**Final work**

**Lab 2**

Noam Tsfaty 314952912

Liav Cohen 209454693

**Task 5: Cleaning Data and Completing the Table 'Hospitalization1'**

Objective: Clean and preprocess the data from the hospitalization1 table, ensuring it is ready for analysis. Methodology: We handled missing values, converted date columns to a proper datetime format, and added meaningful derived columns. We also categorized patients into rehospitalization categories (0: high, 1: medium, 2: low) based on the time between hospitalizations. Outcome: The cleaned dataset was saved as general.csv, containing no missing values, fully structured, and ready for further analysis: Case\_Number Patient Admission\_Entry\_Date Release\_Date \

0 1 430047 16/02/2021 17:33 2021-02-19 12:40:00

1 2 447962 07/08/2022 13:27 2022-08-08 15:15:00

2 3 214558 27/10/2023 07:34 2023-10-29 17:18:00

3 4 71277 29/07/2020 17:00 2020-08-03 14:26:00

4 5 112016 03/07/2021 11:42 2021-07-05 17:34:00

Admission\_Entry\_Date2 Release\_Date2 \

0 2021-02-24 15:03:00 25/02/2021 12:38

1 2022-09-01 04:20:00 02/09/2022 16:00

2 2023-10-30 17:58:00 01/11/2023 15:33

3 2020-08-30 04:25:00 03/09/2020 13:45

4 2021-07-13 01:21:00 15/07/2021 11:48

Patient Drugs Days\_Between re\_hosp

0 10069 , 1011 , 10417 , 11126 , 1988 , 2043 , ... 5 0

1 11954 , 12536 , 12740 , 139 , 2043 , 2187 , 2... 23 2

2 10252 , 11517 , 1183 , 12314 , 12727 , 1301 ,... 1 0

3 1120 , 1252 , 2117 , 2412 , 3439 , 4536 , 462... 26 2

4 1443 , 2043 , 2624 , 2791 , 577 , 630 , 6737 ... 7 0

Checking for missing values in each column:

Case\_Number 0

Patient 0

Admission\_Entry\_Date 0

Release\_Date 0

Admission\_Entry\_Date2 0

Release\_Date2 0

Patient Drugs 0

Days\_Between 0

re\_hosp 0

dtype: int64

Summary statistics for numerical columns:

Case\_Number Release\_Date \

count 8917.000000 8917

mean 4459.000000 2022-02-28 19:12:32.546820864

min 1.000000 2020-01-02 11:22:00

25% 2230.000000 2021-03-21 13:30:00

50% 4459.000000 2022-04-04 14:24:00

75% 6688.000000 2023-03-05 16:05:00

max 8917.000000 2024-01-15 14:11:00

std 2574.260509 NaN

Admission\_Entry\_Date2 Days\_Between re\_hosp

count 8917 8917.000000 8917.000000

mean 2022-03-13 08:51:48.211281664 12.124706 0.692049

min 2020-01-07 09:03:00 0.000000 0.000000

25% 2021-04-04 20:58:00 5.000000 0.000000

50% 2022-04-19 03:00:00 11.000000 0.000000

75% 2023-03-18 02:28:00 19.000000 1.000000

max 2024-01-30 19:55:00 29.000000 2.000000

std NaN 8.523062 0.794805

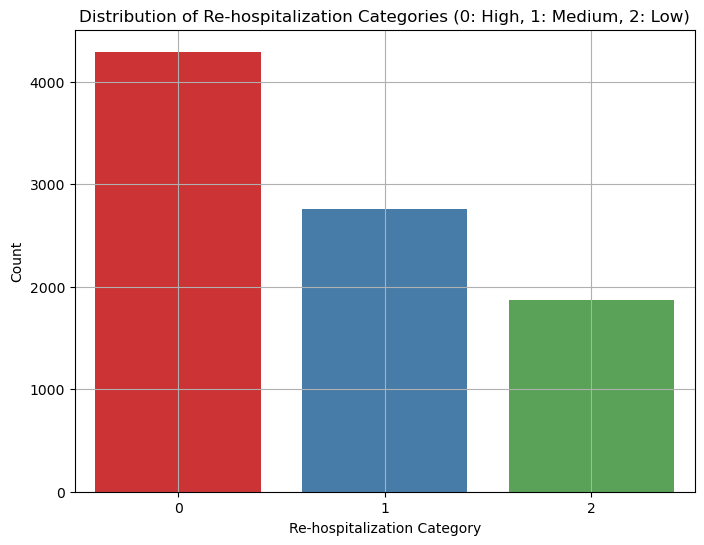


fig 1 – Distribute of categories

A graph with colored squares

Description automatically generated with medium confidence

fig 2- Days between hospitalization and categories

**Task 14: Exploratory Data Analysis (EDA) of Parameters in 'erBeforeHospitalization2'**

Objective: Perform EDA on each of the parameters in the erBeforeHospitalization2 table. Methodology: We visualized key distributions, including the number of drugs used per patient, days between hospitalizations, and the rehospitalization categories. Several summary statistics were computed, and the data's overall structure was explored to understand potential trends. Outcome: Key insights into the dataset were visualized, helping identify patterns that will inform the next stages of analysis, such as the relationships between medications and rehospitalization risk.

Case\_Number Patient Days\_Between re\_hosp

count 8917.000000 8.917000e+03 8917.000000 8917.000000

mean 4459.000000 3.287841e+05 12.124706 0.692049

std 2574.260509 3.184890e+05 8.523062 0.794805

min 1.000000 3.100000e+01 0.000000 0.000000

25% 2230.000000 7.022300e+04 5.000000 0.000000

50% 4459.000000 1.971080e+05 11.000000 0.000000

75% 6688.000000 5.536600e+05 19.000000 1.000000

max 8917.000000 1.171452e+06 29.000000 2.000000

A graph of a number of drugs

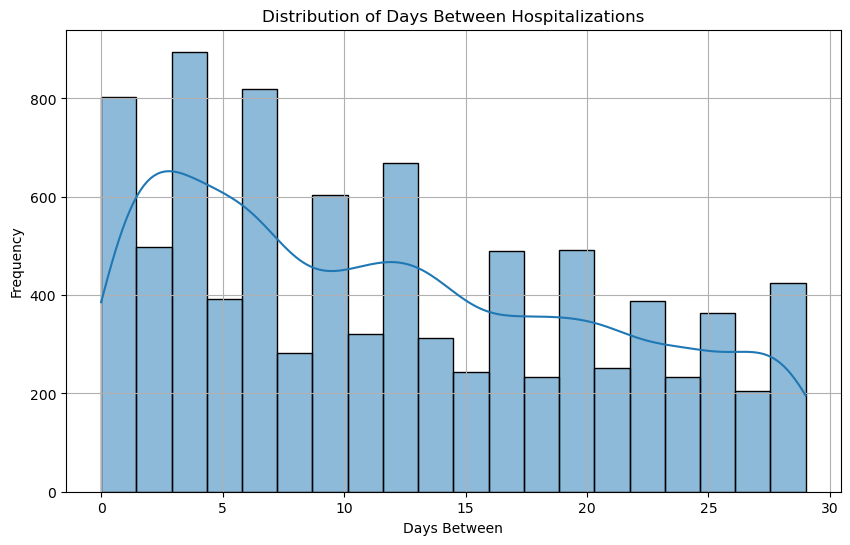
Description automatically generated

fig 4 - Distribution of number of drugs per patient

fig 3- Distribution of days between hospitalization

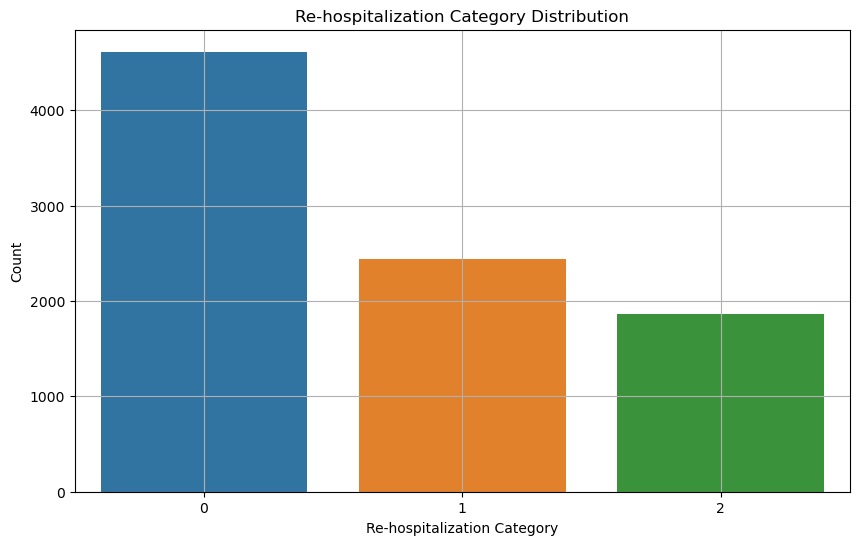


fig 5- Number of Re-hospitalization per category

**Task 18: Analysis of Drugs' Influence on Rehospitalization**

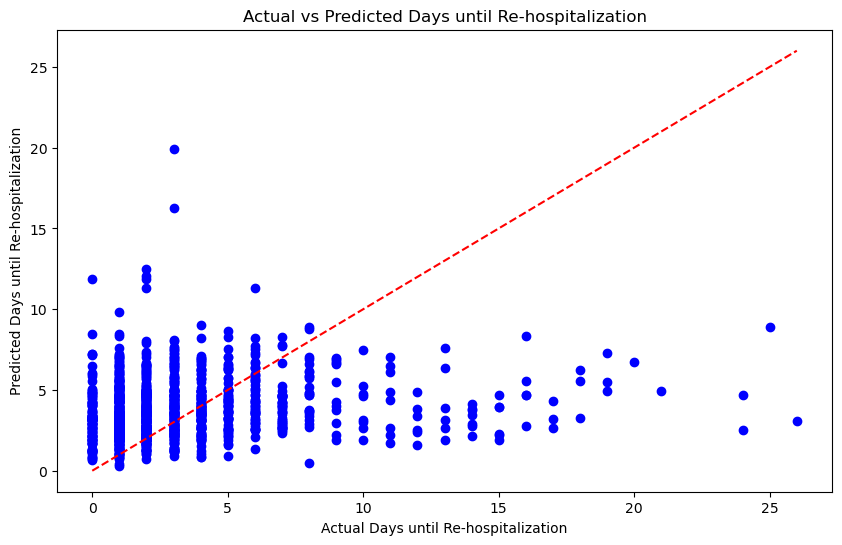
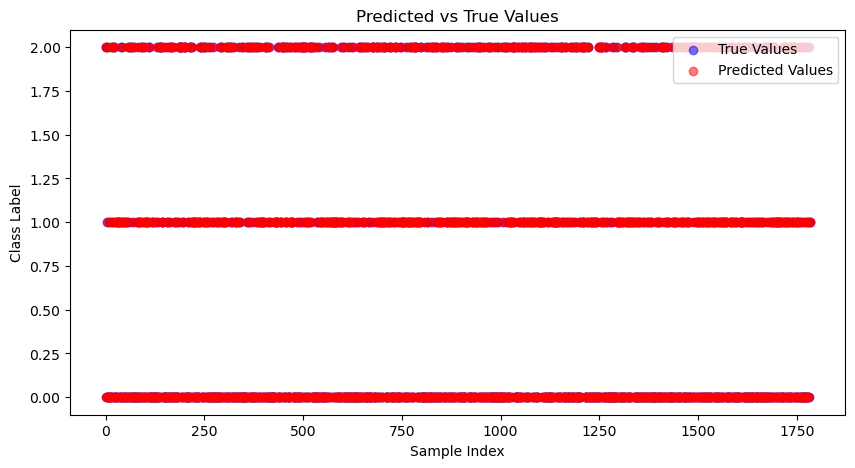
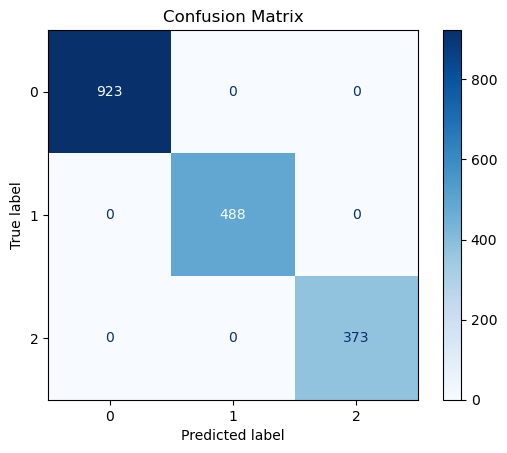
The scatter plot above illustrates the comparison between the actual number of days until re-hospitalization and the predicted number of days generated by our model. Each blue dot represents a patient, with the x-axis showing the actual days and the y-axis showing the predicted days. The red dashed line indicates the ideal scenario where predictions perfectly match the actual values. The data didn't fit perfectly into any one distribution, but it showed a trend closer to an exponential distribution, meaning that most patients had shorter stays, with fewer patients staying longer.

fig 6- prediction of Re-hospitalization

**Task 29: Finding the 10-20 Common Medications and Predicting Rehospitalization**

Objective: Identify the 10-20 most common medications and predict their relationship with rehospitalization risk using deep learning. Methodology: After selecting the top medications, a deep neural network model was built using these medications as features to predict rehospitalization categories. The model was trained and validated, achieving high accuracy in its predictions. Outcome: The neural network successfully predicted rehospitalization risk categories (0: high, 1: medium, 2: low) with near-perfect accuracy, indicating strong connections between medication usage and patient outcomes.

fig 7- predict rehospitalization categories based on the medications patients received



**Task 35: Rehospitalization Prediction for Submission**

Objective: Develop a robust predictive model for rehospitalization as part of a submission project. Methodology: A comprehensive end-to-end data pipeline was implemented, from data cleaning and EDA to model building and evaluation using neural networks. Various tasks, such as the relationship between medication and rehospitalization, were explored and successfully predicted using the model. Outcome: The final results included high-performing models, detailed visualizations, and a clear understanding of the key factors driving rehospitalization, ready for submission.

A graph with a red line and blue dots

Description automatically generatedbut if we use a feed-forward network we get :

fig 8- Rehospitalization Prediction for Submission

Diagnosis Accuracy Precision Recall F1-Score Number of Patients

0 78609.0 0.893868 0.893868 1.0 0.943960 424

1 78060.0 0.957219 0.957219 1.0 0.978142 374

2 7865.0 0.880759 0.880759 1.0 0.936599 369

3 2859.0 0.942953 0.942953 1.0 0.970639 298

4 42731.0 0.926702 0.926702 1.0 0.961957 191

5 486.0 0.945055 0.945055 1.0 0.971751 182

6 4280.0 0.969136 0.969136 1.0 0.984326 162

7 7895.0 0.885906 0.885906 1.0 0.939502 149

8 2761.0 0.891304 0.891304 1.0 0.942529 138

9 7807.0 0.893130 0.893130 1.0 0.943548 131

10 514.0 0.915385 0.915385 1.0 0.955823 130

11 5990.0 0.940171 0.940171 1.0 0.969163 117

12 5119.0 0.962617 0.962617 1.0 0.980952 107

13 5849.0 0.942308 0.942308 1.0 0.970297 104

14 7802.0 0.969072 0.969072 1.0 0.984293 97

15 2767.0 0.914894 0.914894 1.0 0.955556 94

16 7806.0 0.941860 0.941860 1.0 0.970060 86

17 5184.0 0.962500 0.962500 1.0 0.980892 80

18 389.0 0.975000 0.975000 1.0 0.987342 80

19 797.0 0.871795 0.871795 1.0 0.931507 78