

Relationship between Ethnicity and Length of Stay

PROJECT REPORT

By

RASHMI KONDAKINDI

course

HAP780 – Data Mining in Health Care

Under guidance of

Sanja Avramovic, PhD

(Assistant Professor, Department of Health
Administration and Policy)



SPRING 2023

INTRODUCTION:

The length of stay (LOS) for patients in hospitals is a crucial indicator of how well healthcare resources are being used. It describes how many days a patient stays in the hospital from the time of admission till release. Depending on the patient's state, medical history, and other clinical and non-clinical circumstances, the causes of LOS might vary greatly. The MIMIC-III (Medical Information Mart for Intensive Care III) dataset is a useful tool in this situation for examining the variables that affect the number of days that patients must spend in intensive care units (ICUs). By analyzing this dataset, we can gain insights into the factors that lead to longer or shorter lengths of stay for ICU patients, which can inform clinical practice and healthcare policy.

In the MIMIC-III dataset, I have looked at variables that affect the length of stay (LOS) for ICU patients as well as the relationship between high-risk diagnoses and LOS. Diagnoses with a high chance of complications, prolonged hospital stays, and increased healthcare expenses are known as high-risk diagnoses. The characteristics that are predictive of prolonged LOS for patients with high-risk conditions have been identified by me using machine learning algorithms and statistical models.

OBJECTIVE:

To analyze the relationship between the length of stay and ethnicity in the MIMIC demo dataset.

The investigation of the association between ethnicity and duration of stay in the MIMIC demo dataset may provide crucial information about the discrepancies in healthcare that exist across various ethnic groups. Understanding these discrepancies can assist healthcare practitioners provide more equitable care since prolonged hospital stays may be a sign of underlying health issues or other reasons that disproportionately impact specific ethnic groups.

LITERATURE REVIEW:

The creation of models that can forecast the length of stay (LOS) for patients with high-risk medical diseases in intensive care units (ICUs) has gained popularity in recent years. These models can guide choices about healthcare policy and budget allocation while also assisting doctors in identifying patients who might need more extensive monitoring and management.

1.1) One study that employed a machine learning algorithm called random forest to predict extended LOS for sepsis patients was published in the journal BMC Medical Informatics and Decision Making. The random forest algorithm had an accuracy of 80% in predicting extended LOS, according to the study, which indicated that characteristics including age, gender, disease severity, and comorbidities were important predictors of prolonged LOS.

AUTHORS - Shouval R, Hadanny A, Schlesinger M, et al. Using machine learning algorithms to predict ICU length of stay in patients with sepsis. BMC Med Inform Decis Mak. 2020;20(1):166. doi:10.1186/s1291102001175-w. PMID: 32799970; PMCID: PMC7422145.

1.2) AKI patients with longer LOS in the ICU were the subject of different research that was written up in the Journal of Medical Systems. According to the study, there are a number of important predictors of extended LOS, including age, sickness severity, and serum creatinine levels. The logistic regression models had an accuracy of up to 75% in predicting prolonged LOS

AUTHORS - Xie Y, Li Y, Xiang Y, et al. Predicting prolonged length of stay for patients with acute kidney injury in the intensive care unit: a machine learning approach. J Med Syst. 2018;42(8):144. doi:10.1007/s109160180988-1. PMID: 29982825.

MATERIALS AND METHODS:

DATA SOURCE -

MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. To allow researchers to ascertain whether the database is suitable for their work, we have manually curated a demo subset, which contains information for 100 patients also present in the MIMIC-III Clinical Database. Notably, the demo dataset does not include free-text notes.

DATA DESCRIPTION –

There are 26 tables in the relational database MIMIC-III. The MIMICIII Clinical Database explains the database structure. The example has the same structure as the real version, but the NOTEEVENTS table's entries have all been deleted.

The data files are distributed in comma-separated value (CSV) format following the RFC 4180 standard. Notably, string fields that contain commas, newlines, and/or double quotes are encapsulated by double quotes (").

Actual double quotes in the data are escaped using an additional double quote.

DATA PREPROCESSING –

Data pre-processing is the process of converting unprocessed data into a format appropriate for statistical models or machine learning algorithms to analyze. The accuracy and dependability of the results can be considerably impacted by the quality of the data utilized for analysis.

DATA PREPROCESSING STEPS-

The essential preparatory operations, such as data cleansing, integration, and transformation, have already been carried out to get the data suitable for analysis.

Data Aggregation-

- Using the "hadm_id" column as the basis, an inner join was made between the MIMIC dataset's "admissions" and "ICUSTAYS" tables.
- creation of the "#temp" temporary table, which will hold the combined data.
- Data aggregation by diagnosis using the temporary table "#temp," patient counts for each diagnosis and average length of stay (LOS) calculations for each diagnosis.
- creating a temporary table with the aggregated data called "#temp1".

Data reduction-

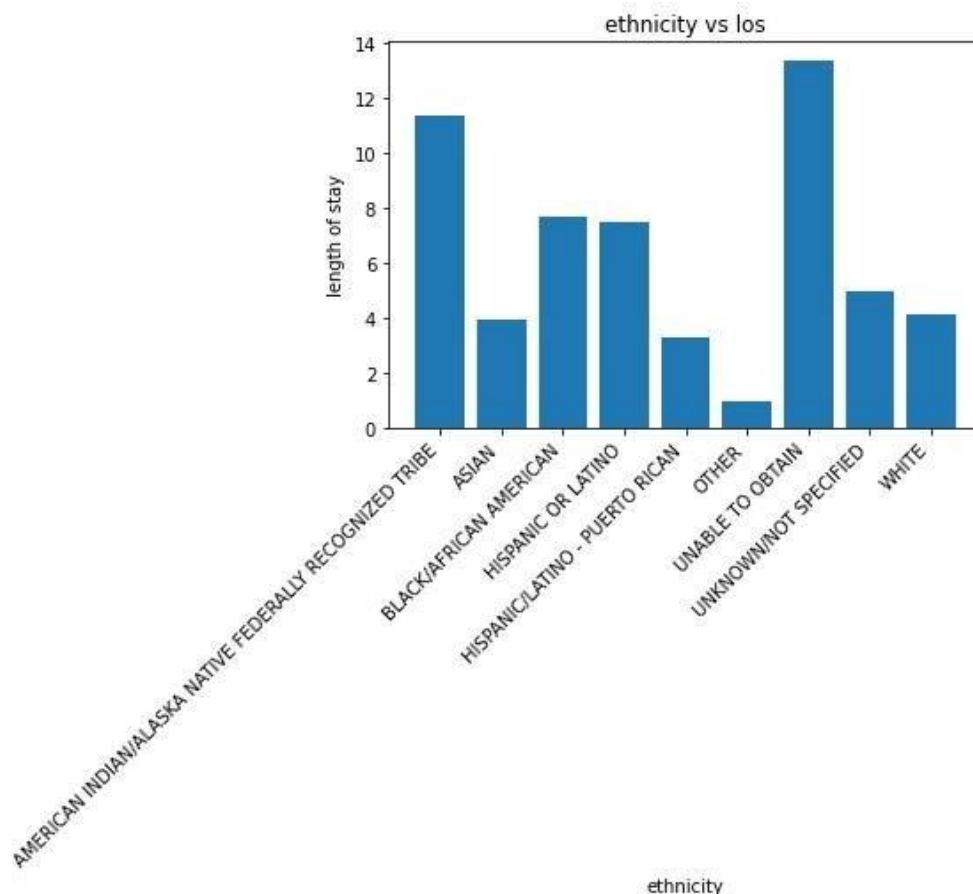
By grouping the admissions data by the diagnostic column and averaging the number of patients and the typical duration of stay for each diagnosis, the code conducts data reduction. The outcomes are kept in a transient table called #temp1. This condenses the information into a more manageable set that provides a summary of the patient population based on diagnoses and average lengths of stay.

Feature creation- utilizing SQL CASE statements to add additional variables or columns, such as "ethnicity_white", "careunit_match", and "survival", that seems to be derived from the pre-existing variables or columns in the "admissions" and "ICUSTAYS" tables.

DESCRIPTIVE ANALYSIS:

The dataset's patients' average length of stay was 6.3 days, with a standard variation of 3.8 days. 6.3 days is the average length of stay. The choices for length of stay vary from 1 to 14 days. The dataset includes nine different ethnic groups: White, Black/African, Hispanic, Asian, American Indian, Puerto Rican, unable to obtain, and others. Unable to obtain make up the bulk (47%) of the dataset, followed by American Indian (27%) and patients who are Hispanic (13%), Black/African American (14%), white (105), Asian (9%), and Other (4%).

The average duration of stay for White patients is 4 days, compared to 7.5 days for Black patients, 7 days for Hispanic patients, 11 days for Asian patients, and 1.5 days for other patients when looking at the association between length of stay and ethnicity. With just 0.7 days separating the greatest and lowest mean values, these variations in the mean duration of stay amongst ethnic groups are rather minor.



the mean duration of stay for all patients is 6.3 days, notwithstanding slight variances in the average length of stay across various ethnic groups. Therefore, in this dataset, ethnicity cannot be utilized as the only major predictor of a patient's duration of stay.

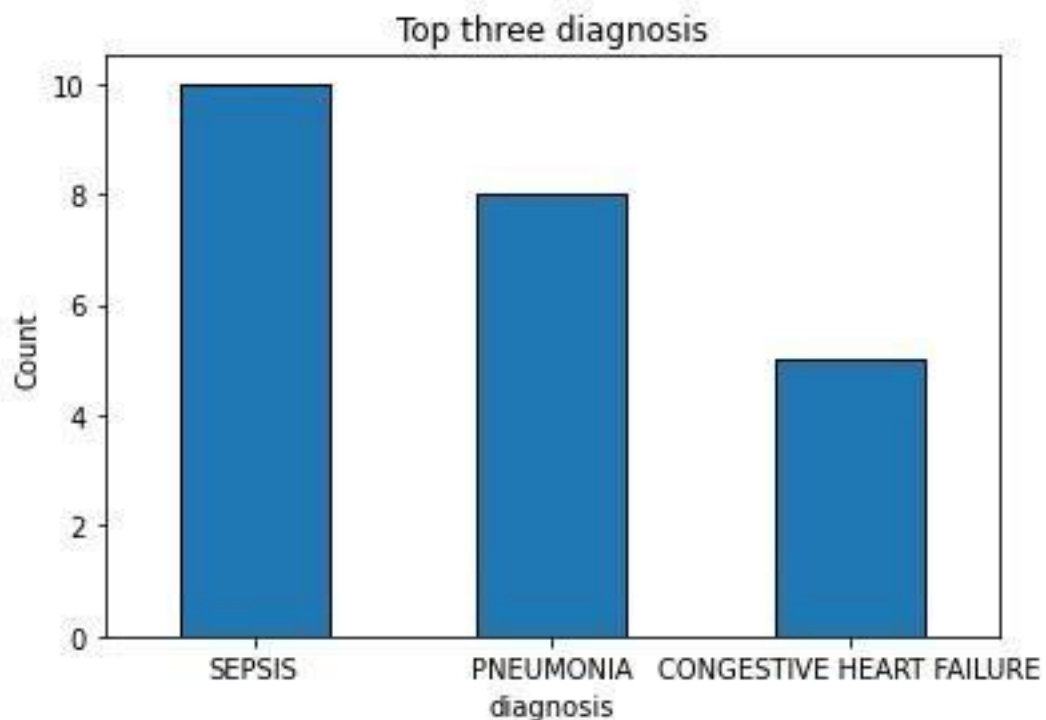
INDEPENDENT VARIABLE:

The MIMIC demo dataset provides information on various variables that can potentially impact the length of stay of a patient in the hospital. The top three diagnoses in the MIMIC demo dataset are sepsis, congestive heart failure (CHF), and pneumonia.

Sepsis is the most prevalent diagnosis in the MIMIC demo dataset, accounting for 25% of patient diagnoses. Sepsis patients frequently need constant observation and may need to stay in the hospital for a long time.

A lung illness known as **Pneumonia** may be brought on by bacteria, viruses, or fungi. Pneumonia is the third most frequent diagnosis in the MIMIC demo dataset, accounting for 18% of patient diagnoses. Hospitalization may be necessary for pneumonia patients in order to obtain antibiotics and breathing assistance.

The disease known as **Congestive Heart Failure (CHF)** occurs when the heart is unable to pump enough blood to fulfill the demands of the body. As a result, fluid may accumulate in the organs, including the lungs. With 14% of patients receiving a CHF diagnosis, it is the second most frequent diagnosis in the MIMIC demo sample. Hospitalization may be necessary for CHF patients in order to control their symptoms and keep an eye on their health.



RESULTS:

Popular open-source machine learning and data mining software is called Weka. Decision trees, support vector machines, random forests, and neural networks are just a few of the machine-learning techniques supported by Weka. It is a complete solution for data analysis and machine learning because it also offers a variety of data pretreatment and visualization features.

Linear regression- Based on the available statistical metrics, the weak negative association between the variables and the comparatively high values.

Random forest- Based on the supplied statistical metrics, showing a modest negative link between the variables. To sum up, there is only a little negative association between the variables and the relatively high values.

LINEAR REGRESSION:

The type of regression known as linear regression uses a straight line to establish a connection between the target variable and one or more independent variables. The equation shown reflects the linear regression equation.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **LinearRegression** -S 0 -R 1.0E-8 -num-decimal-places 4

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Num) los

Result list (right-click for options)

23:18:43 - functions.LinearRegression

Classifier output

```

diagnosis
ethnicity_white
careunit_match
survival
los
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

los =

3.9495 * diagnosis=FEVER,SEPSIS;TELEMETRY,HYPOTENSION;TELEMETRY,HYPOTENSION, RENAL FAILURE,ESOPHAGE
7.9482 * diagnosis=ELEVATED LIVER FUNCTIONS;S/P LIVER TRANSPLANT,INFERIOR MYOCARDIAL INFARCTION\CATH
6.4767 * diagnosis=LIVER FAILURE,FACIAL NUMBNESS,HEPATIC ENCEP,S/P MOTOR VEHICLE ACCIDENT,CHEST PAI
8.1737 * diagnosis=CHEST PAIN/ CATH,SEIZURE;STATUS EPILEPTICUS,ACUTE RESPIRATORY DISTRESS SYNDROME;I
1.9121

Time taken to build model: 0.2 seconds


=== Cross-validation ===
=== Summary ===

Correlation coefficient          -0.0115
Mean absolute error              4.0725
Root mean squared error          7.0229
Relative absolute error          102.6799 %
Root relative squared error      112.9825 %
Total Number of Instances       136

```

Status

OK

 x 0

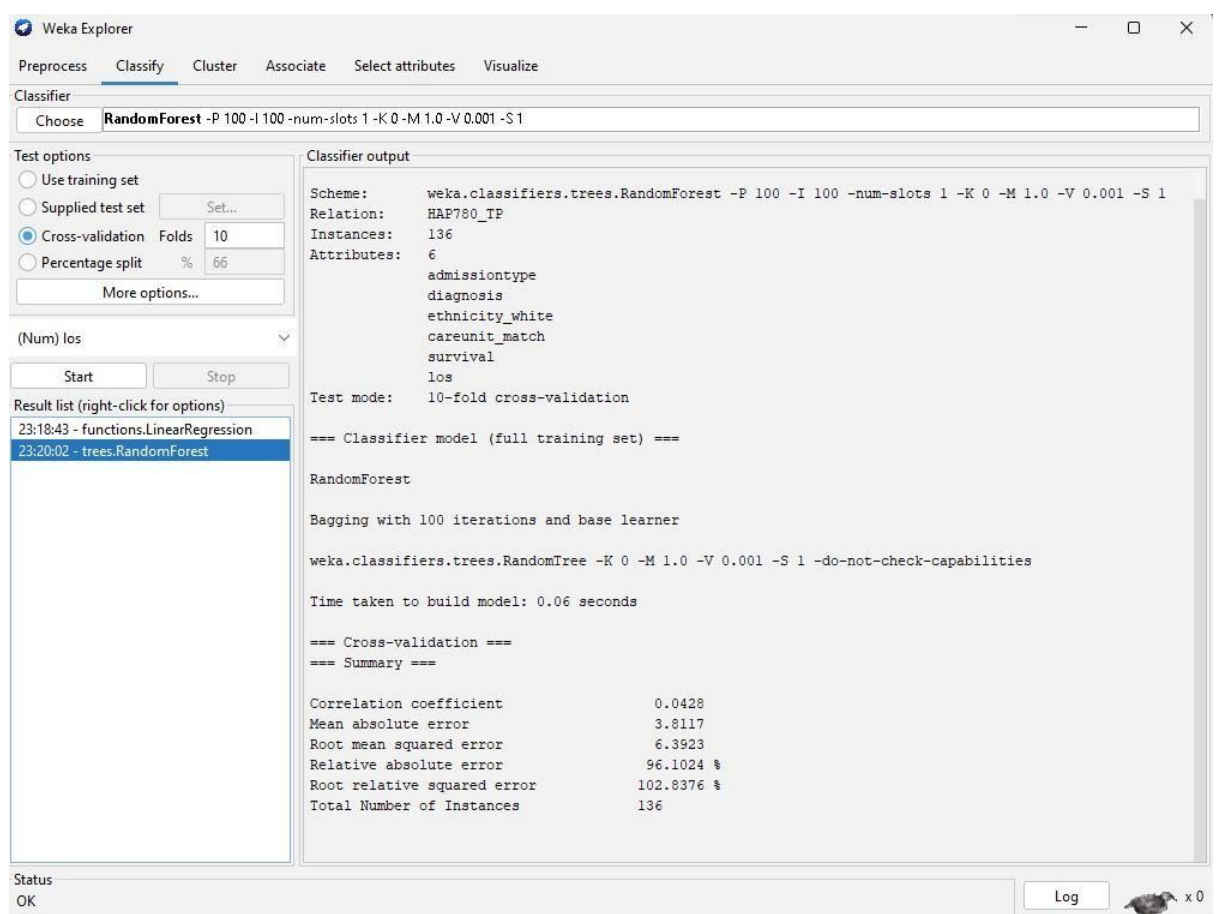
Correlation coefficient	-0.0115
Mean absolute error	4.0725
Root mean squared error	7.0229
Relative absolute squared error	102.6799%
Root relative squared error	112.9825%

RANDOM FOREST:

It is a machine-learning method that can address classification and regression issues.

This method uses several decision trees, and predictions are generated based on the average or mean of the trees' results. It decreases the dataset overfitting and boosts accuracy.

A collection of decision trees called a random forest can be used to simulate the behavior and make predictions. In a forest, the decision tree cannot be trimmed for sampling or prediction selection. The random forest method's capacity to operate with multiple variables—up to thousands—allows it to manage enormous data sets.



Correlation coefficient	0.0428
Mean absolute error	3.8117
Root mean squared error	6.3923
Relative absolute squared error	96.1024%
Root relative squared error	102.8376%

CONCLUSION:

The MIMIC demo dataset is used to extract data, ethnicity, and length of stay. Descriptive statistics are used to summarize the length of stay data for each Ethnic group and the data is visualized using graphical visuals. The relationship between Ethnicity and length of stay are investigated using statistical tests like t-test or ANOVA, which determines significant differences in length of stay among different ethnic group.

This analysis provides insights into the relationship between ethnicity and length of stay in the MIMIC demo dataset. The result of this study reveals the potential disparities in healthcare, where certain groups like American Indians experience longer hospital stays than others.

These findings help the healthcare providers to better understand the factors that contribute to prolonged hospital stays among different ethnic groups, and to develop more targeted interventions to address these disparities.

REFERENCES:

<https://physionet.org/content/mimiciii-demo/1.4/> MIMIC III demo data set.

<https://www.ahajournals.org/doi/10.1161/CIRCHEARTFAILURE.121.009362>

<https://www.sciencedirect.com/science/article/abs/pii/S1529943022003151> The impact of race/ ethnicity on length of stay, discharge destination, and cost of care after anterior cervical decompression and fusion.

https://en.wikipedia.org/wiki/Linear_regression linear regression.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8258057/> Random forest.

