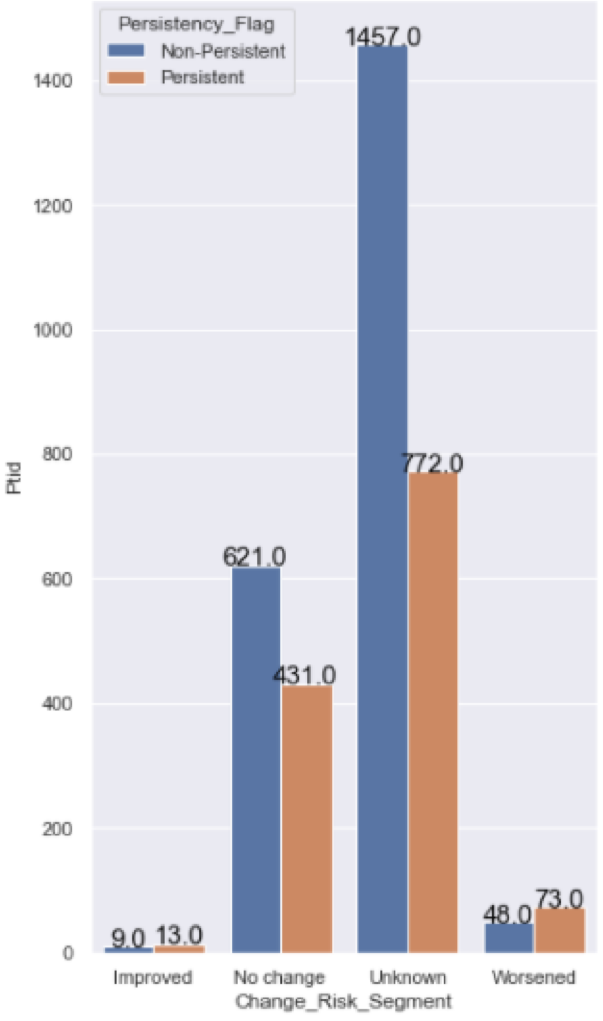


Hypotheses 9: Patients that presents VLR_LR as Risk Segment during the therapy and presents Unknown value in Risk Segment after the therapy are more persistent of a drug than patients that present other types of Risk Segment and Change in Risk Segment.

FALSE There are 641 more persistent patients that presents other types of Risk Segment and Change on Risk Segment than persistent patients that presents VLR_LR as Risk Segment during the therapy and presents Unknown value in Risk Segment after the therapy .

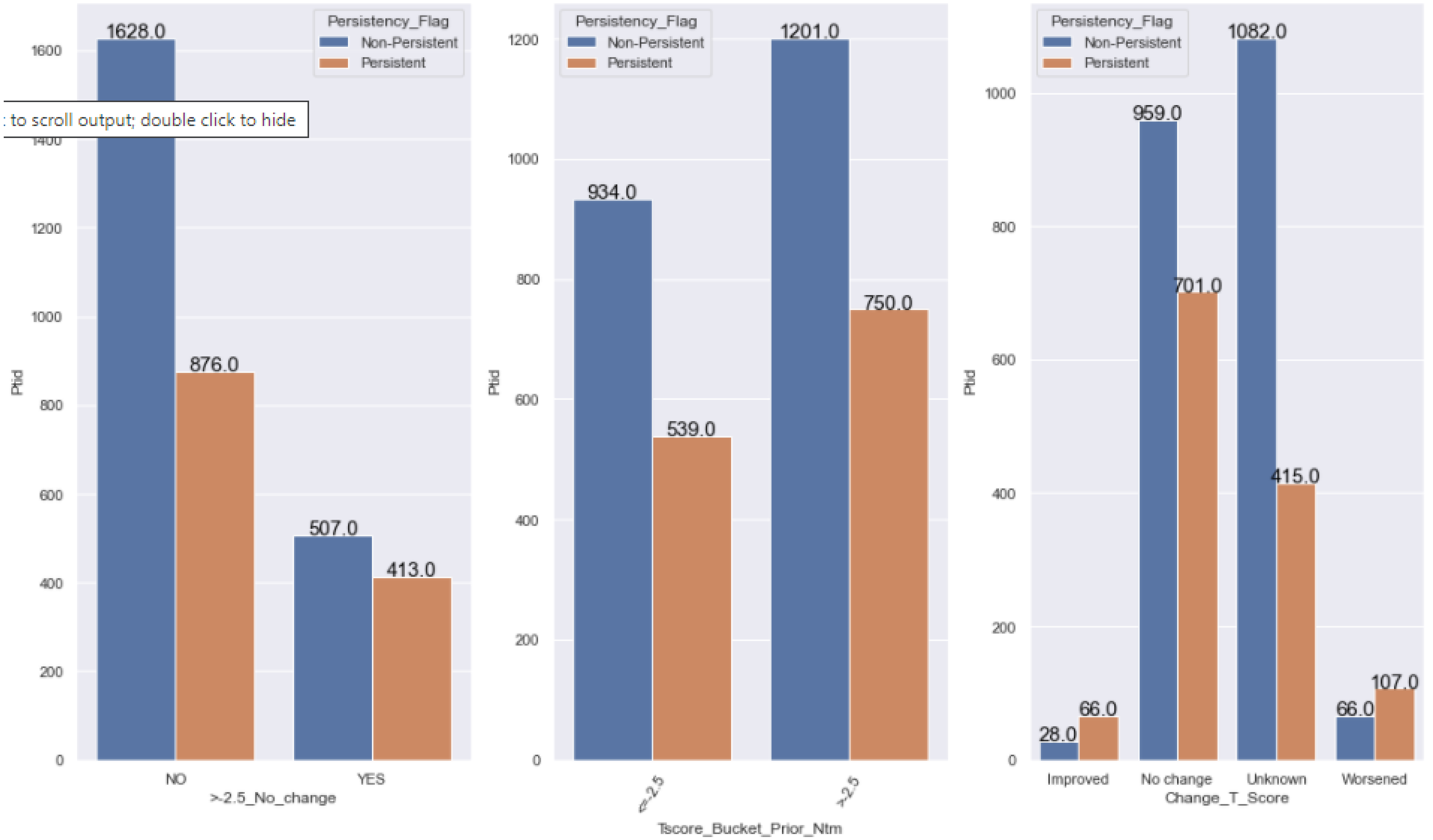






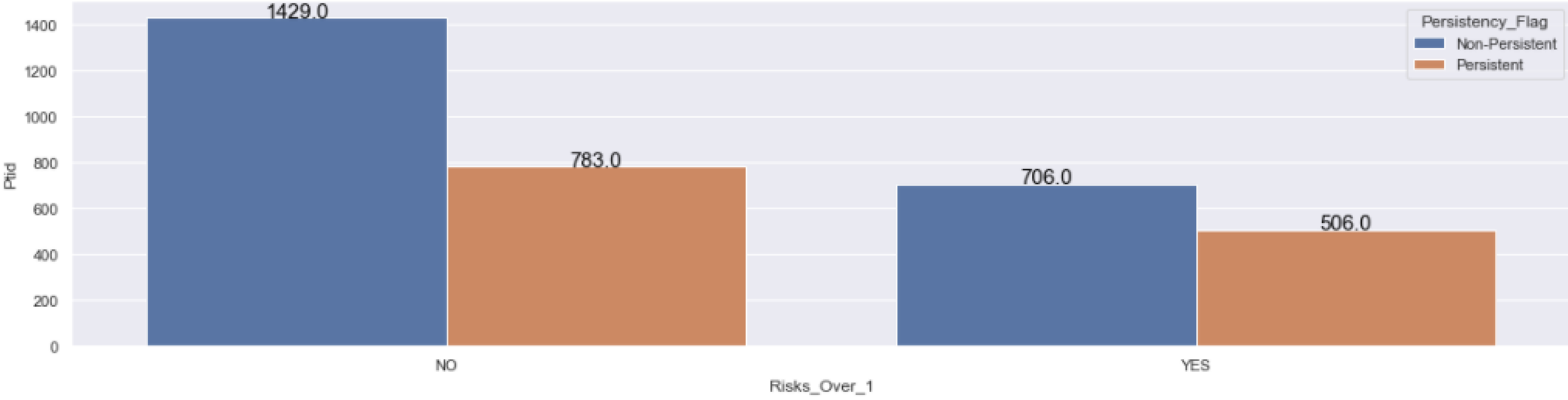
Hypotheses 10: Patients that presents T Socre of >-2.5 before the therapy and presents No Change T Score status after the therapy are less persistent of a drug than patients that presents other types of T Score and T Score status.

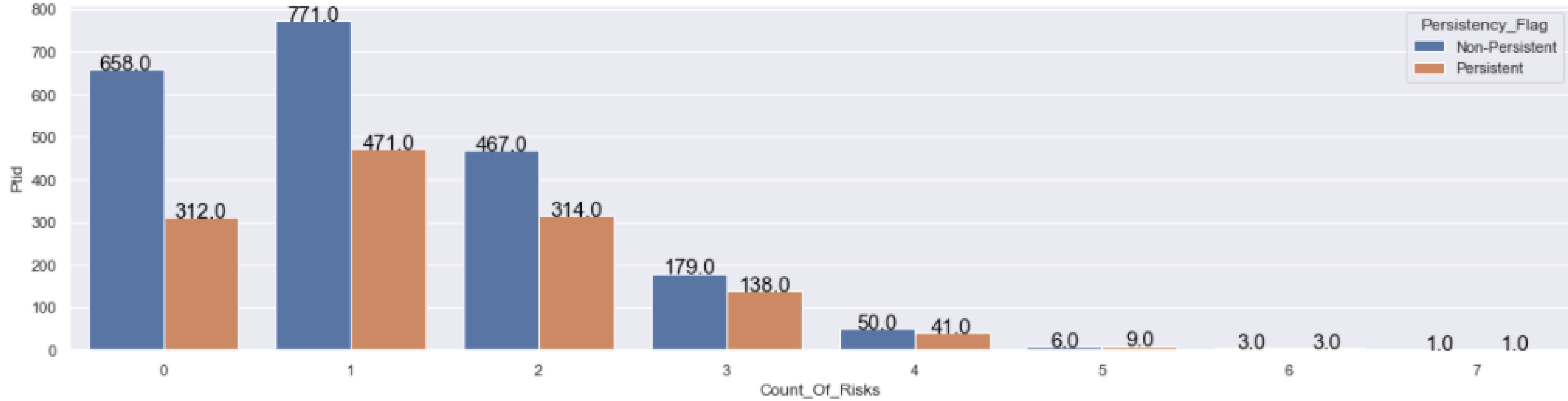
TRUE There are 463 more persistent patients that presents other types of T Score and Change on T Score than persistent patients that presents T Score of >-2.5 before the therapy and presents No Change T Score status after the therapy.



Hypotheses 11: Patients that presents the amount of risk factor higher than 1 are more persistent of a drug than patients that presents the amount of risk factor lower than 1.

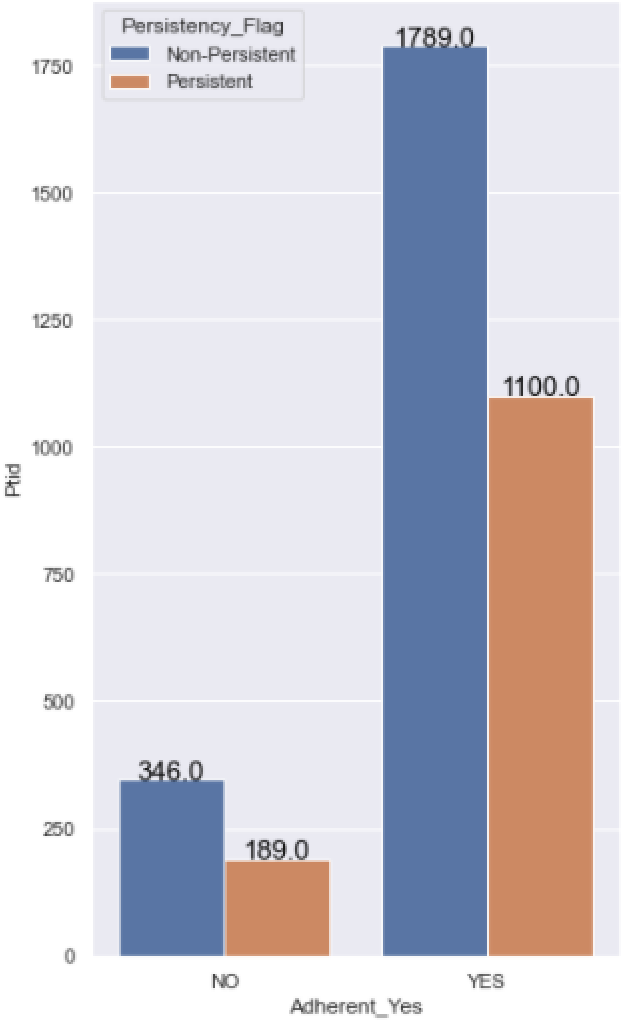
FALSE There are 277 more persistent patients that do not presents the amount of risk factor higher than 1 than persistent patients that presents the amount of risk factor higher than 1.

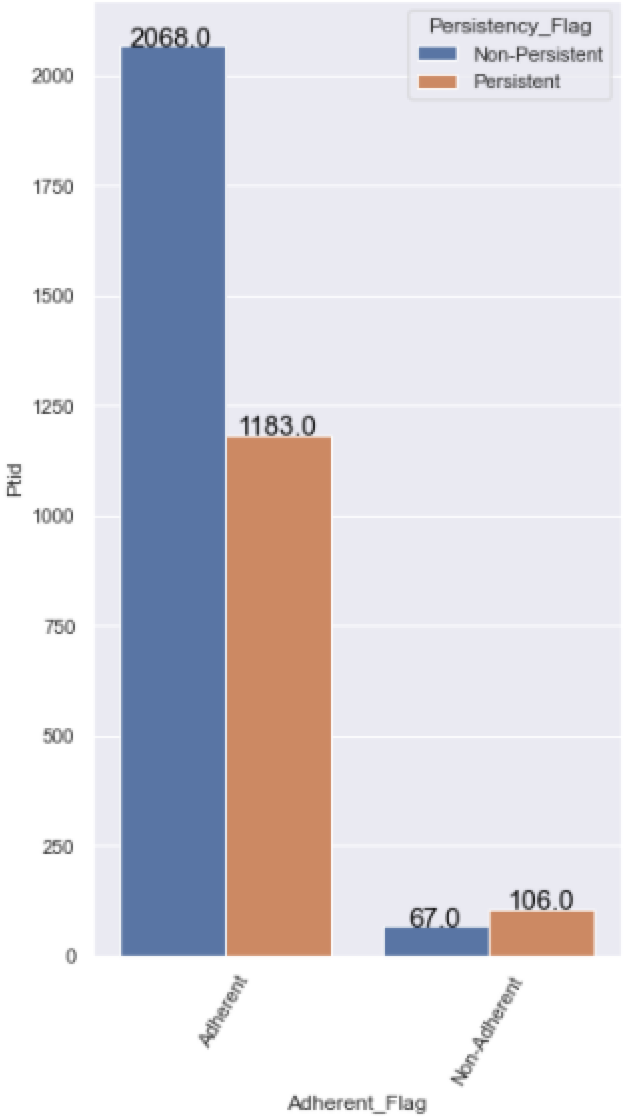


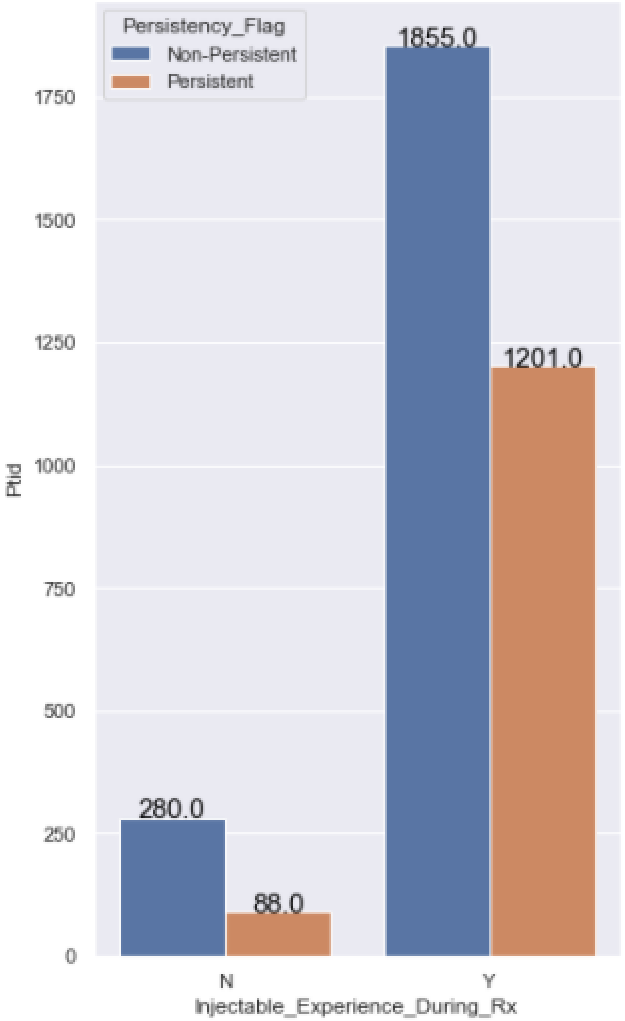


Hypotheses 12: Patients adherent for therapies and that used drugs during the therapy are more persistent of a drug than patients in different situations related to both attributes.

TRUE There are 189 more persistent patients adherent for therapies and that used drugs before the therapy than persistent are not adherent for therapies and that did not use drugs before the therapy.

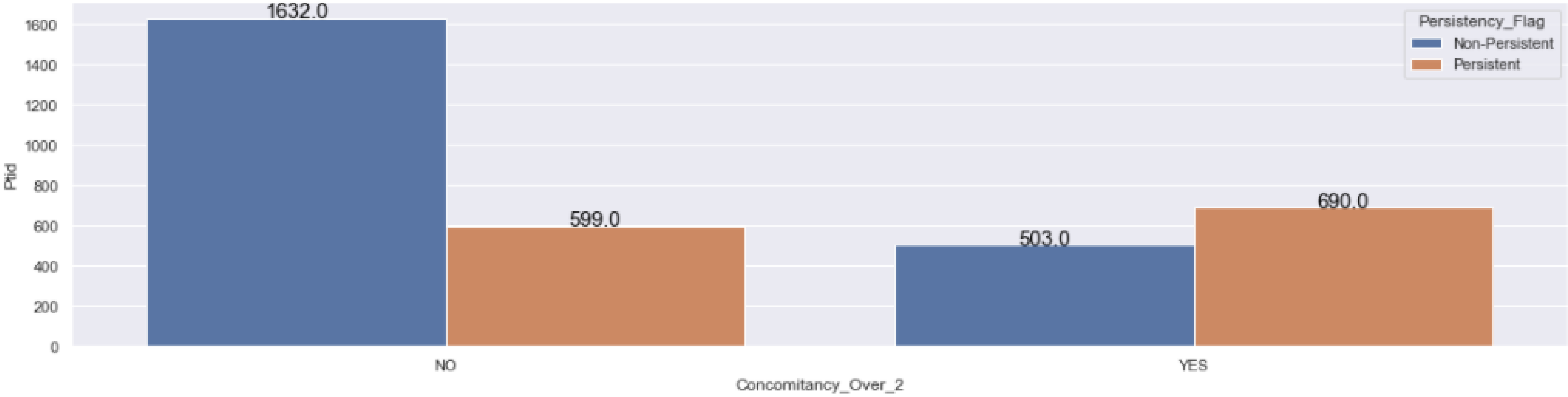


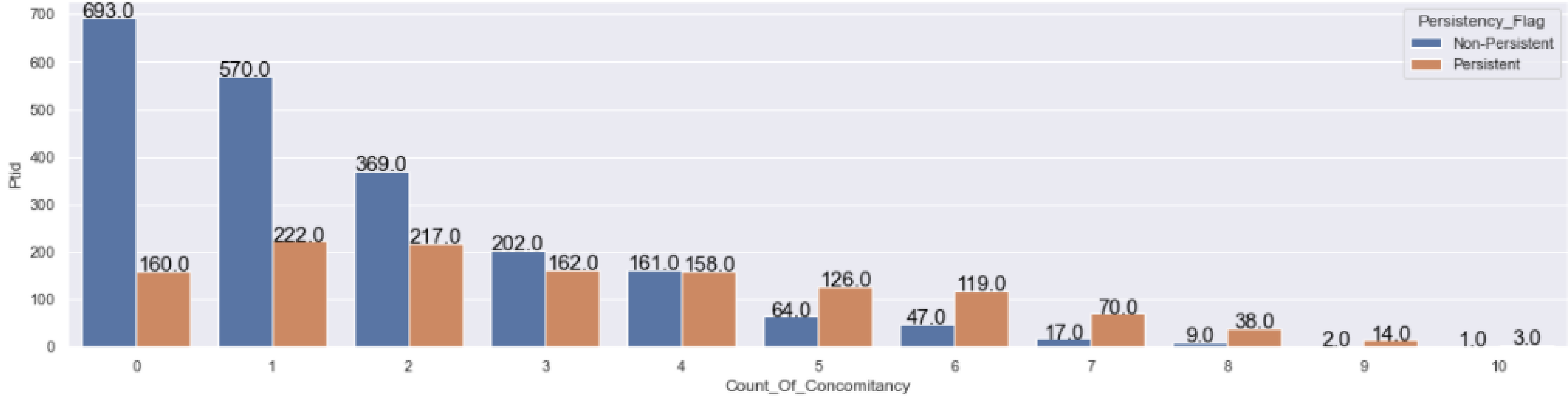




Hypotheses 13: Patients that used more than 2 concomitancy drugs before the therapy are less persistent of a drug than patients that used more than 2 concomitancy drugs before the therapy.

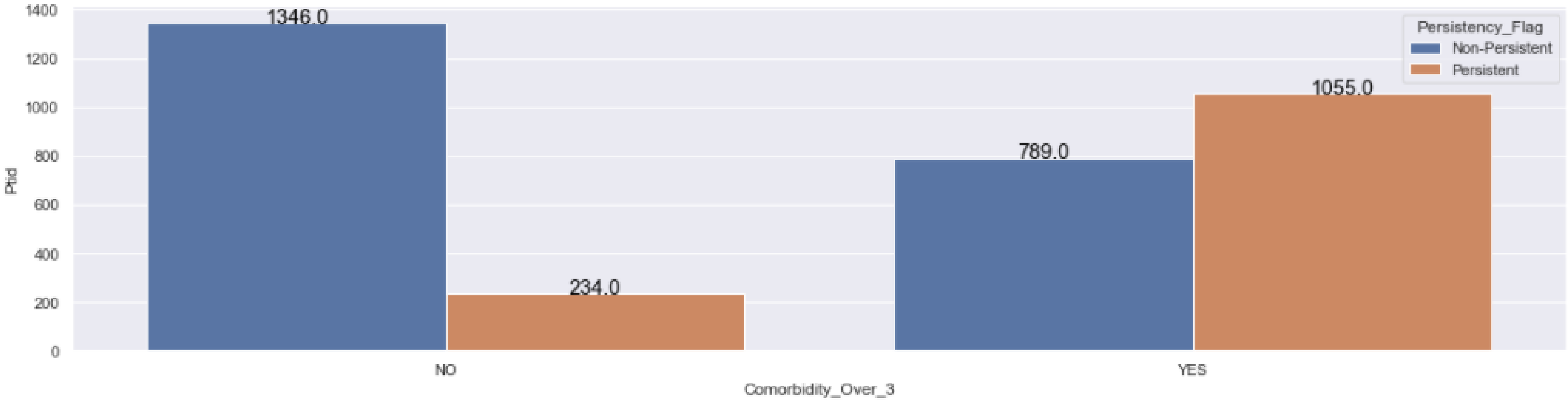
FALSE There are 91 more persistent patients that used more than 2 concomitancy drugs before the therapy persistent patients that used less than 3 concomitancy drugs before the therapy.

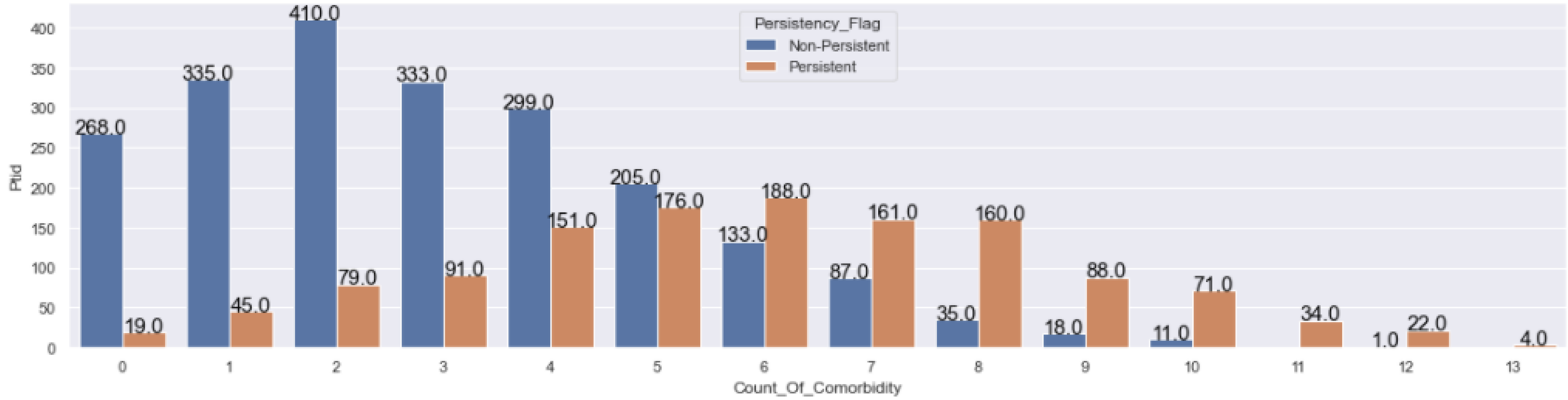




Hypotheses 14: Patients that presents more than 3 different types of comorbidity are more persistent of a drug than patients that present less than 3 different types of comorbidity.

TRUE There are 821 more persistent patients that presents more than 3 different types of comorbidity than persistent patients that do not presents more than 3 different types of comorbidity.

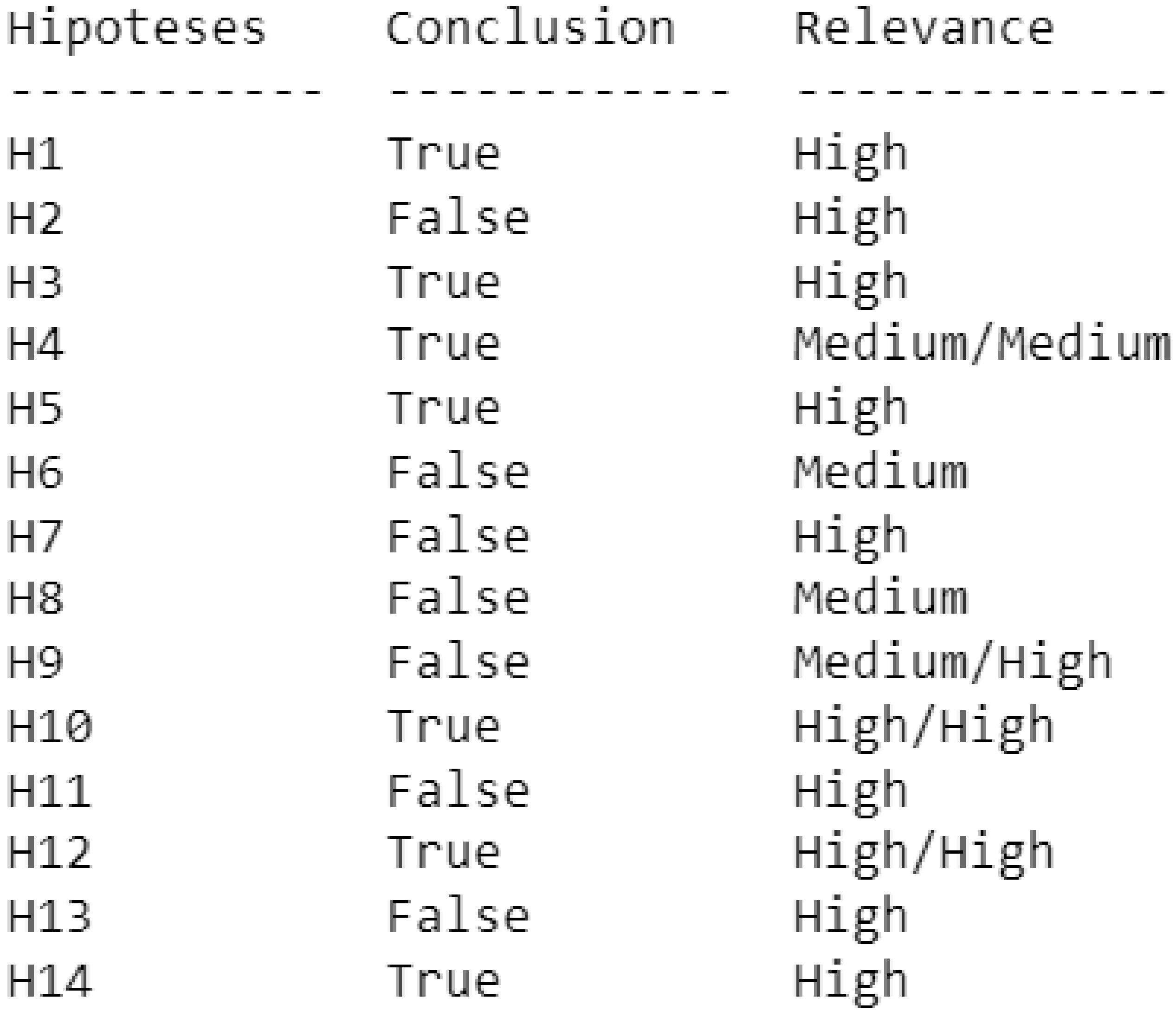




3.4.4. Summary of Hypothesis

This summary of hypotheses informs which hypotheses were validated as true and which were not validated.

Relevance indicates which attributes are important for training the classification model, i.e., it tells which attributes can help in achieving a high classification accuracy (higher number of successful classifications).



```
Hipoteses    Conclusion    Relevance
-----------  ------------  -------------
H1           True          High
H2           False         High
H3           True          High
H4           True          Medium/Medium
H5           True          High
H6           False         Medium
H7           False         High
H8           False         Medium
H9           False         Medium/High
H10          True          High/High
H11          False         High
H12          True          High/High
H13          False         High
H14          True          High
```

3.5. Multivariate Analysis – Numerical Variables

In this numerical analysis matrix, all rows and all columns are all numerical attributes.

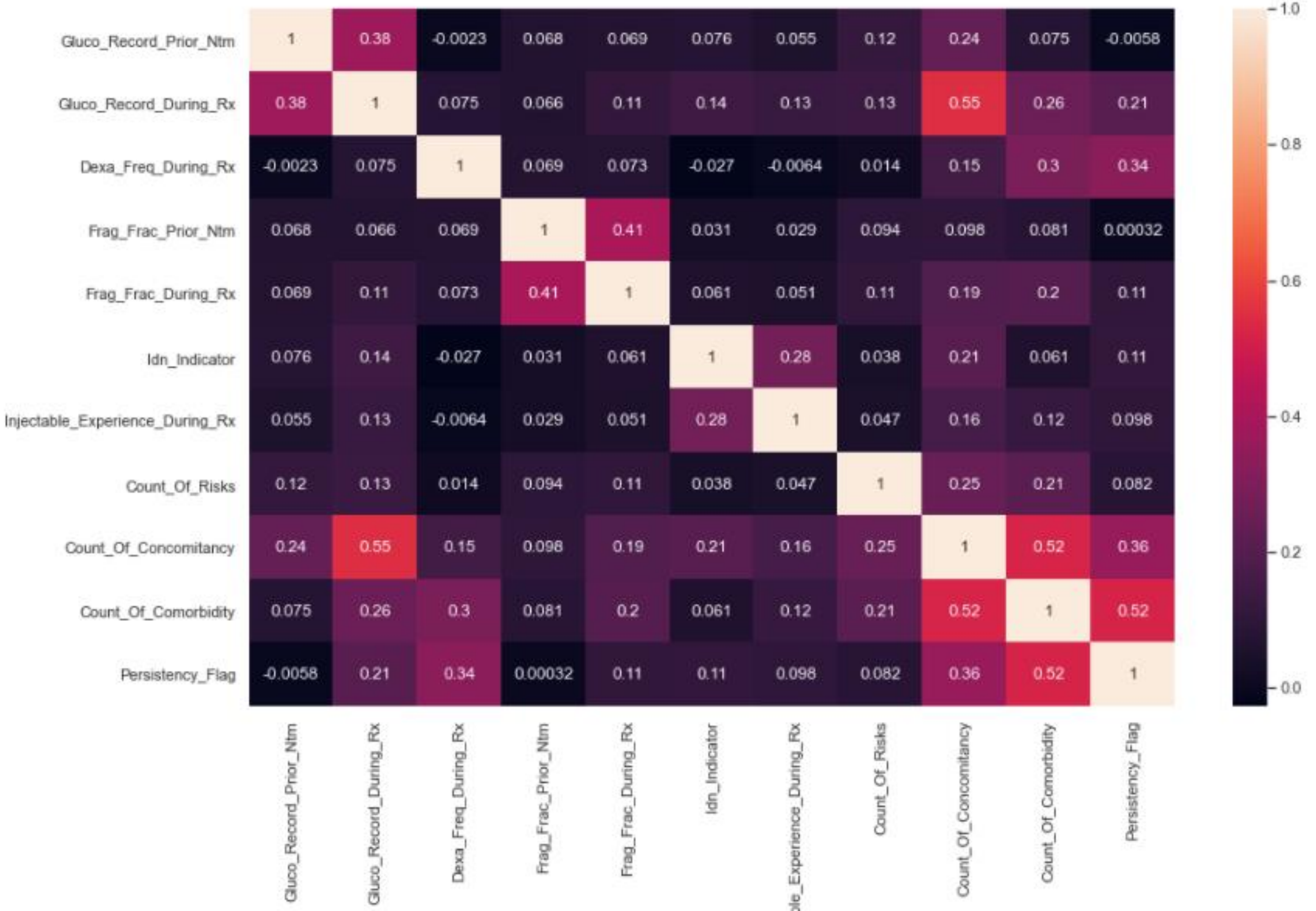
This graph informs the correlation between the attributes.

The closer to the white the higher is the positive correlation between them.

The closer to black the higher the negative correlation between them.

If the attributes present a high correlation between them (both positive and negative) it means that they have the same type of information. With this we can eliminate one of them, reducing the dimensionality of the dataset.

This dimensionality reduction represents better chances for the classification model to obtain a high accuracy (higher number of successful classifications).



	Gluco_Record_Prior_Ntm	Gluco_Record_During_Rx	Dexa_Freq_During_Rx	Frag_Frac_Prior_Ntm	Frag_Frac_During_Rx	Idn_Indicator	le_Experience_During_Rx	Count_Of_Risks	Count_Of_Concomitancy	Count_Of_Comorbidity	Persistency_Flag
Gluco_Record_Prior_Ntm	1	0.38	-0.0023	0.068	0.069	0.076	0.055	0.12	0.24	0.075	-0.0058
Gluco_Record_During_Rx	0.38	1	0.075	0.066	0.11	0.14	0.13	0.13	0.55	0.26	0.21
Dexa_Freq_During_Rx	-0.0023	0.075	1	0.069	0.073	-0.027	-0.0064	0.014	0.15	0.3	0.34
Frag_Frac_Prior_Ntm	0.068	0.066	0.069	1	0.41	0.031	0.029	0.094	0.098	0.081	0.00032
Frag_Frac_During_Rx	0.069	0.11	0.073	0.41	1	0.061	0.051	0.11	0.19	0.2	0.11
Idn_Indicator	0.076	0.14	-0.027	0.031	0.061	1	0.28	0.038	0.21	0.061	0.11
Injectable_Experience_During_Rx	0.055	0.13	-0.0064	0.029	0.051	0.28	1	0.047	0.16	0.12	0.098
Count_Of_Risks	0.12	0.13	0.014	0.094	0.11	0.038	0.047	1	0.25	0.21	0.082
Count_Of_Concomitancy	0.24	0.55	0.15	0.098	0.19	0.21	0.16	0.25	1	0.52	0.36
Count_Of_Comorbidity	0.075	0.26	0.3	0.081	0.2	0.061	0.12	0.21	0.52	1	0.52
Persistency_Flag	-0.0058	0.21	0.34	0.00032	0.11	0.11	0.098	0.082	0.36	0.52	1

4. Conclusion & Demonstration

Here it is presented the hypothesis and insights conclusion, also presents the performance of the machine learning classification model.

4.1. Hypothesis and Insights Conclusion

Based on all the Insights created from this dataset, patients who have a higher probability of persisting of a drug for the entire treatment, have the following characteristics:

- Female;
- Patients older than 65 years;
- Caucasian non-Hispanic Patients;
- Patients mapped to IDN;
- Patients that received the drug prescription from General Practitioner Specialty;
- Patients who did not use Glucocorticoid and had Fragility Fracture;
- Patients who had no Dexa Scan;
- Patient’s adherent for therapies and that used drugs before the therapy.

4.2.1. Data Preparation

In this step we demonstrate the performance of the chosen model witch classifies the target variable Persistency Flag.

To do such demonstration, we had to do a data preparation of the dataset, in order to improve the accuracy of the classification model. The data preparation methods are:

- Rescaling variables: Reduce the importance of the variables with higher range. Variables with a high range can prejudice the classification model training.

- Transforming variables: Transform all categorical variables in numerical variables. This measure is taken because the classification model just work with numerical variables.

- Feature Selection: Select the most important variables for the model.

- Balancing Dataset: Equalizes the number of classes, making the number of values of the class ‘Persistent’ equal to the number of values of the class ‘Non – Persistent’.

4.2.2. Machine Learning Models

The target variable Persistent Flag is of type binary class, i.e., it has only two classification values, in this case, Persistent or Non-Persistent.

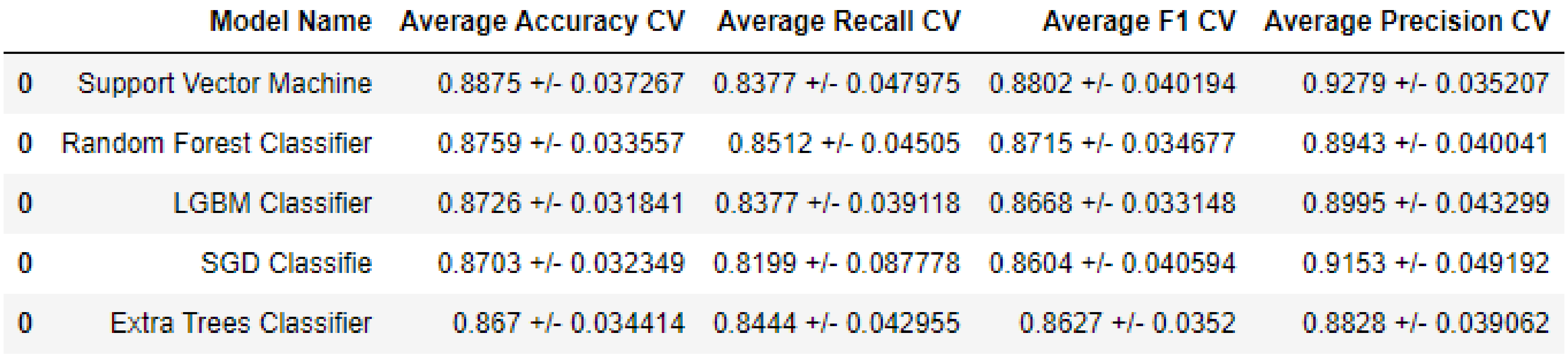
Due to this fact some Machine Learning algorithms used at this business problem are:

- **Extra Trees Classifier;**
- **LGBM Classifier:**
- **Stochastic Gradient Descent;**
- **Random Forest;**
- **Support Vector Machine.**

Those Machine Learning algorithms generally present very good results with this type of classification.

4.2.3. Machine Learning Models Performance

This slide presents a comparison among all ML models applied on this dataset.



	Model Name	Average Accuracy CV	Average Recall CV	Average F1 CV	Average Precision CV
0	Support Vector Machine	0.8875 +/- 0.037267	0.8377 +/- 0.047975	0.8802 +/- 0.040194	0.9279 +/- 0.035207
0	Random Forest Classifier	0.8759 +/- 0.033557	0.8512 +/- 0.04505	0.8715 +/- 0.034677	0.8943 +/- 0.040041
0	LGBM Classifier	0.8726 +/- 0.031841	0.8377 +/- 0.039118	0.8668 +/- 0.033148	0.8995 +/- 0.043299
0	SGD Classifie	0.8703 +/- 0.032349	0.8199 +/- 0.087778	0.8604 +/- 0.040594	0.9153 +/- 0.049192
0	Extra Trees Classifier	0.867 +/- 0.034414	0.8444 +/- 0.042955	0.8627 +/- 0.0352	0.8828 +/- 0.039062

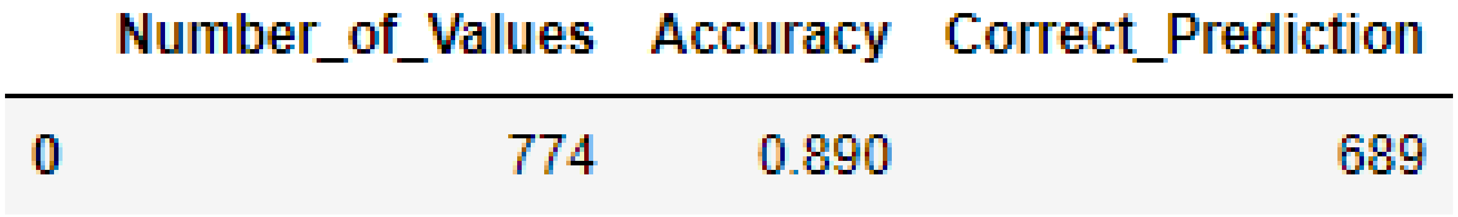
Despite the model **Support Vector Machine** do not presents the highest value for 'Average Recall CV (Cross Validation), the model **Support Vector Machine** presents the highest values for 'Average Accuracy CV', 'Average F1 CV' and 'Average Precision CV'.

Due those facts the model **Support Vector Machine** was chosen as final classification model.

4.2.4. Support Vector Machine – Business Performance

- **Accuracy**

With the accuracy of the ML model it is possible to find out the exact number of correct predictions made by the same model.



	Number_of_Values	Accuracy	Correct_Prediction
0	774	0.890	689

From **774** values, the ML model was able to correctly predict **689** values of the attribute 'Persistency_Flag'.

4.2.5. Support Vector Machine – Business Performance

- **Confusion Matrix & Classification Report**

With the Confusion Matrix it is possible to verify how many values of the attribute 'Persistency_Flag' were correctly and erroneously classified in each class, in other words, how many values equal to Persistent (1) and equal to Non-Persistent (0) were correctly classified, and how many of each class were not.

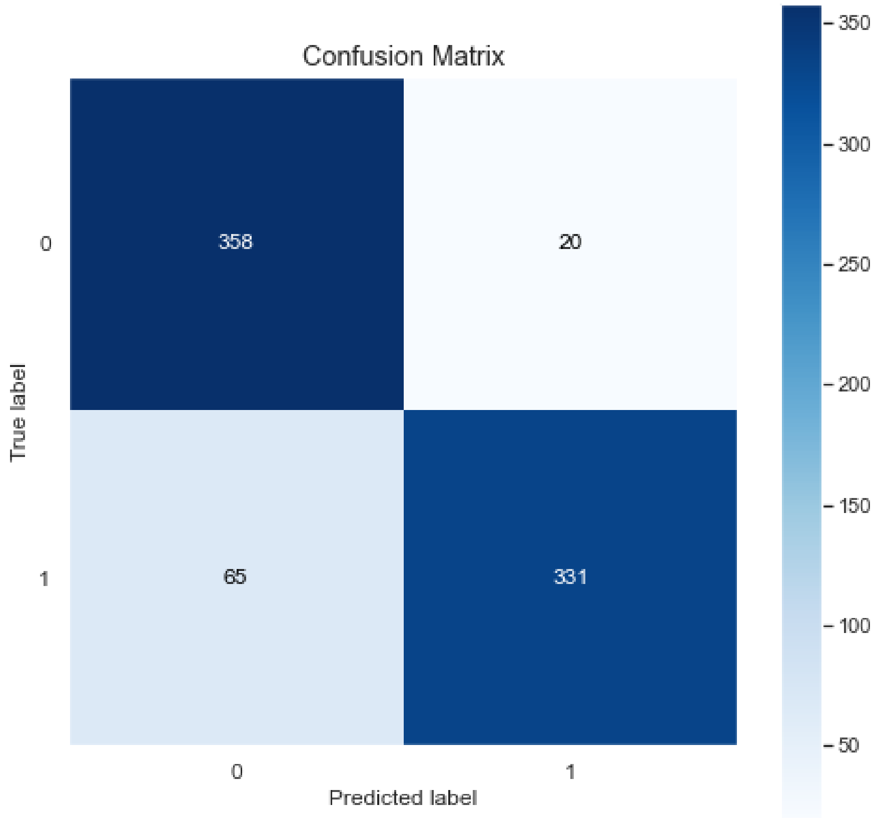
According to the classification report and the confusion matrix:

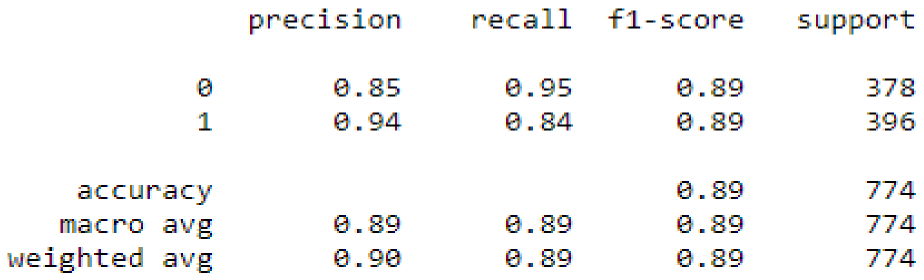
358 values that are originally class Non-Persistent (0) have been **correctly classified** as Non-Persistent, which represents **95%** of the Non-Persistent class;

331 values that are originally class Persistent (1) have been **correctly classified** as Persistent, which represents **84%** of the Persistent class;

20 values that are originally class Non-Persistent (0) have been **incorrectly classified** as Persistent;

65 values that are originally class Persistent (1) have been **incorrectly classified** as Non-Persistent.




```
              precision    recall  f1-score   support

           0       0.85      0.95      0.89       378
           1       0.94      0.84      0.89       396

    accuracy                           0.89       774
   macro avg       0.89      0.89      0.89       774
weighted avg       0.90      0.89      0.89       774
```

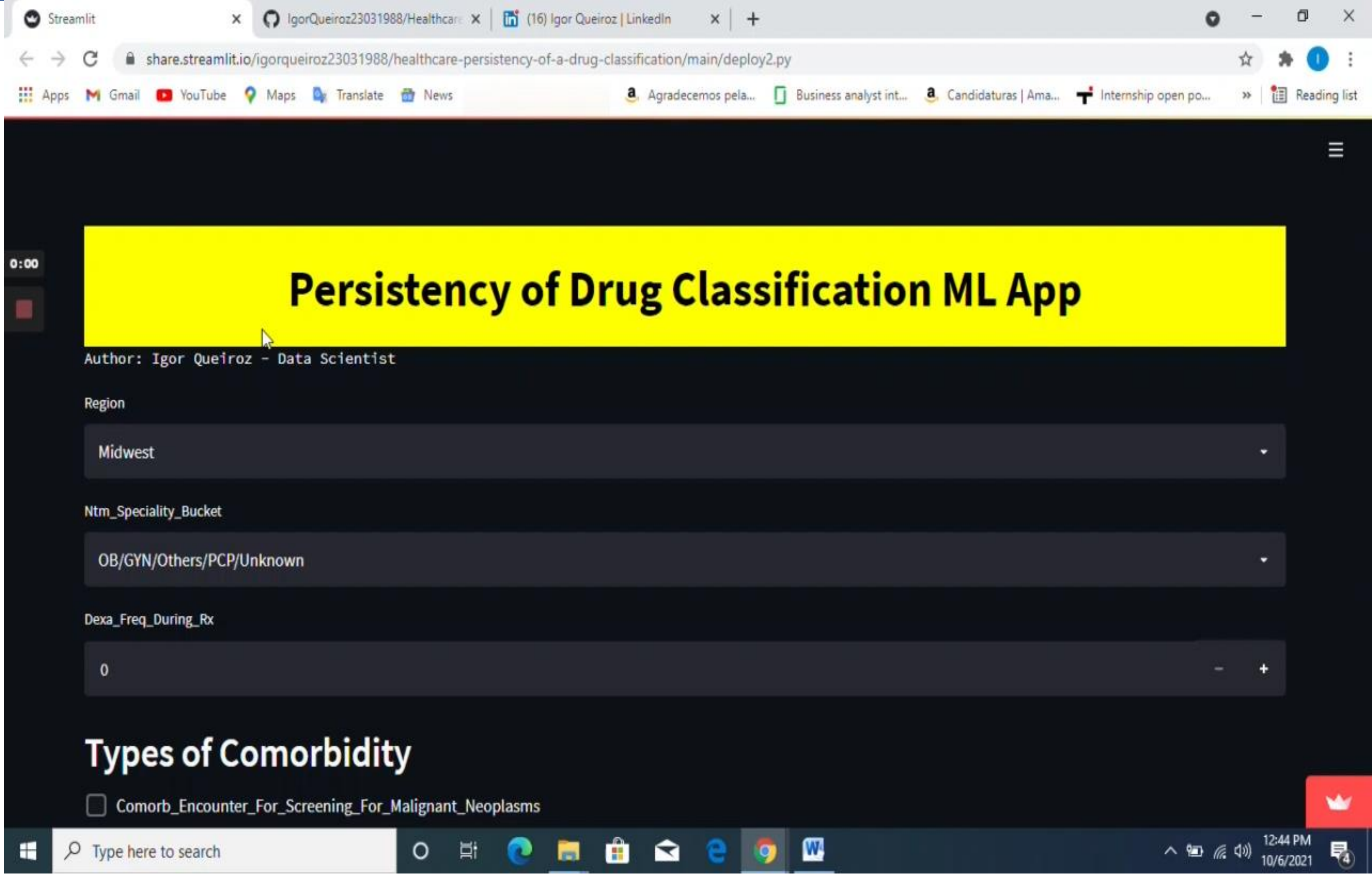
4.2.6. Persistency of a Drug Classification ML WebApp

- Here is presented a WebApp created in order to classify if futures patients will use the drugs during the entire treatment or if they won't, if they are going to stop to take the drug in the middle of the treatment, for example.

- It was created with the most important patient characteristics with the aim of get a high accuracy classification.

- It is possible to check the WebApp in this link: https://share.streamlit.io/igorqueiroz23031988/healthcare-persistency-of-a-drug-classification/main/deploy2.py.

- Below, there is a video explaining about the business problem and how the ML model works, also with examples of predictions.



Persistency of Drug Classification ML App

Author: Igor Queiroz - Data Scientist

Region

Midwest

Ntm_Speciality_Bucket

OB/GYN/Others/PCP/Unknown

Dexa_Freq_During_Rx

0

Types of Comorbidity

☐ Comorb_Encounter_For_Screening_For_Malignant_Neoplasms

5. Next Steps

- Business User Model Workshop;
- Collect Usability Feedback;
- Increase Model Accuracy by 10%.

Thank You