**Task 9: EDA for all parameters in table Hospitalization 1**

Description of the data:

The table has 7033 rows with 12 columns.

1. Patient - contains a unique numeric id code for the patient. A patient code can repeat several times, each time for a different hospitalization period.
2. Department id - Contains the hospitalization department number from 1 to 5.
3. Admission Medical Record - a unique numeric id for the hospitalization medical record.
4. Admission entry date - date and time of admission to the department.
5. Release date - date and time of release from the department.
6. Reception type - 3 categories:
   1. Urgent
   2. Invited (planned hospitalization)
   3. Day hospitalization
7. Patient origin - 5 categories:
   1. From home
   2. From institute
   3. Other
   4. From clinic
8. Release Type - 2 categories:
   1. Released home
   2. Released to institute
9. Release doctor code - the id code of the releasing doctor.
10. Hospitalization duration - the time the patient was hospitalized in the department in days (Release date minus admission entry date).
11. Admission diagnoses - ICD9 diagnosis code or codes that were given at time of admission (numeric or combination of numbers and letters or characters like dots).
12. Release diagnoses - ICD9 diagnosis code or codes that were given at time of release (numeric or combination of numbers and letters or characters like dots).

Methodology:

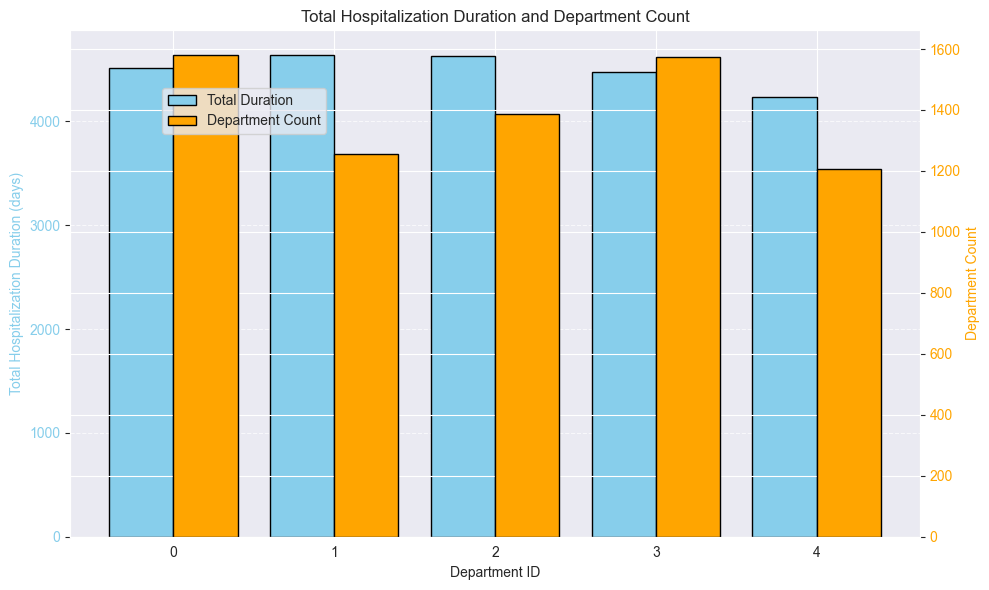
1. Importing the data from the excel file to a pandas data frame (df).
2. Translating the Hebrew column names to English.
3. Checking for missing values (nulls).

| Column | null count |
| --- | --- |
| Reception type | 64 |
| Release doctor code | 114 |
| Admission diagnoses | 461 |
| Release diagnoses | 29 |

1. Removing rows that have 3 or 4 null values. 27 rows removed.
2. Replaced the nulls in the ‘release doctor code’ column with the most common release doctor code in the department at the same day of release. If no other release doctor code exists on the same day, it takes the most common code on the days before and after, increasing the number of days until all nulls are filled.
3. Replacing the nulls in the columns ‘admission diagnoses’ and ‘release diagnoses’ with the value ‘Unknown’
4. Replacing the nulls in the ‘reception type’ column with the most frequent value ‘Urgent’ (90.2% of the column values). No missing values remain in the df.
5. Checking that there are no repeated values in the column ‘admission medical record’.
6. Translating the Hebrew categorical values to English.
7. Looking for Hebrew characters in other possible string containing columns.
8. Converting the admission diagnoses and release diagnoses column values from comma separated to lists.
9. Changing the data type of the categorical features from integer/real to category.
10. Looking for non-logical statistical values in the data, like logical min and max values in the numerical columns and the same value count in all columns. No issues were detected at this point. Before the null removal we Identified negative values in the hospitalization duration column due to switching of admission and release dates. But, these rows were removed during the null removal process.
11. Feature engineering
    1. Creating a new column named 'Admission Number' to classify the hospitalization number of each patient according to his admission date (earliest admission date of a patient = 1st hospitalization = Admission Number = 1, etc.)
    2. Creating date only columns from the ‘Admission entry date’ and ‘Release date’ date time columns.
12. Visualizations of the data and analyzing it.
13. Exporting the cleaned data frame to a CSV file.

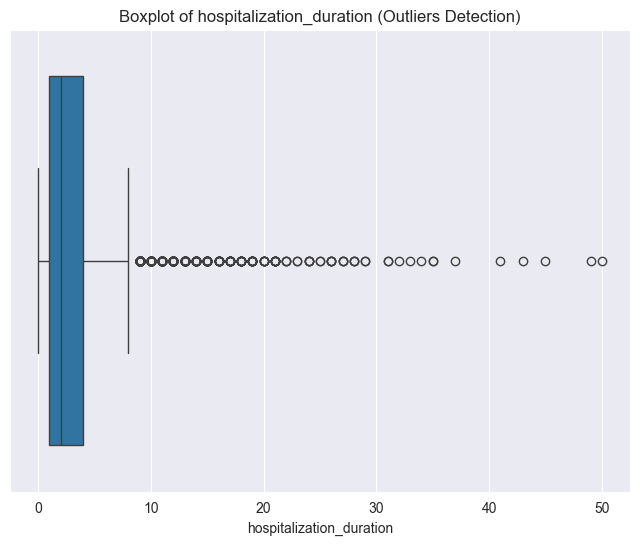
**Results:**

**Figure 1:** Total hospitalization duration and department count



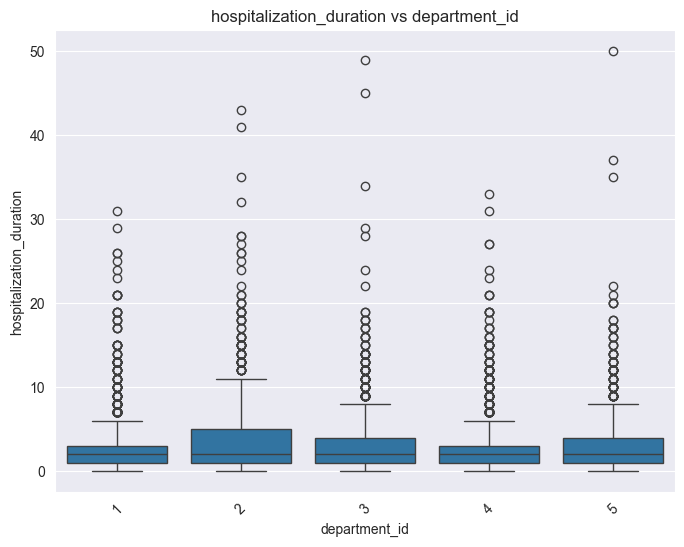
We can see that although there are some differences in department count (number of hospitalizations\patients per department) the differences in total hospitalization duration between the departments is smaller.

**Figure 2:** Boxplot of general hospitalization duration



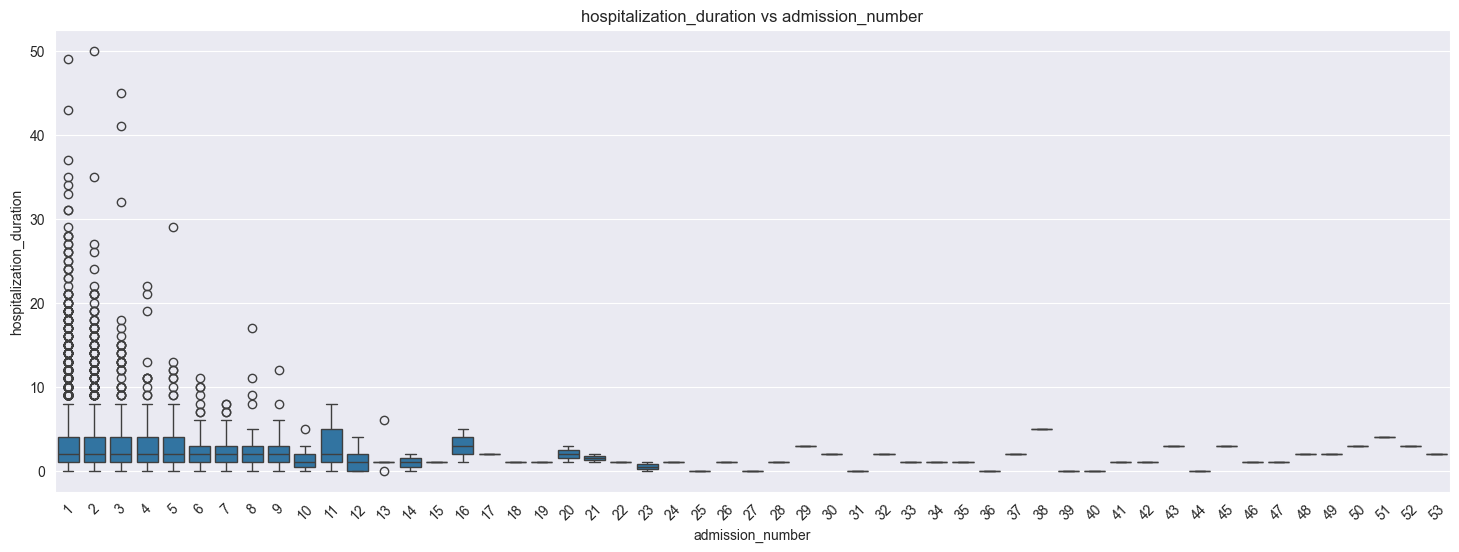
When box-plotting the general hospitalization duration we can see that the majority of hospitalizations are short.

**Figure 3**: Hospitalization duration per department



When box-plotting the hospitalization duration against the department id, we see that the mean hospitalization duration for all departments is about the same but department 2 has a higher percentage of long hospitalization durations.

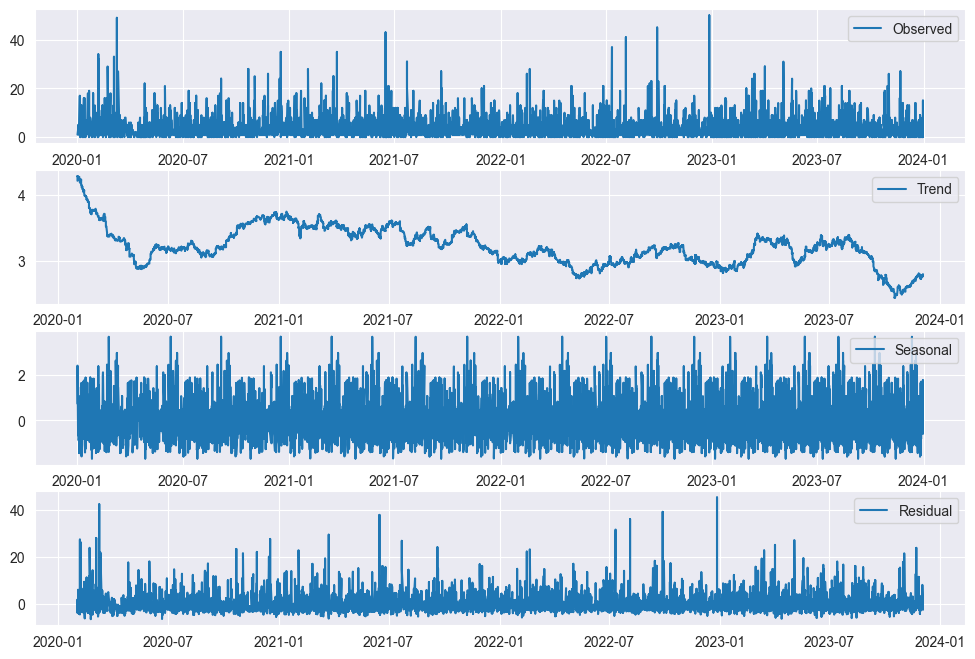
**Figure 4**: Hospitalization duration and Admission number



When box-plotting hospitalization duration with categorical splitting of admission number, we see that up to the 5th hospitalization (admission number) the box plots are the same (with the exception of the outliers). From the 6th hospitalization, as the number of hospitalizations per category gets lower, we see a more sporadic behavior.

The general trend is that as the admission number gets higher the hospitalization duration is shorter.

**Figure 5:** Seasonality and cyclical patterns analysis



When performing seasonal decomposition analysis on a seasonal period of 1 year we get a seasonal repetitive pattern that begins by repeating about every 3 months but at the end of the analysis period decreases to about every 2 months. When looking at the trend plot there seems to be a general slight decrease over time.

**Conclusions:**

The EDA was performed successfully.

The resulting data frame consists of 7006 rows (hospitalizations) out of the original 7033 rows. The row removal was due to 3 or 4 missing values in each row.

The remaining missing data was logically filled (see methodology section).

All Hebrew strings were translated to English.

New features were created for admission number (the patient hospitalization \ rehospitalization number) and admission / release date with the date only (without the HH:MM:SS data).

Although there are small differences (less than 12%) in department counts (number of hospitalizations \ patients per department) the differences in total hospitalization duration between the departments is smaller (less than 9%). Meaning the workload between the departments is relatively even.

While the mean hospitalization duration for all departments is about the same, department 2 has a higher percentage of long hospitalization durations.

Up to the 5th hospitalization (admission number) the hospitalization duration distributions are the same (with the exception of the outliers). From the 6th hospitalization, as the number of hospitalizations per category gets lower, we see a more sporadic behavior. The general trend is that as the admission number gets higher the hospitalization duration is shorter.

Seasonal decomposition analysis on a seasonal period of 1 year results in a seasonal repetitive pattern of about 3 months but at the end of the analysis period decreases to about 2 months. When looking at the trend plot there seems to be a general slight decrease over time.

**Task 24 - Correlation between department occupancy and rehospitalization**

Description of the data:

Units occupancy rate contains 6736 rows with 4 columns.

1. Date - from 01-Jan-2021 to 31-Dec-2023. Each date appears 5 times, one per department id.
2. Department id - department 1 to 5.
3. Occupancy count - the number of beds occupied in the department.
4. Occupancy rate - occupancy count divided to 40 (the number of beds per department).

Methodology:

1. Importing the Units occupancy rate data from the excel file to a pandas data frame (df).
2. Translating the Hebrew column names to English.
3. Checking for null values. No null values found.
4. Feature engineering:
   1. Creating a feature for hospitalization count that specifies in each row the total hospitalizations a patient had.
   2. Creating a feature for the previous release date.
   3. Creating a feature for duration between hospitalizations - previous release date minus admission date.
   4. Creating a feature for duration classification that classifies the duration between hospitalizations to short, medium or long by tertiaries.

Short - 0 to 8 days (0 to 33%)

Medium - 9 to 24 days (34% to 66%)

Long - 25 days and more (67% to 100%)

1. Creating a processed data frame based on hospitalization1 with new engineered features, containing only rehospitalizations (not including the rows of the 1st hospitalization). Total rows - 2503.
2. Merging to a combined data frame, the df created in section 5 and the Units occupancy rate df by the combined values of department id and admission date (date in the Units occupancy rate df). So that each row that has a certain admission date and a certain department gets the values of occupancy count and occupancy rate of the same date and department.
3. Correlation analysis between occupancy rate and hospitalization duration per department.
4. Training a classification model.
   1. Target feature - Duration classification. Encoded using Label encoder.
   2. Training features - Occupancy rate, Hospitalization duration. Standardized using Standard scalar.
   3. splitting data to 70% train 30% test.
   4. Model - NeuralNetworkClassifier - Torch nn
      1. Hidden size = 10
      2. Learning rate = 0.01
      3. Number of epochs = 5000
5. Evaluating the model classification.
6. Training the model with the feature occupancy rate only (same parameters).

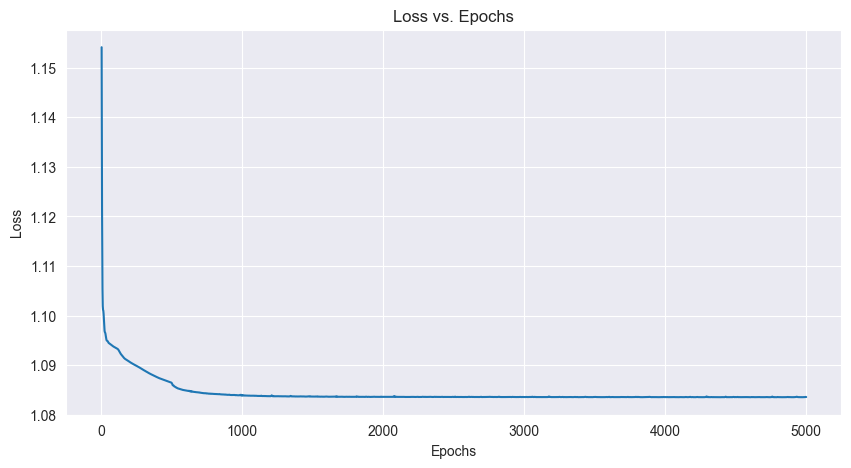
**Results:**

**Table1:**

| Department | Correlation between occupancy rate and hospitalization duration |
| --- | --- |
| 1 | 0.01919 |
| 2 | -0.03087 |
| 3 | -0.00029 |
| 4 | -0.03588 |
| 5 | -0.00948 |

We can see that there is no correlation between the department occupancy rate and the patient hospitalization duration.

**Figure 1:** Training loss vs epochs

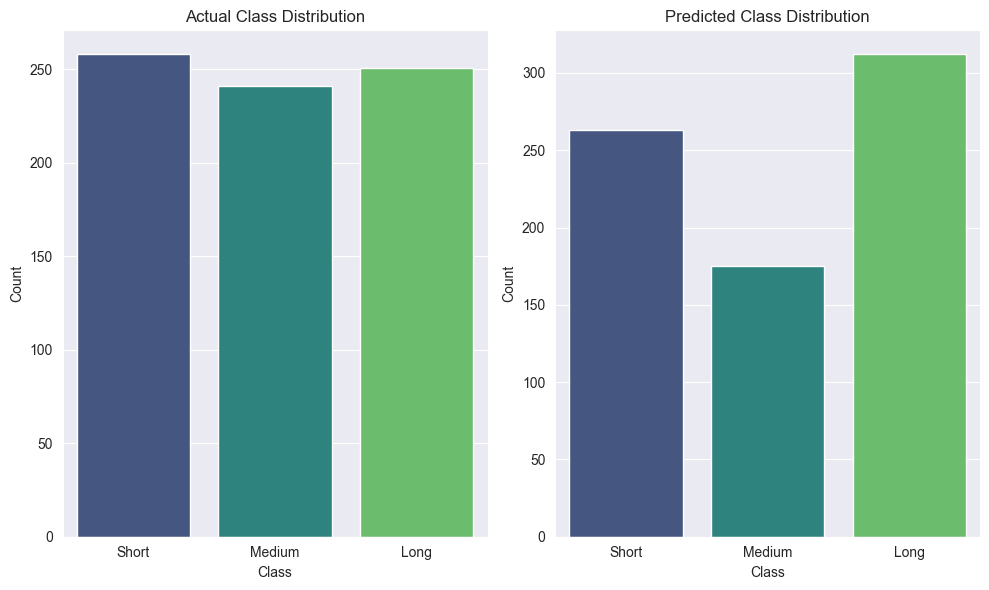
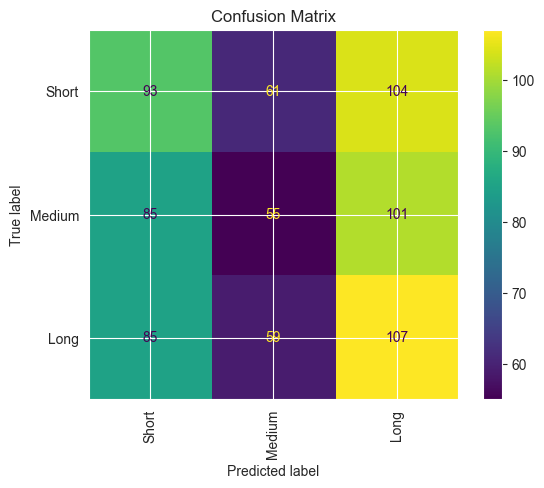


We can see a decrease in the loss up to 1000 epochs. After 1000 epochs the loss remains constant at 1.0835.

**Table 2:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 | 0.35 | 0.36 | 0.36 | 258 |
| 1 | 0.31 | 0.23 | 0.26 | 241 |
| 2 | 0.34 | 0.43 | 0.38 | 251 |
| accuracy |  |  | 0.34 | 750 |
| macro avg | 0.34 | 0.34 | 0.33 | 750 |
| weighted avg | 0.34 | 0.34 | 0.33 | 750 |

We can see that all scores are very low and are very similar between the classes (with the medium class being a bit lower), meaning the model prediction is very low. There is no correlation between the occupancy rate and hospitalization duration to the duration between hospitalizations.

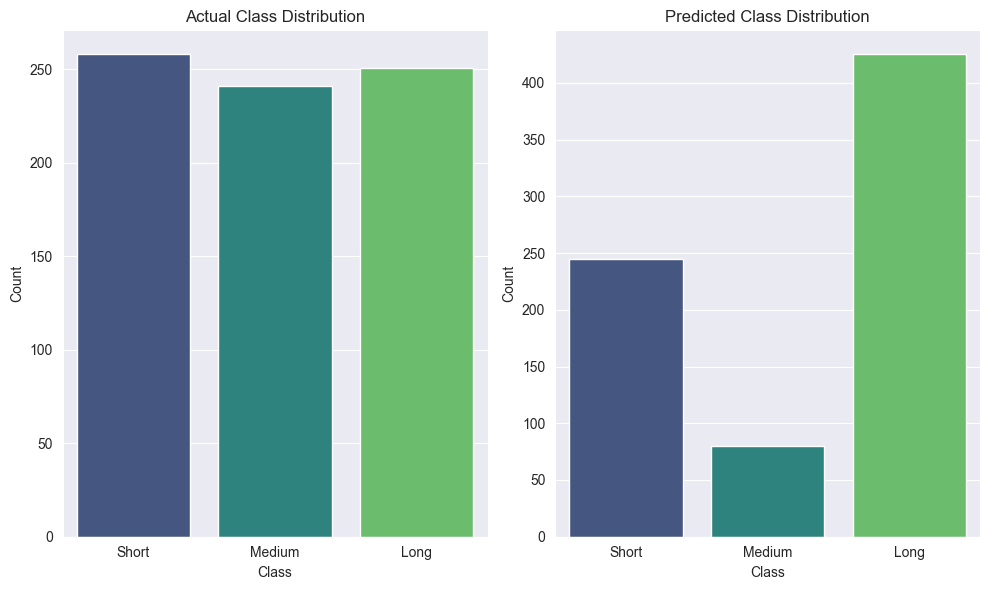
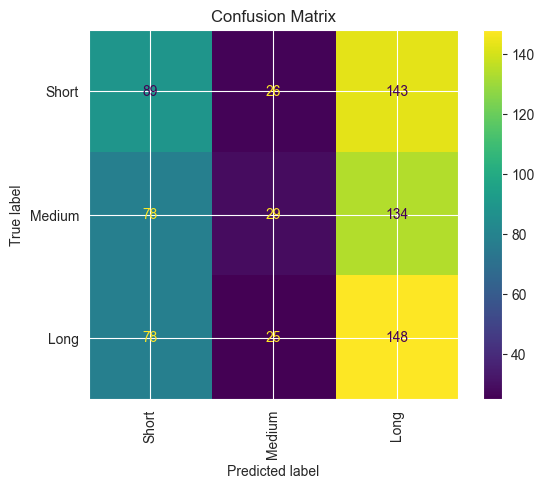
**Figure 2:** Actual class distribution vs predicted class distribution and Confusion matrix

We can see that the predicted classifications for the short and long durations between hospitalizations is much more accurate than the one for the medium durations. But as we learn from the model evaluation scores, these relatively good results do not come from an accurate model prediction but from the even distribution between target classes. Moreover, from the confusion matrix we can see that from a total of 750 samples in the test set only 255 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 34.0%.

**Table 3:** Model evaluation scores on the test set for a model with the occupancy rate only.

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 | 0.36 | 0.34 | 0.35 | 258 |
| 1 | 0.36 | 0.12 | 0.18 | 241 |
| 2 | 0.35 | 0.59 | 0.44 | 251 |
| accuracy |  |  | 0.35 | 750 |
| macro avg | 0.36 | 0.35 | 0.32 | 750 |
| weighted avg | 0.36 | 0.35 | 0.33 | 750 |

We can see that all scores are still very low and that the recall and f1-score for the medium class is even lower than the previous model. There is no correlation between the occupancy rate to the duration between hospitalizations.

**Figure 3:** Actual class distribution vs predicted class distribution and Confusion matrix for model with the occupancy rate only

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are not accurate, especially the medium and long classes. From the confusion matrix we can see that from a total of 750 samples in the test set only 266 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 35.5%.

**Conclusions:**

There is no correlation between the department occupancy rate and the patient hospitalization duration.

The classification model did not find a correlation between the occupancy rate and hospitalization duration to the duration between hospitalizations.

The task was to find the relationship between the department occupancy and rehospitalization. We chose the duration between hospitalizations as a representative measurement to rehospitalization. The question is, if there is no better feature to represent rehospitalizations.

**Task 26 - Finding a connection between the age and sex of the patient to:**

1. **1. time between hospitalizations**

**2. time between 1st and 2nd hospitalization (task 16)**

1. **number of hospitalization days in first hospitalization (task 17)**
2. **number of hospitalization days in 2nd hospitalization (task 18)**
3. **number of hospitalization days in 2nd and above hospitalizations (hospitalization duration in rehospitalizations)**

**EDA of General data table**

General data table - Description of the data:

The table has 4535 rows with 12 columns.

1. Patient - contains a unique numeric id code for the patient. A patient code can repeat several times, each time for a different hospitalization period.
2. Age - patient age in years.
3. Gender - patient gender (Male / Female).
4. Income source - the source that covers the hospitalization expenses (HMO / self funding / Prison authority / Tourist).
5. Weight - patient weight in Kg.
6. Height - patient height in cm.
7. BMI - patient body mass index (weight in kilograms (kg) divided by height in meters (m) squared.
8. Chronological illnesses - if the patient has a chronological illnesses (1) or not (0).
9. Education - patient education
10. Number of children - patient number of children
11. Marital status - patient marital status
12. Medications - patient list of regular medications

**Methodology:**

1. Importing the data from the excel file to a pandas data frame (df).
2. Translating the Hebrew column names to English.
3. Checking for missing values (nulls).

| Column | null count |
| --- | --- |
| Income source | 11 |
| Weight | 817 |
| Height | 837 |
| BMI | 1031 |
| Education | 1894 |
| Number of children | 162 |
| Marital status | 61 |
| Medications | 40 |

1. Fixing errors and filling nulls in the height and weight columns.
2. Some heights were entered in meters instead of centimeters. These heights were identified by being between 1 and 2.5 and the value was multiplied by 100.
3. Some heights were entered without the hundred digit (eg. 80 instead of 180). This was identified by the height value being between 40 and 95 and the weight value being between 30 and 120. 100 was added to the height value.
4. Fixing switched values of height and weight. For some patients height values were entered as weight and weight as height. This was identified by the height value being smaller than the weight value and the weight value being between 120 to 190. The weight and height values were switched.
5. Replacing null values of height and weight with the mean value of the same age and gender. If there is no valid value for the same age and gender, the mean value of the age minus i and gender is taken. i starts at 1 and increases by 1 until a valid value is found.
6. Fixing errors and filling nulls in the BMI column.
7. Some BMI values are very high, hence are not logical. In these cases, if the BMI value is over 55 the BMI value is replaced by methods 17c or 17d.
8. Null BMI values are replaced by methods 17c or 17d.
9. If height and weight values are not null, the BMI is calculated from their values (weight / (height/100)2.
10. If height or weight is null, the mean BMI value of the same age and gender is taken. If there is no valid value for the same age and gender, the mean value of the age minus i and gender is taken. i starts at 1 and increases by 1 until a valid value is found.
11. Replacing null values of features ‘education’, ‘income source’, ‘number of children’ and ‘marital status’ by their mean values for the same age and gender. If there is no valid value for the same age and gender, the mean value of the age, the age minus i, the age plus i and gender is taken. i starts at 0 and increases by 1 until a valid value is found.
12. Converting ‘medications’ column from comma separated strings to lists and replacing null values by the 3 most common medication codes from the same gender, age and age±i (i start at 1 and increase by 1 until a value is found).
13. Checking for missing values. No missing values found.
14. Feature engineering (same as task 24):
15. Creating a feature for hospitalization count that specifies in each row the total hospitalizations a patient had.
16. Creating a feature for the previous release date.
17. Creating a feature for duration between hospitalizations - previous release date minus admission date.
18. Creating a feature for duration classification that classifies the duration between hospitalizations to short, medium or long by tertiaries.

Short - 0 to 8 days (0 to 33%)

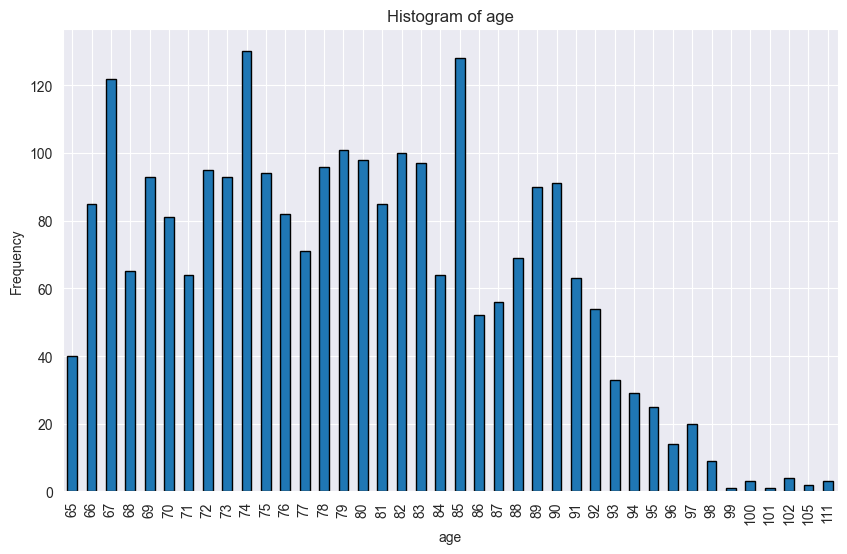
Medium - 9 to 24 days (34% to 66%)

Long - 25 days and more (67% to 100%)

1. Creating a processed data frame based on hospitalization1 with new engineered features, containing only rehospitalizations (not including the rows of the 1st hospitalization). Total rows - 2503.
2. Merging to a combined data frame (df\_combined\_26), from the df created in section 10 and the clean general data df by the combined values of patient id.
3. Age feature analysis by plotting the feature description and a histogram of age frequencies. Minimum age is 65 and maximal age is 111.
4. Feature engineering of feature age binned from the age feature in order to minimize the age feature and ease the classification. A total of 7 age bins were created with the following age groups: 65-69, 70-74, 75-79, 80-84, 85-8', 90-9', 95+. This classification gives an even frequency of the 5 first age bins while retaining a variety in age groups for finding better classification relationships by the model.

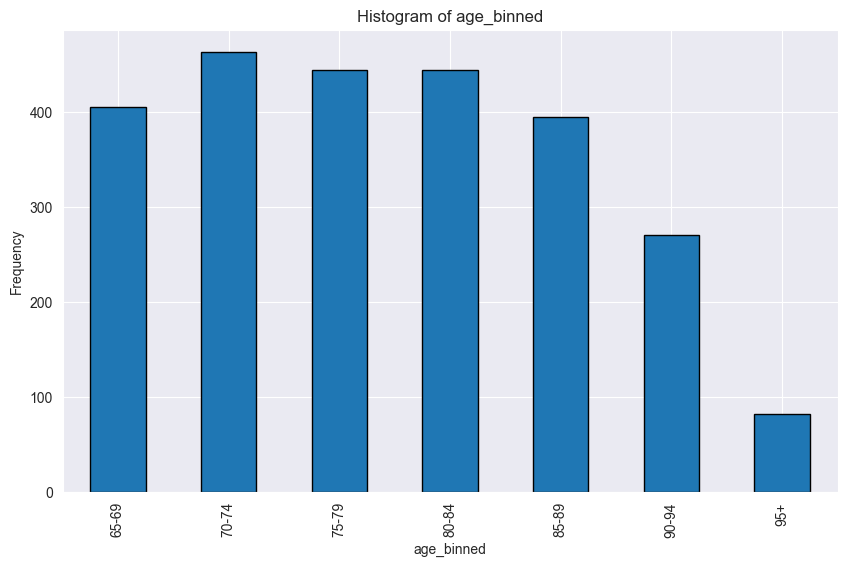
**Results:**

**Figure 1:** Histogram of age frequencies

****

We can see that the minimum age is 65 and the maximum age is 111 and that up to the age of ~90 the age frequencies vary around the same frequencies. Above the age 92 we see a decrease of frequency with the rise of age.

**Figure 2:** Histogram of age binned frequencies



We can see that this age binned classification gives an even frequency for the 5 first age bins while retaining a variety in age groups for finding better classification relationships by the model.

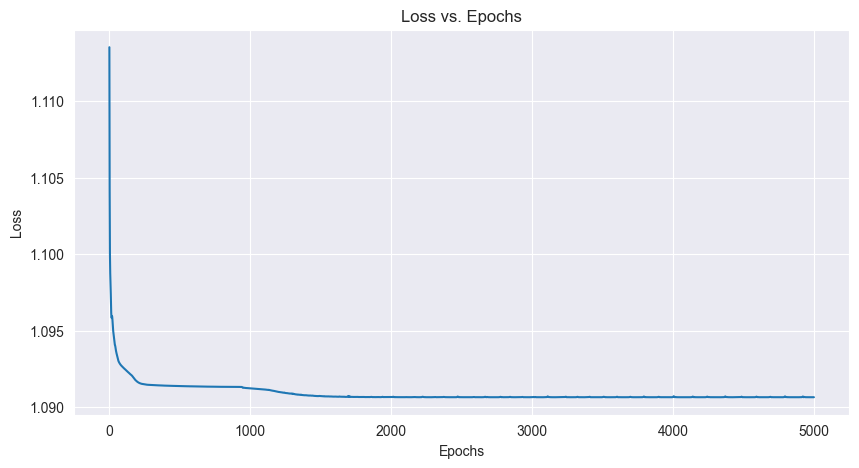
**Task 26. a. 1. - Finding a connection between the age and sex of the patient to time between hospitalizations**

**Methodology:**

1. Defining the features age binned and gender as the training features.
2. Defining the feature duration classification (duration between hospitalizations) as target.
3. Training a classification model.
   1. Target feature - Duration classification. Encoded using Label encoder.
   2. Training features - age binned and gender. Encoded using Label encoder.
   3. Splitting data to 70% train 30% test.
   4. Model - NeuralNetworkClassifier - Torch nn
      1. Hidden size = 10
      2. Learning rate = 0.01
      3. Number of epochs = 5000
4. Evaluating the model classification.

**Results:**

**Figure 1:** Training loss vs epochs

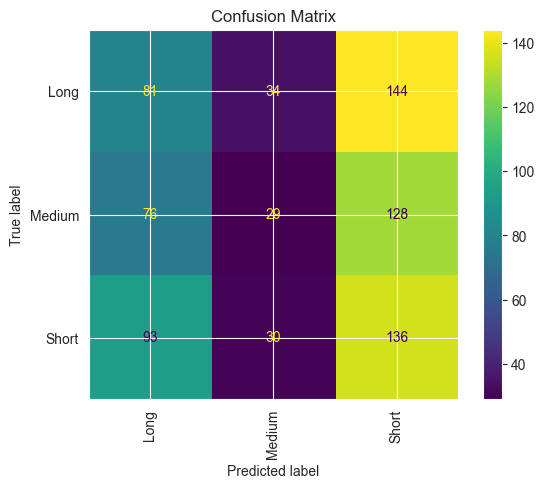
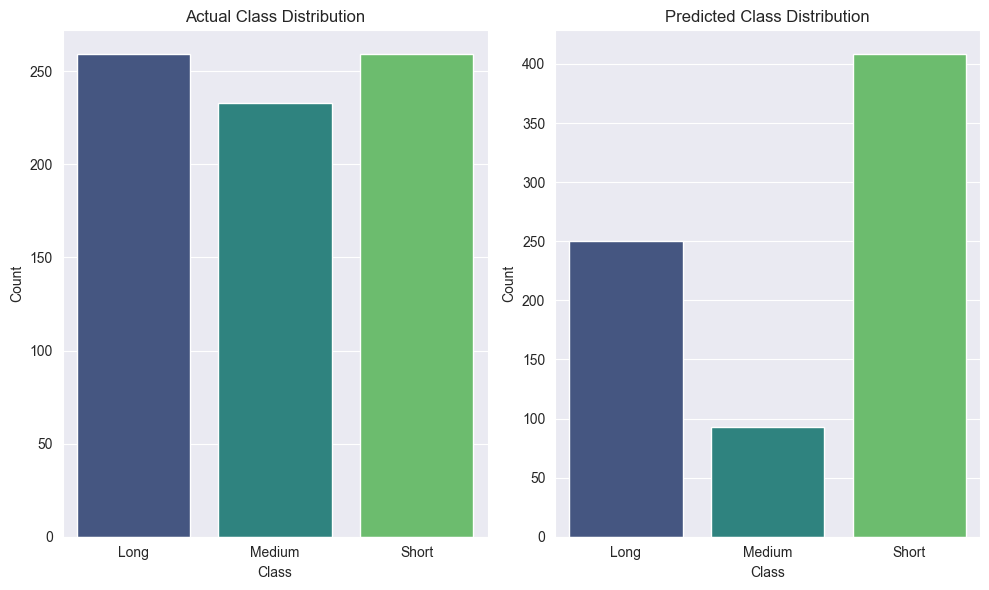


We can see a decrease in the loss up to 2000 epochs then remaining constant at 1.0906.

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.32 | 0.31 | 0.32 | 259 |
| 1 (Medium) | 0.31 | 0.12 | 0.18 | 233 |
| 2 (Short) | 0.33 | 0.53 | 0.41 | 259 |
| accuracy |  |  | 0.33 | 751 |
| macro avg | 0.32 | 0.32 | 0.30 | 751 |
| weighted avg | 0.32 | 0.33 | 0.31 | 751 |

We can see that all scores are very low and that the recall and f1-score for the medium class is even lower. There is no correlation between age and gender to the duration between hospitalizations.

**Figure 2:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are not accurate, especially the medium and long classes. From the confusion matrix we can see that from a total of 751 samples in the test set only 246 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 32.8%.

**Conclusions:**

There is no correlation between the age and gender of the patient and the duration between hospitalizations.

**Task 26. a. 2. - Finding a connection between the age and sex of the patient to the time between 1st and 2nd hospitalization**

**Methodology:**

1. Creating a subset df from the ‘df\_combined\_26’ df that contains only the 2nd hospitalizations.
2. Creating a feature named ‘1st\_to\_2nd\_dur\_class’ for classifying the durations between hospitalization 1 and hospitalization 2 by tertiaries of the ‘duration between hospitalizations’ feature.

Short - 0 to 8 days (0 to 33%) - 469 values

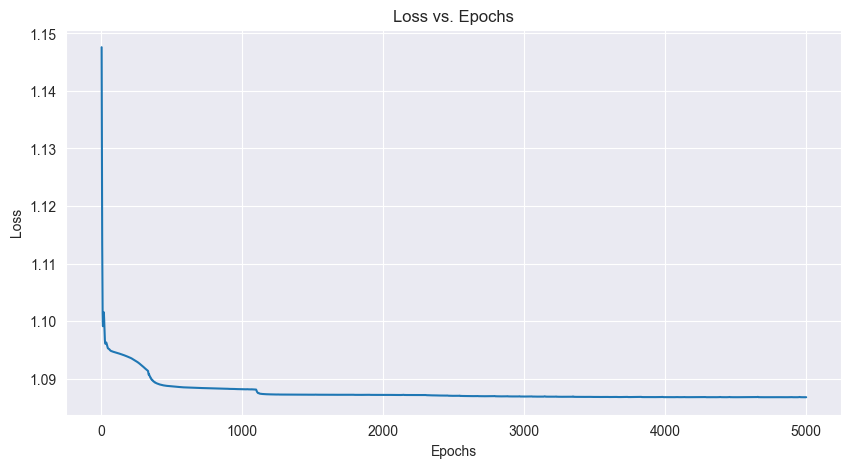
Medium - 9 to 27 days (34% to 66%) - 448 values

Long - 28 days and more (67% to 100%) - 425 values

1. Training a classification model.
   1. Target feature - 1st\_to\_2nd\_dur\_class. Encoded using Label encoder.
   2. Training features - age binned and gender. Encoded using Label encoder.
   3. Splitting data to 70% train 30% test.
   4. Model - NeuralNetworkClassifier - Torch nn
      1. Hidden size = 10
      2. Learning rate = 0.01
      3. Number of epochs = 5000
2. Evaluating the model classification.

**Results:**

**Figure 1:** Training loss vs epochs

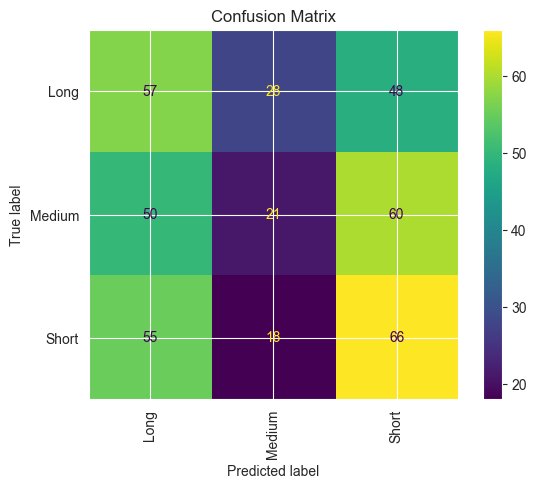


We can see a decrease in the loss up to 3200 epochs then remaining constant at 1.0868.

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.35 | 0.43 | 0.39 | 133 |
| 1 (Medium) | 0.31 | 0.16 | 0.21 | 131 |
| 2 (Short) | 0.38 | 0.47 | 0.42 | 139 |
| accuracy |  |  | 0.35 | 403 |
| macro avg | 0.35 | 0.35 | 0.34 | 403 |
| weighted avg | 0.35 | 0.36 | 0.34 | 403 |

We can see that all scores are very low. There is no correlation between age and gender to the duration between 1st and 2nd hospitalization.

**Figure 2:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are not accurate. From the confusion matrix we can see that from a total of 403 samples in the test set only 144 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 35.7%.

**Conclusions:**

There is no correlation between the age and gender of the patient and the duration between the 1st and 2nd hospitalizations.

**Task 26. b. - Finding a connection between the age and sex of the patient to the number of hospitalization days in first hospitalization (task 17)**

**Methodology:**

1. Merging to a combined data frame (df\_combined\_26\_b), from the cleaned hospitalization1 df and the clean general data df by the combined values of patient id.
2. Creating a subset df from the ‘df\_combined\_26\_b’ df that contains only the 1st hospitalizations.
3. Feature engineering of feature age binned from the age feature in order to minimize the age feature and ease the classification. A total of 7 age bins were created with the following age groups: 65-69, 70-74, 75-79, 80-84, 85-8, 90-94, 95+. This classification gives an even frequency of the 5 first age bins while retaining a variety in age groups for finding better classification relationships by the model.
4. Creating a feature named ‘1st\_hosp\_dur\_class’ for classifying the hospitalization durations of the 1st hospitalizations by tertiaries of the ‘hospitalization durations’ feature.

Short - 0 to 1 days (0 to 33%) - 1649 values

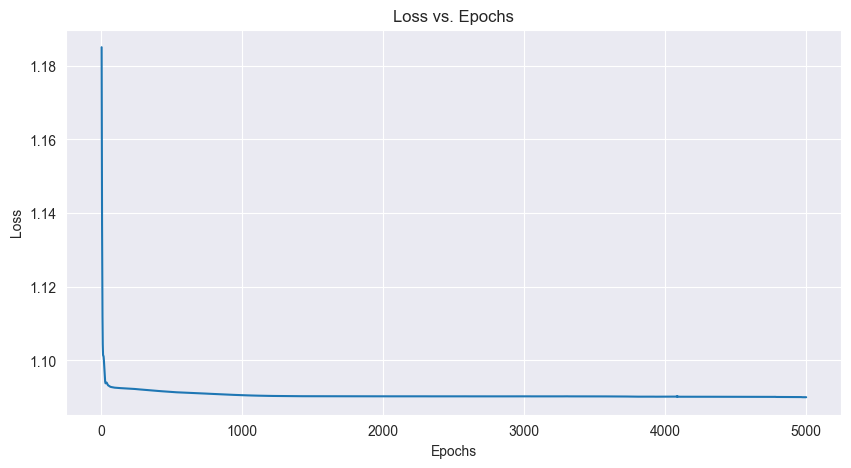
Medium - 2 to 3 days (34% to 66%) - 1503 values

Long - 4 days and more (67% to 100%) - 1354 values

1. Training a classification model.
   1. Target feature - 1st\_to\_2nd\_dur\_class. Encoded using Label encoder.
   2. Training features - age binned and gender. Encoded using Label encoder.
   3. Splitting data to 70% train 30% test.
   4. Model - NeuralNetworkClassifier - Torch nn
      1. Hidden size = 10
      2. Learning rate = 0.01
      3. Number of epochs = 5000
2. Evaluating the model classification.

**Results:**

**Figure 1:** Training loss vs epochs

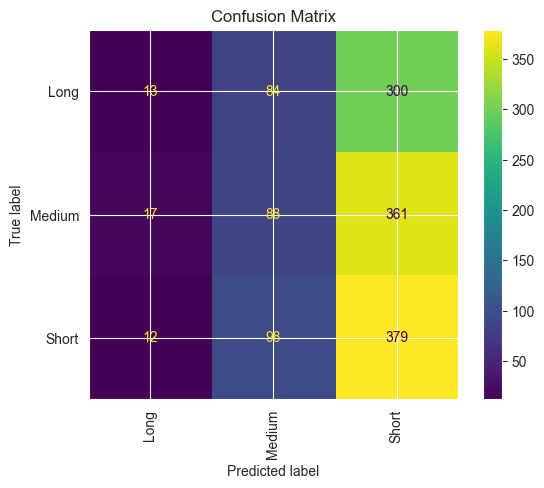


We can see a decrease in the loss to 1.0905 at epoch 1000 and a very slow decrease to 1.0900 At epoch 4830.

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.31 | 0.03 | 0.06 | 397 |
| 1 (Medium) | 0.33 | 0.19 | 0.24 | 466 |
| 2 (Short) | 0.36 | 0.78 | 0.50 | 489 |
| accuracy |  |  | 0.36 | 1352 |
| macro avg | 0.33 | 0.33 | 0.26 | 1352 |
| weighted avg | 0.34 | 0.36 | 0.28 | 1352 |

We can see that all scores are very low and that the recall and f1-score for the 0 class (long) is close to 0. There is no correlation between age and gender to the first hospitalization duration.

**Figure 2:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are very far from the actual values. From the confusion matrix we can see that from a total of 1352 samples in the test set only 480 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 35.5%.

**Conclusions:**

There is no correlation between the age and gender of the patient and the 1st hospitalization duration.

**Task 26. c. - Finding a connection between the age and sex of the patient to the number of hospitalization days in 2nd hospitalization (task 18)**

1. Creating a feature named ‘2nd\_hosp\_dur\_class’ for classifying the hospitalization durations of the 2nd hospitalizations by tertiaries of the ‘hospitalization durations’ feature.

Short - 0 to 1 days (0 to 33%) - 489 values

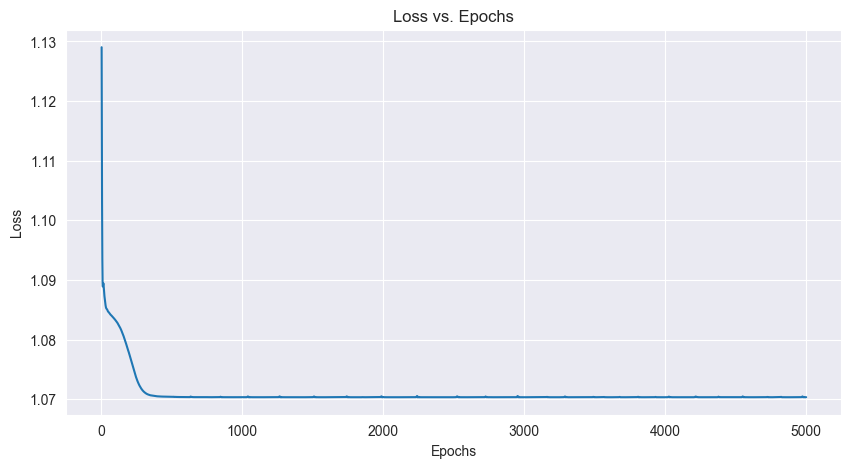
Medium - 2 to 3 days (34% to 66%) - 448 values

Long - 4 days and more (67% to 100%) - 405 values

1. Training a classification model.
   1. Target feature - 2nd\_hosp\_dur\_class. Encoded using Label encoder.
   2. Training features - age binned and gender. Encoded using Label encoder.
   3. Splitting data to 70% train 30% test.
   4. Model - NeuralNetworkClassifier - Torch nn
      1. Hidden size = 10
      2. Learning rate = 0.01
      3. Number of epochs = 5000
2. Evaluating the model classification.

**Results:**

**Figure 1:** Training loss vs epochs

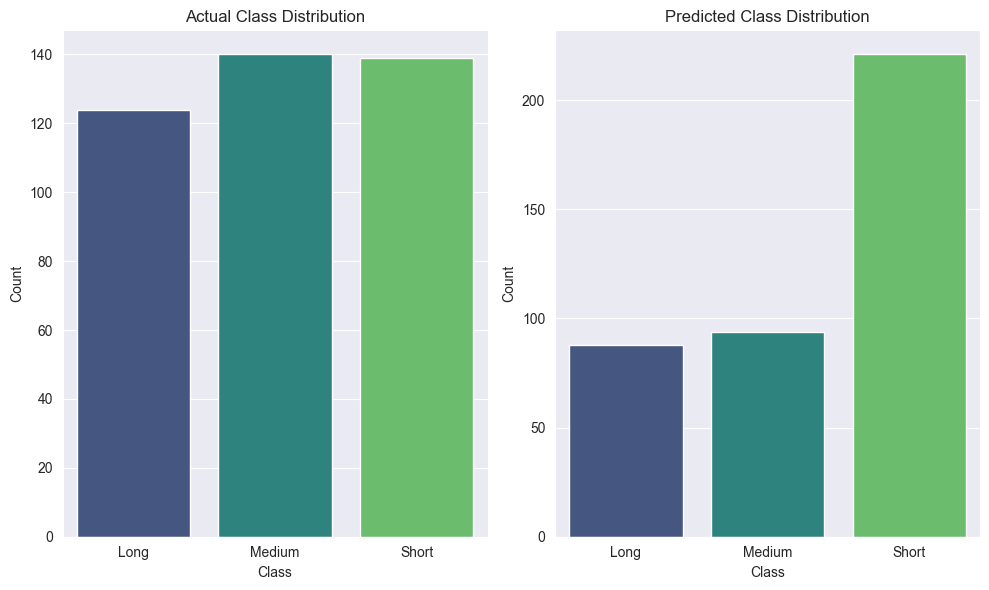
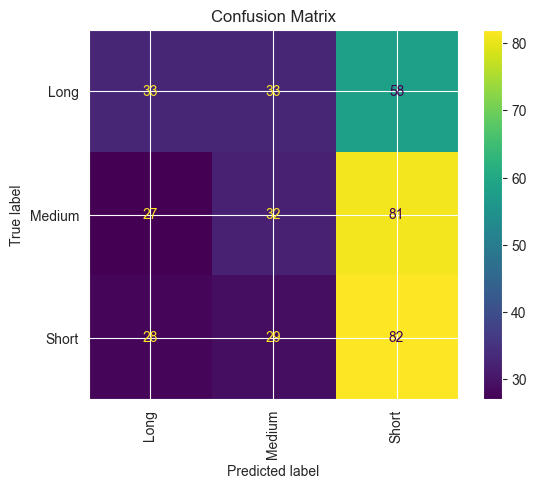


We can see a decrease in the loss to 1.0703 at epoch 670 and stayed constant after.

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.38 | 0.27 | 0.31 | 124 |
| 1 (Medium) | 0.34 | 0.23 | 0.27 | 140 |
| 2 (Short) | 0.37 | 0.59 | 0.46 | 139 |
| accuracy |  |  | 0.36 | 403 |
| macro avg | 0.36 | 0.36 | 0.35 | 403 |
| weighted avg | 0.36 | 0.36 | 0.35 | 403 |

We can see that all scores are very low. There is no correlation between age and gender to the 2nd hospitalization duration.

**Figure 2:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are very far from the actual values. From the confusion matrix we can see that from a total of 403 samples in the test set only 147 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 36.5%.

**Conclusions:**

There is no correlation between the age and gender of the patient and the 2nd hospitalization duration.

**Task 26. d. - Finding a connection between the age and sex of the patient to number of hospitalization days in 2nd and above hospitalizations (hospitalization duration in rehospitalizations)**

**Methodology:**

1. Creating a subset df from the ‘df\_combined\_26’ df that contains only the 2nd and above hospitalizations.
2. Creating a feature named ‘2nd\_plus\_hosp\_dur\_class’ for classifying the hospitalization durations of the 2nd and above hospitalizations by tertiaries of the ‘hospitalization durations’ feature.

Short - 0 to 1 days (0 to 33%) - 964 values

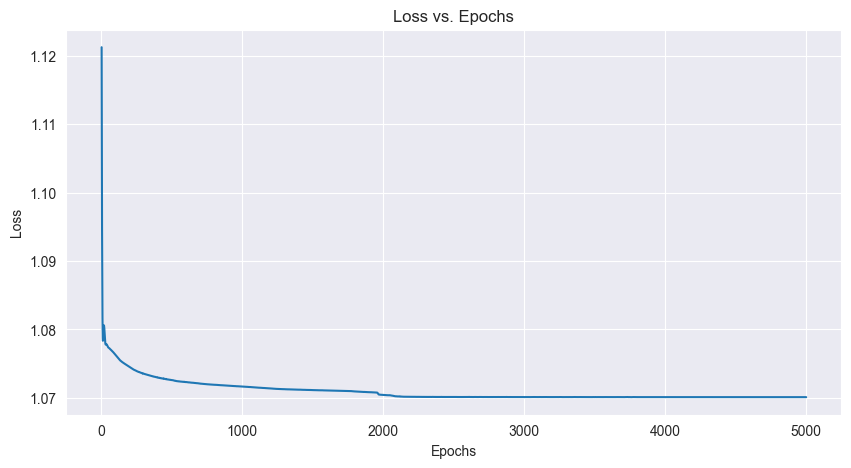
Medium - 2 to 3 days (34% to 66%) - 836 values

Long - 4 days and more (67% to 100%) - 700 values

1. Training a classification model.
   1. Target feature - 2nd\_hosp\_dur\_class. Encoded using Label encoder.
   2. Training features - age binned and gender. Encoded using Label encoder.
   3. Splitting data to 70% train 30% test.
   4. Model - NeuralNetworkClassifier - Torch nn
      1. Hidden size = 10
      2. Learning rate = 0.01
      3. Number of epochs = 5000
2. Evaluating the model classification.

**Results:**

**Figure 1:** Training loss vs epochs

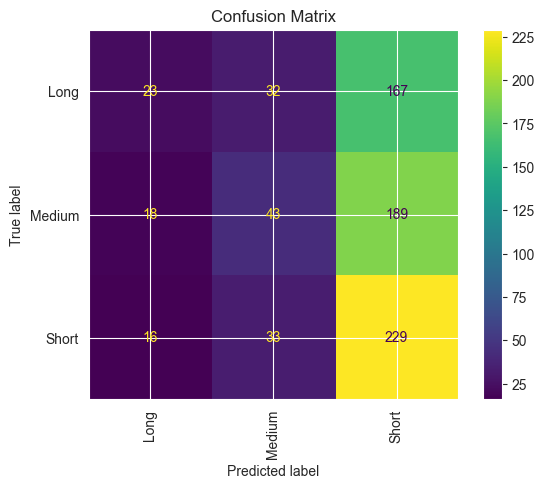


We can see a decrease in the loss to 1.0711 at epoch 1560 and stayed constant after.

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.40 | 0.10 | 0.16 | 222 |
| 1 (Medium) | 0.40 | 0.17 | 0.24 | 250 |
| 2 (Short) | 0.39 | 0.82 | 0.52 | 278 |
| accuracy |  |  | 0.39 | 750 |
| macro avg | 0.40 | 0.37 | 0.31 | 750 |
| weighted avg | 0.40 | 0.39 | 0.33 | 750 |

We can see that all scores are very low and that the recall and f1-score for the 0 class (long) is even lower. There is no correlation between age and gender to the hospitalization duration in rehospitalizations.

**Figure 2:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are very far from the actual values. From the confusion matrix we can see that from a total of 750 samples in the test set only 295 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 39.3%.

**Conclusions:**

There is no correlation between the age and gender of the patient and the hospitalization duration in rehospitalizations.

**Task 37 - Dimensionality Reduction of hospitalization 1 - in regards to hospitalization duration classification as the target prediction**

1. **Dimension reduction using Autoencoder**
2. **Classification without dimension reduction as a base-line**

**Task 37 a. Dimensionality Reduction of hospitalization 1 using Autoencoder - in regards to hospitalization duration classification as the target prediction**

**Methodology:**

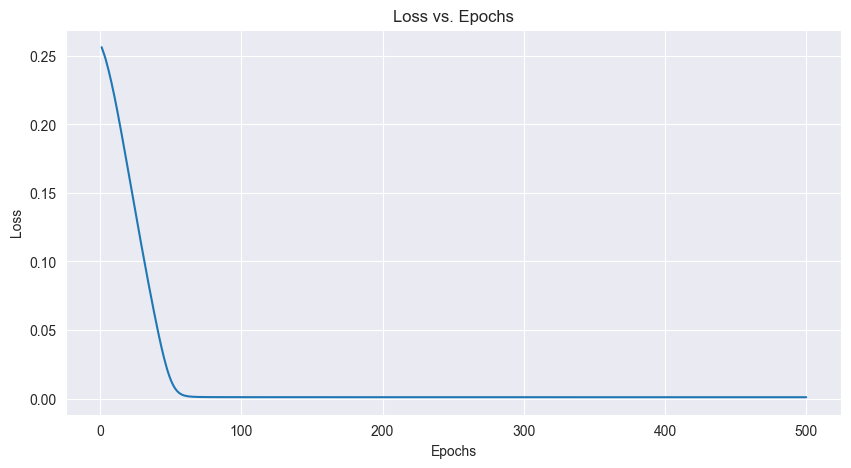
1. Splitting the admission diagnosis list and release diagnosis list to separate columns so that each diagnosis gets a column.
2. Filling empty values with empty strings.
3. Creating an hospitalization duration classification (‘hosp\_dur\_class’) by tertiaries:
   1. Short - 0 to 1 days (0 to 33%) - 2613 values
   2. Medium - 2 to 3 days (34% to 66%) - 2339 values
   3. Long - 4 days and more (67% to 100%) - 2054 values
4. Setting the relevant features to be processed by the autoencoder:

'Patient', 'department\_id','reception\_type', 'patient\_origin', 'Admission\_Entry\_Date', 'patient\_origin', 'Release\_Type', 'release\_doctor\_code', 'admission\_diagnoses\_1', 'admission\_diagnoses\_2', 'admission\_diagnoses\_3', 'admission\_diagnoses\_4', 'admission\_diagnoses\_5', 'admission\_diagnoses\_6', 'release\_diagnoses\_1','release\_diagnoses\_2','release\_diagnoses\_3','release\_diagnoses\_4','release\_diagnoses\_5','release\_diagnoses\_6','release\_diagnoses\_7','release\_diagnoses\_8','release\_diagnoses\_9','release\_diagnoses\_10'

1. Setting the target feature as ‘hosp\_dur\_class’.
2. Preprocessing the train features to tensors.
3. Training the autoencoder to create a 3 dimensional (hidden\_size = 3) numpy representation of the train features (500 epochs, learning rate 0.01).
4. Defining the model and training the classifier.
   1. Encoding the target feature using label encoder.
   2. Splitting data to 70% train 30% test and converting them to tensors.
   3. Model - NeuralNetworkClassifier - Torch nn
   4. Hidden size = 10
   5. Learning rate = 0.01
   6. Number of epochs = 5000
5. Evaluating the model classification.

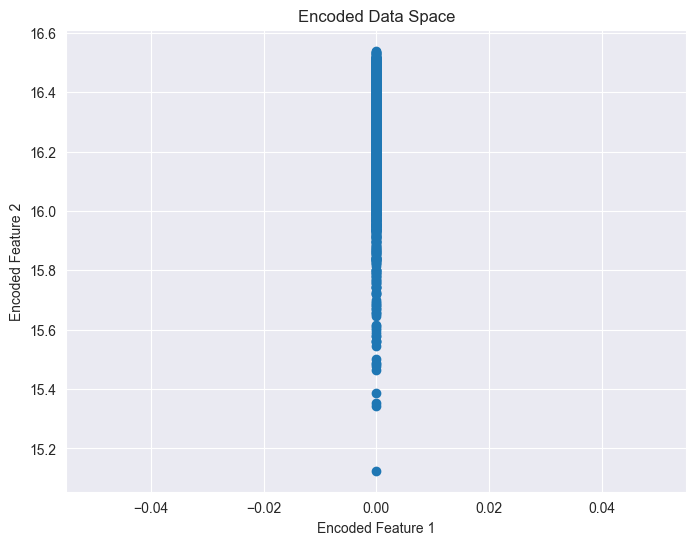
**Results:**

**Figure 1:** Training loss vs epochs of Autoencoder training

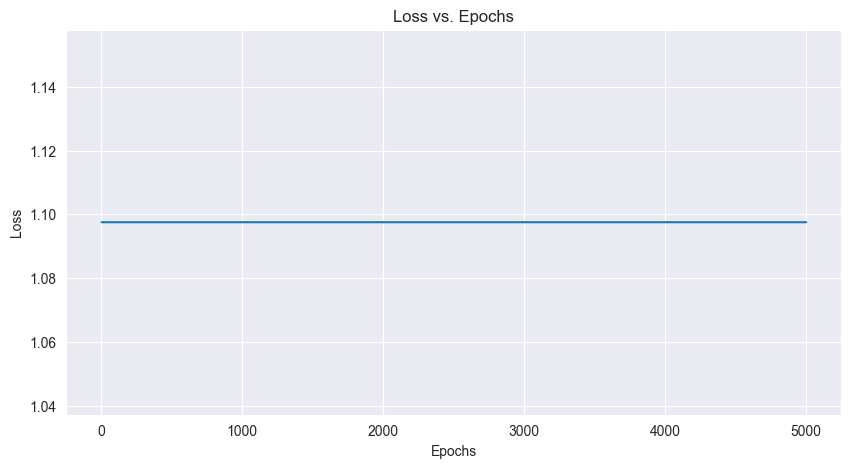


We can see a decrease in the loss to 0.0010 at epoch 80 and stayed constant after.

**Figure 2:** Encoded data space



**Figure 3:** Training loss vs epochs of classification training

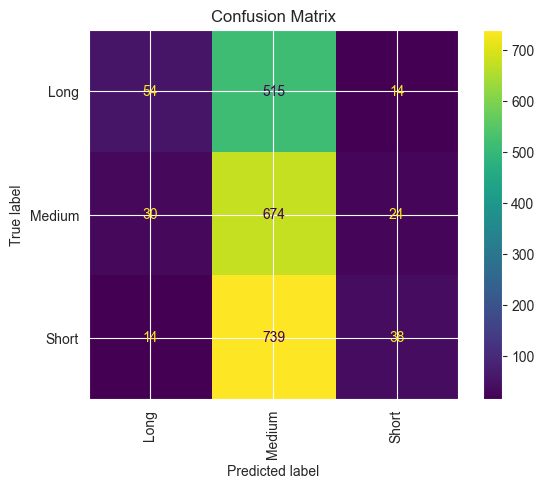


We can see that the loss stays constant, hence the model is not able to converge.

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.55 | 0.09 | 0.16 | 583 |
| 1 (Medium) | 0.35 | 0.93 | 0.51 | 728 |
| 2 (Short) | 0.50 | 0.05 | 0.09 | 791 |
| accuracy |  |  | 0.36 | 2102 |
| macro avg | 0.47 | 0.36 | 0.25 | 2102 |
| weighted avg | 0.46 | 0.36 | 0.25 | 2102 |

We can see that all scores are very low and that the recall and f1-score for the 0 class (long) is even lower.

**Figure 4:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are very far from the actual values. From the confusion matrix we can see that from a total of 2102 samples in the test set only 766 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 36.4%.

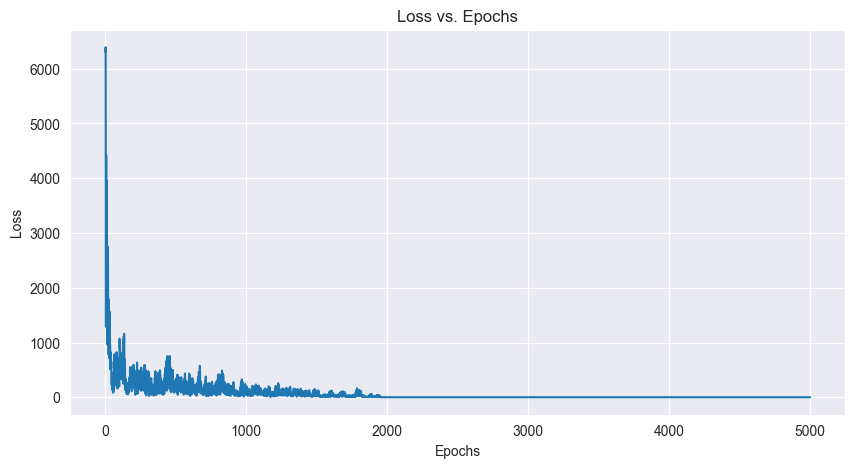
**37 b. Classification without dimension reduction as a base-line**

**Methodology:**

1. Encoding categorical features using label encoder.
2. Encoding target features using label encoder.
3. Converting datetime type feature to integer and standardizing it.
4. Defining the model and training the classifier.
   1. Splitting data to 70% train 30% test and converting them to tensors.
   2. Model - NeuralNetworkClassifier - Torch nn
   3. Hidden size = 10
   4. Learning rate = 0.01
   5. Number of epochs = 5000
5. Evaluating the model classification.

**Results:**

**Figure 1:** Training loss vs epochs of classification training

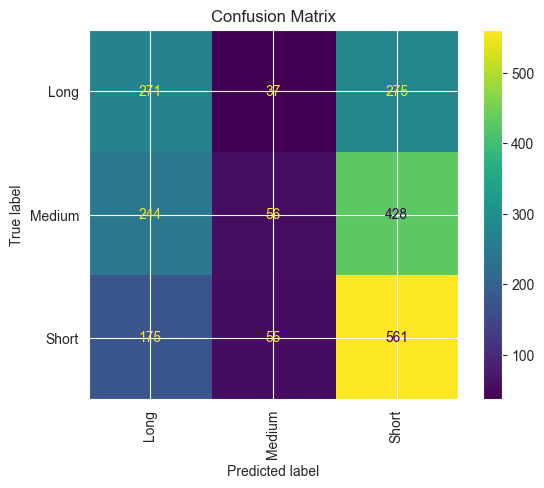


We can see a decrease in the loss with lots of fluctuations up to epoch 1960 and a slow decrease up to 1.0606 at epoch 5000

**Table 1:** Model evaluation scores on the test set

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| 0 (Long) | 0.39 | 0.46 | 0.43 | 583 |
| 1 (Medium) | 0.38 | 0.08 | 0.13 | 728 |
| 2 (Short) | 0.44 | 0.71 | 0.55 | 791 |
| accuracy |  |  | 0.42 | 2102 |
| macro avg | 0.40 | 0.42 | 0.37 | 2102 |
| weighted avg | 0.41 | 0.42 | 0.37 | 2102 |

We can see that all scores are very low and that the recall and f1-score for the 1 class (medium) is even lower.

**Figure 4:** Actual class distribution vs predicted class distribution and Confusion matrix

In the actual vs predicted class distribution plots, we can see that the predicted classifications for all classes are very far from the actual values. From the confusion matrix we can see that from a total of 2102 samples in the test set only 888 (diagonal sum) were predicted correctly, meaning an overall accuracy of only 42.2%.

**Conclusions:**

Performing dimensional reduction needs to be oriented for a given classification task. We have selected to try to find a relationship between the possible relevant features to the hospitalization duration (after converting it to 3 classes).

As we are in a deep learning course, we chose to use an autoencoder for performing dimension reduction. Reducing to two and three dimensions gave the same result (we showed here only the 3D reduction).

Both classification after a dimensional reduction using an autoencoder and without, resulted in poor classification accuracy. Meaning there is no relationship between the used features from the hospitalization 1 data set and the hospitalization duration.