**Task 20 - Classify Doctor rank (Senior/Not Senior) to rehospitalization**

The objective of this task is to infer and understand if the doctors' experience and professional observance has any effect on getting rehospitalized and the duration of the hospitalizations.

In order to proceed with this task, we created an EDA for the 2 additional important datasets: hDoctor which tells us the doctors' id who released patients, and Doctors' rank if they are Seniors.That way we can understand the data like the previous task and prepare the dataset for model training by cleaning and transforming it.

**Preprocessing**

* Normalizing Date Columns: Many columns had date values that needed to be standardized to ensure consistency across the dataset. We used the normalize\_date\_column function to convert date fields into a usable format.
* Encoding Categorical Variables: Some features were categorical and needed to be encoded for machine learning models. We applied label encoding and one-hot encoding where appropriate using the encode\_columns and encode\_and\_correlate functions.
* Merging Datasets: We integrated various datasets, including hospitalization records and rehospitalization status, using left joins to append relevant patient status information.
* Based on Doctor rank data(Senior/Not Senior), the ranks are split almost equally to 4 categories: Yes, No, ? and Depends from which date.
* Based on the PIE chart, They're split almost equally in amount of doctors.

Through Gradient Boosting Model, we trained the data to predict rehospitalization depentant on the doctor rank.

**Conclusions:**

The accuracy is 50% and ROC is 0.48 , while based on the report the precision for both classes is close to the 50% and the model recalls True outputs better, meaning we need Feature Extraction/Engineering in the preprocessing phase.

**Task 25 –**

Summary: Analyzing the Correlation Between Doctors and Rehospitalization Rates

This task analyzed the relationship between doctors and the rehospitalization rates of patients. The data included the number of patients treated by each doctor and the corresponding rehospitalization counts.

Pearson Correlation: A moderate positive correlation was found with a correlation coefficient of 0.496 and R-squared of 0.24, indicating that 24% of the rehospitalization variance is explained by the number of patients treated.

Neural Network: A neural network model achieved an R-squared of 0.1338, explaining around 13.4% of the rehospitalization variance. The analysis suggests a moderate linear relationship, with the neural network providing additional, though limited, predictive power.

**Task 22 –**

**Research Question**

Is there statistical relationship between DayOfWeek and Rehospitalization?

**Target variable**

* Type: discrete
* Possible target labels: "rehospitalized", "non-rehospitalized"

**Feature variable**

* Type: discrete
* Possible feature labels: "Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"

**Analysis output**

Conditions for statistical relationship test are not met, because of definitive bias:

Number of rehospitilied patients: 7033 VS number of non-rehospitilized patients: 0

We are unable to create "contingency table" that is a requirement for Chi-Squared or Fisher's Tests.

**Conclusion**

It is impossible to state presence of statistical relationship between day-of-week and rehospitalization.