DATA SCIENCE QHO636

Individual COURSEWORK

Tutor: Dr Hakim Mezali Student: Maria Moraru, 10103766

BSC Computer Science SEMESTER B, London, May 2021/2022

Solent University

QA Higher Education

Contents

[Model 2 Inpatient Readmission 3](#_Toc104805005)

[Introduction 3](#_Toc104805006)

[Methods 3](#_Toc104805007)

[Dataset Preparation - Exploratory Data Analysis 4](#_Toc104805008)

[Analysis and Results 4](#_Toc104805009)

[Data Modelling and Visualisation 6](#_Toc104805010)

[Evaluation 7](#_Toc104805011)

[Limitations and Challenges 8](#_Toc104805012)

[Conclusion 8](#_Toc104805013)

[Model 3 Predicting house prices in the US 8](#_Toc104805014)

[Introduction 8](#_Toc104805015)

[Methods 8](#_Toc104805016)

[Dataset Preparation - Exploratory Data Analysis 8](#_Toc104805017)

[Analysis and Results 9](#_Toc104805018)

[Data Modelling and Visualisation 10](#_Toc104805019)

[Evaluation 10](#_Toc104805020)

[Limitations and Challenges 10](#_Toc104805021)

[Conclusion 10](#_Toc104805022)

[Appendix 15](#_Toc104805023)

[References 43](#_Toc104805024)

# Model 2 Inpatient Readmission

## Introduction

In this coursework I will predict which patients are at risk of emergency admission within 45 days after being discharged. I received the subset from my seminar tutor, Dr. Hakim Mezali.

The Government is allocating the highest budget to the health system in UK. Contrary to that, there is a huge back lock for waiting for an outpatient appointment, between 6-9 months. The NHS system in UK has two main issues, first is their bad management and second is the ageing population. I will test which of the explanatory variables will better predict if a patient is at risk of emergency admission within 45 days after being discharged. This research worth doing to save NHS money by preventing the risk of a patient to be admitted to A&E within 45 days after being discharged, by giving them outpatient appointments.

The subchapters in this coursework are as instructed on SOL page in the Assignment template and the numbering (a, b, c, d) is as instructed in the coursework provided at the seminar.

The length of stay of patients is an important measure of hospital management. Hospitals have limited resources, so they need to allocate beds and clinician time efficiently. The pandemic serves as an example of this notion. More than ever before, it is in the best interest of patients, hospitals and the public to limit hospital stays and to have an idea if a patient is going to be readmitted to hospital.[[1]](#footnote-1)

Being able to predict if a patient will be readmitted with the information available as soon as they enter the hospital and are diagnosed can therefore have many positive effects for a hospital and its efficiency.[[2]](#footnote-2)

## Methods

For this subset, a logistic regression model or Sigmoid Function will be used to predict patients that are at risk of emergency admission within 45 days after discharge. I used Python language in Anaconda with Jupyter notebook.

Regression only works with numerical values. y is a dependent variable, a number, quantitative: to predict whether a patient is at risk of emergency admission (coded as 1) or not at at risk of emergency admission (coded as 0).

  Logistic Regression Linear Regression works with numeric continuous data and therefore it is not suitable for predicting qualitative or categorical outcomes. To predict binary categorical outcomes, we use a particular classification technique called Logistic Regression. Since Logistic Regression predict binary outcomes, such as Pass/Fail, True/False, Yes/No, we use a Logistic Response Function.

The probabilities on the Logistic Response Function range from 0 to 1. We need to set the threshold or range values for which either 0 or 1 is predicted. For example, we can set this threshold to be 0.4 and thus for probabilities from 0 to 0.39 the value of 0 will be predicted and for probabilities from 0.4 to 1 the value of 1 will be predicted. [[3]](#footnote-3)

P(0) and P(1) can be used to represent two different outcomes. As such:

𝑃(0)=1−𝑃(1)P(0)=1−P(1)

𝑃(𝑦=1)=11+𝑒−(β0+β1𝑥1+β2𝑥2+⋯+β𝑘𝑥𝑘)P(y=1)=11+e−(β0+β1x1+β2x2+⋯+βkxk)

𝑂𝑑𝑑𝑠=𝑒−(β0+β1𝑥1+β2𝑥2+⋯+β𝑘𝑥𝑘)Odds=e−(β0+β1x1+β2x2+⋯+βkxk)

𝐿𝑜𝑔𝑖𝑡=𝑙𝑛(𝑂𝑑𝑑𝑠)=β0+β1𝑥1+β2𝑥2+⋯+β𝑘𝑥𝑘Logit=ln(Odds)=β0+β1x1+β2x2+⋯+βkxk

The full Python code can be found in Appendix, Figure 30.

### Dataset Preparation - Exploratory Data Analysis

(a)In step one, the exploratory analysis I found that there are no missing values or outliers, all independent variables are numerical and concluded that the data is clean.

There are 22 columns and excepting the first and the ID column, we are left with 20 independent variables as explained below:

* PrRespInfect: Previous history of respiratory infection
* PrRefCon: Cause of emergency admission is same as previous admission
* no\_emergency\_admissions: The number of emergency admissions prior to the current admission
* no\_AnEAtten: The number of A&E attendances prior to the current admission
* noOutpatientsAtten: The number of outpatient attendances prior to the current admission
* DiseaseYN: Existence of specific diseases, e.g., alcohol related disorders, COPD, diabetes, sickle cell diseases
* Severity index (ranging from 0 – 8)
* Previous hospital length of stay (los)
* Age, gender, religion (Christian, Muslim, Hindu, Sikh, and Other) ▪ Index of Multiple Deprivation (IMD) Score
* Seven Indicators related to IMD
* Barriers to housing, Crime and Disorder, Education Skills and Training, Employment, Health Deprivation and Disability, Income, and Living Environment

There are 1489 cases as we can see from the data shape. The distribution is parametric.

Prior to building a Logistic Regression model, data is split into training and test sets. The model is then built on the training set and tested on the test set. For this dataset I divided 20% for testing and 80% for training.

One way to ensure that the model performs well is to scale the features or columns of data that make up the independent variables. I scaled the features to improve the model performance.

## Analysis and Results

A confusion matrix can be used to assess the performance of a Logistic Regression model.

|  |  |  |
| --- | --- | --- |
|  | Prediction = 0 | Prediction = 1 |
| Actual = 0 | True Negative | False Positive |
| Actual = 1 | False Negative | True Positive |

Sensitivity=TruePositive / (TruePositive+FalseNegative)

Sensitivity analysis shows the percentage of positive cases that are correctly identified.

Specificity=TrueNegative / (TrueNegative+FalsePositive)

Specificity analysis shows the percentage of negative cases that are correctly identified. Setting the model threshold high will result in lower sensitivity and higher specificity. Setting the model threshold low will result in higher sensitivity and lower specificity.[[4]](#footnote-4)

(b) Prior to building a Logistic Regression model, I split the data into training (80%) and test sets (20%). The Logistic Regression model can now be built on the training set and tested on the test set. To evaluate the model and get his effectiveness, I used the confusion matrix.

**MODEL 1:** Considering all 22 dependent variables and 1191 cases (80% of the data), I found that the model performance is 97%, which is very good.

The outcome of the confusion matrix for the trained dataset was array ([[112, 7],

[ 2, 177]], dtype=int64)

The value of 112 in the top left-hand corner of the above matrix shows the number of times that our model correctly predicts 0 when the actual values are 0. The value of 177 in the bottom right-hand corner of the above matrix shows the number of times our model correctly predicts 1 when the actual values are 1. We can calculate the accuracy of our model as follow.

(112+177) / (112+177+7+2) = 0.9697986577181208

The above score shows that our model correctly predicts the binary values in about 97% of the times. Another way of performing the same calculation is accuracy\_score(y\_test, y\_pred) which prints the accuracy score.

Misclassification rate: (2+7) / (112+7+2+177) = 0.03

Sensitivity: 112/119 = 0.94

Specificity: 177/179 = 0.98

**MODEL 2:** Considering first 12 dependent variables and 1191 cases (80% of the data), I found that the model performance is 59%, which is very good.

Text

Description automatically generated

A picture containing shape

Description automatically generated

Figure A

Figure A shows that the model considering difference of days since last admission to A&E, previous history of respiratory infection, cause of emergency admission is same as previous admission, the number of emergency admissions prior to the current admission, the number of outpatient attendances prior to the current admission, existence of specific diseases, religion, length of stay, age, gender and barriers to housing are the most important features in predicting a patients readmission.

**MODEL 3:** Considering first 11 dependent variables and 1191 cases (80% of the data), I found that the model performance is 59%, as showed in Figure B, which is good.

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Figure B

**MODEL 4:** Considering first 10 dependent variables and 1191 cases (80% of the data), I found that the model performance is 24%, which is bad.

**MODEL 5:** Considering first 9 dependent variables and 1191 cases (80% of the data), I found that the model performance is 95%, which is very good.

**MODEL 6:** Considering first 8 dependent variables and 1191 cases (80% of the data), I found that the model performance is 83%, which is good.

**MODEL 7:** Considering first 7 dependent variables and 1191 cases (80% of the data), I found that the model performance is 7.71%, which is very bad model.

**MODEL 8:** Considering first 6 dependent variables and 1191 cases (80% of the data), I found that the model performance is 18.12%, which is very bad model.

**MODEL 9:** Considering first 5 dependent variables and 1191 cases (80% of the data), I found that the model performance is 23.82%, which is very bad model.

**MODEL 10:** Considering first 4 dependent variables and 1191 cases (80% of the data), I found that the model performance is 85.90%, which is very good model.

**MODEL 11:** Considering first 3 dependent variables and 1191 cases (80% of the data), I found that the model performance is 93.95%, which is very good model.

**MODEL 12:** Considering first 2 dependent variables and 1191 cases (80% of the data), I found that the model performance is 5.70%, which is very bad model.

The response variable has two classes, readmitted within 45 days of discharge, and not readmitted within 45 days, which is coded as ‘1’ and ‘0’, respectively.

Let be the probability of belonging to class for given explanatory fields, e.g., daydiff, age, gender, etc.

I will use a cut-off point of 0.5.

The Significance level (P value of Alpha) for ‘daydiff’ is 0, which is less than the cut-off point and we use this in our modelling.

### Data Modelling and Visualisation

The equations is as follows

Log [p(c1│Readmission\_45) / 1- p(c1│ Readmission\_45)] = -0.0857 + daydiff + 2 PrRespInfect + 6.26 noOutpatientsAtten + 15 DiseasesYN + 28.15 GroupedReligion + 31.01 speldur + 4.20 BarriersToHousing + 2.77 CrimeandDisorderScore + 3.94 EmploymentScore + 5.71 LivingEnvironmentScore + 2.96 LivingEnvironmentScore + 2.36 severity

The probability that a patient has an emergency admission within 45 days of discharge can be determined by

p(c1│Readmission\_45) = exp (-0.0857 + daydiff + 2 PrRespInfect + 6.26 noOutpatientsAtten + 15 DiseasesYN + 28.15 GroupedReligion + 31.01 speldur + 4.20 BarriersToHousing + 2.77 CrimeandDisorderScore + 3.94 EmploymentScore + 5.71 LivingEnvironmentScore + 2.96 LivingEnvironmentScore + 2.36 severity) / 1+ exp (-0.0857 + daydiff + 2 PrRespInfect + 6.26 noOutpatientsAtten + 15 DiseasesYN + 28.15 GroupedReligion + 31.01 speldur + 4.20 BarriersToHousing + 2.77 CrimeandDisorderScore + 3.94 EmploymentScore + 5.71 LivingEnvironmentScore + 2.96 LivingEnvironmentScore + 2.36 severity)

Using the logreg function the model is as follows from Figure C.

Log [p(c1│Readmission\_45) / 1- p(c1│ Readmission\_45)] = -0.0857 – 0.09PrRespInfect + 0.05CrimeandDisorderScore + 0.83EducationSkillsandTrainingScore – 1.46 HealthDeprivationandDisability – 0.04 IncomeScore -26.56 LivingEnvironmentScore + 0.54severity

Table

Description automatically generated

Calendar

Description automatically generated with medium confidence

Figure C

## Evaluation

For Regression Equation: Readmission\_45 = 0.837 -0.003 \* daydiff

The model predicted the Readmission\_45 of 0.66 in that market

R squared value of the model is 43.06

Conclusion: 43% of the data fit the multiple regression model. The independent variables ‘daydiff’ have explained the variation of ‘Readmission\_45’ by 43%. There is 57% that remains unexplained.

The Mean Absolute Error is 0.318.

For Equation: Readmission\_45 = 0.8361+ (-0.003 \* daydiff) + (0.031 \* PrRespInfect)

43% of the data fit the multiple regression model. The independent variables ‘daydiff’ have explained the variation of ‘Readmission\_45’ by 43%. There is 57% that remains unexplained. The Mean Absolute Error is 0.318.

### Limitations and Challenges

It was hard to understand how to buid a model with different independent variables. I needed that in order to find the best model using Forward and Backward elimination.

In Backwards elimination, I got the error of continuous values for the last 10 columns and had to start building models from the first 11 columns, 10 columns, all the way down to the 2nd and 1st independent variable.

## Conclusion

Model 1 has the best accuracy score and will be used in predicting if a patient will be readmitted to hospital within 45 days. The variables included in the model are the most important features in predicting a patients readmission.

# Model 3 Predicting house prices in the US

## Introduction

In this chapter, I will develop a linear model to predict American house prices. All the figures can be found in Appendix. In the Figure 1, I imported the libraries and requested information about the dataset.

## Methods

For this dataset, a multiple linear regression (MLR) model will be used to predict House prices in America. I used Python language in Anaconda with Jupyter notebook.

Regression only works with numerical values. y is a dependent variable, a number, quantitavive: to predict house prices.

Multiple linear regression has one y and two or more x variables. It is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable.

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

Assumptions for Multiple Linear Regression: 1. A linear relationship should exist between the Target and predictor variables. 2. The regression residuals must be normally distributed. 3. MLR assumes little or no multicollinearity (correlation between the independent variable) in data.

Formula: Y = β0 + β1X1 + β2X2 + β3X3 + ... + βnXn + e

Y = Dependent variable / Target variable

β0 = Intercept of the regression line

β1, β2,..βn = Slope of the regression lime which tells whether the line is increasing or decreasing

X1, X2,..Xn = Independent variables / Predictor variables

e = Error[[5]](#footnote-5)

### Dataset Preparation - Exploratory Data Analysis

1. **Initial Investigation**

The dataset has 12 independent variables (x) and 1 dependent variable (y) that is called Total.

There are 60 cases and 14 columns (including first column that counts the cases) as seen in Figure 2.

There are 2 missing values for Land and 2 missing values for Total. This were replaced with the mean for for the column as seen in Figure 3.

There are no duplicated rows as seen in Figure 4

When checking for outliers for the numerical independent variables, most were seen in the Acreage column, Full Bath, Land and Fireplce as seen in Figure 5.

Using Interquartile method, I removed the outliers for Acreage, Land and FirstArea as seen in Figure 6.

(b) Converting/ Transforming/ Recoding Data

I used Label encoding for two distinct values (binomial) for ‘Garage’ and One hot encoding for three or more distinct values (polynomial) for ‘Height’, ‘Exterior’, ‘ Fuel’.

I converted the categorical variable ‘Height’ to numerical value and resulted that:

* 0 was attributed to 1Story
* 1 was attributed to BiLevel and 2Stories
* 2 was attributed to 2Storatk and SplitLev
* 3 was attributed to 1.5Story
* 4 was attributed to 1Stryatk, as seen in Figure 7 (Appendix)

**Dividing the Data into Test and Train**

The dataset was divided 33% for testing and 67% for training as seen in Figure 8 (Appendix).

I normalized the data as in Figure 9.

## Analysis and Results

The Linear Regression Model is calculated in Figure 10.

Graphical user interface, text, application, email

Description automatically generated

Figure 10

The linear ecuation

Total = 160503.89 + 65517 \* Land – 37329 \* Acreage + 4082 \* RtAcge + 5986 \* FirstArea + 1342 \* Rooms + 7945 \* Bedrooms + 7672 \* FullBath + 7263 \* HalfBath – 1081 \* Fireplce

c) RtAcge was created in Excel in the ‘House Price’ dataset

### Data Modelling and Visualisation

**Multicollinearity and variance inflation factors (VIF)**

Land is strong, positive, linear correlated with the target variable. Rooms is uncorrelated.

Chart, scatter chart

Description automatically generated

Figure 11

The heatmap shows no multicollinearity and that the variable ‘Land’ is strong, positive, linear correlated with the ‘Total’ variable.

Chart

Description automatically generated

Figure 12

## Evaluation

### Limitations and Challenges

## Conclusion

I evaluated each model, comment the challenges I had and concluded, for each model, not as separate subchapters.

**MODEL 1:** Regression Equation: Total = -47223.32 + 3.73 \* Land + e

Graphical user interface, text, application, email

Description automatically generated

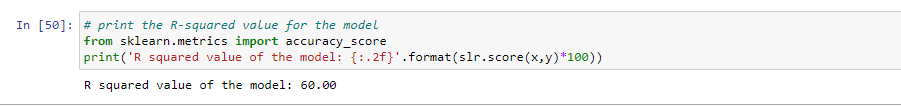


Figure 13

Conclusion: The model predicted the Total by -41408.82 in that market.

With Single Linear Regression the model is:

Regression Equation: Total = -47223.32 + 3.73 \* Land + e

R squared = 60%

Land Variable has explained the variation of Total house prices by 60%. There is 40% that remains unexplained using SLR.

Graphical user interface, text, application

Description automatically generated

The mean absolute error is 30557.

**Multiple Linear Regression**

(d) I will create Multiple Linear Regression (MLR) including RtAcge variable.

**MODEL 2**

4 X’s

Equation: Total = β0 + (β1 \* Land) + (β2 \* Acreage) + (β3 \* FirstArea) + (β4 \* Rooms) + e

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Figure 14

Conclusion: The model predicted the Total of -731386.73 in that market.

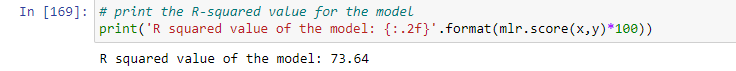


Figure 15

Conclusion: 73.64% of the data fit the multiple regression model.

Land, Acreage, Rooms and FirstArea variable have explained the variation of Total House Price by 73.64%.

With this model, just 26% remains unexplained and so it is better than MODEL 1 SLR.

Text

Description automatically generated

Figure 16

The mean absolute error is 28201 which is better than 30557(for MODEL 1).

**MODEL 3**

When we have 8 independent variables X1, X2, X3, …, X8

Equation: Total = β0 + (β1 \* Land) + (β2 \* Acreage) + (β3 \* FirstArea) + (β4 \* Rooms) + + (β5 \* Bedrooms) + (β6 \* FullBath) + (β7 \* HalfBath) + (β4 \* Fireplce) + e

Y = Dependent variable / Target variable β0 = Intercept of the regression line β1, β2,..βn = Slope of the regression lime which tells whether the line is increasing or decreasing X1, X2,..Xn = Independent variables / Predictor variables e = Error Equation: Total = β0 + (β1 \* Land) + (β2 \* Acreage) + (β3 \* FirstArea) + (β4 \* Rooms) + e

Graphical user interface, text, application, email

Description automatically generated

Figure 16

Total = β0 + (2.82 \* Land) + (-24141.27 \* Acreage) + (40.06 \* FirstArea) + (1751.09 \* Rooms) + (1751.09 \* Bedrooms) + (1751.09 \* FullBath) + (1751.09 \* HalfBath) + (1751.09 \* Fireplce) + e

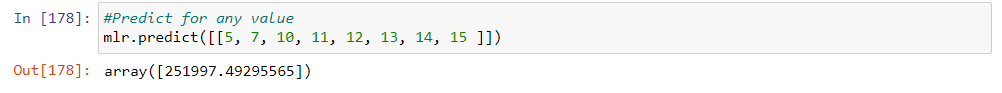


Figure 17

Conclusion: The model predicted the House Price of 251997.4 in that market

Calculating the R squared (coefficient of determination)

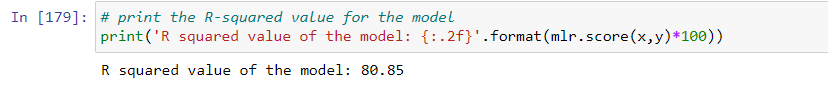


Figure 18

Conclusion: 80.95% of the data fit the multiple regression model. The independent variables ‘Land’, ‘Acreage’, ‘Rooms’, ‘FirstArea’, ‘Bedrooms’, ‘FullBath’, ‘HalfBath’, ‘Fireplce’ have explained the variation of ‘Total’ by 80.95%. There is 19.05% that remains unexplained. This is better than for the previous two models.

Text

Description automatically generated

Figure 19

The absolute mean error 22725 is better than for the previous two models.

P VALUE

Significance Level =Alpha= 5% (95% accuracy and 5% of error)

There are two hypotheses

1. H0 : The variable Land does not predict the Total price
2. Ha: The variable Land does predict the Total price

Chart, scatter chart

Description automatically generated

P value = 0.124 = 12.4% which is higher than 5% so we accept H0 : The variable Land does not predict the Total price

For the Fireplce variable, P value = 0.003 = 0.3% which is less than 5% so we reject H0 : The variable Land does not predict the Total price and accept Ha: The variable Land does predict the Total price

For the Acreage variable P value = 0.274 = 27.4% which is higher than 5% so we accept H0 : The variable Acreage does not predict the Total price.

For the Rooms variable P value = 0.001 = 0.1% which is less than 5% so we reject H0 : The variable Rooms does not predict the Total price and accept Ha: The variable Rooms does predict the Total price (Figure 27 in Appendix)

For the FullBath variable P value = 0.001 = 0.1% which is less than 5% so we reject H0 : The variable FullBath does not predict the Total price and accept Ha: The variable FullBath does predict the Total price (Figure 23 in Appendix)

For the Bedrooms variable P value = 0.524 = 52.4% which is higher than 5% so we accept H0 : The variable Bedrooms does not predict the Total price.( Figure 24 in Appendix)

For the HalfBath variable P value = 0.000 = 0% which is less than 5% so we reject H0 : The variable HalfBath does not predict the Total price and accept Ha: The variable HalfBath does predict the Total price.( Figure 25

In Appendix)

For the RtAcge variable P value = 0.000 = 0% which is less than 5% so we reject H0 : The variable RtAcge does not predict the Total price and accept Ha: The variable RtAcge does predict the Total price. (Figure 26 in Appendix)

(e) Variance Inflation Factors

Which measures the correlation and strength of correlation between the explanatory variables in a regression model. (Figure 28 in Appendix)

The equation resulted is: Total = 35.48 + 4.83Land + 4.46Acreage + 1.18Rooms

Table

Description automatically generated

An increase with 1 unit in the assessed value of the land (pound) will increase the house price by 4.83.

An increase with 1 acre will increase the house price by 4.46.

An increase with 1 room will increase the house price by 1.18.

# Appendix

Text

Description automatically generated with medium confidence

Figure 1

Graphical user interface, application

Description automatically generated

Figure 2

Graphical user interface, text

Description automatically generated

Figure 3

Background pattern

Description automatically generated with medium confidence

Figure 4

A picture containing diagram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Figure 5

Chart, histogram

Description automatically generated

Graphical user interface, text, application

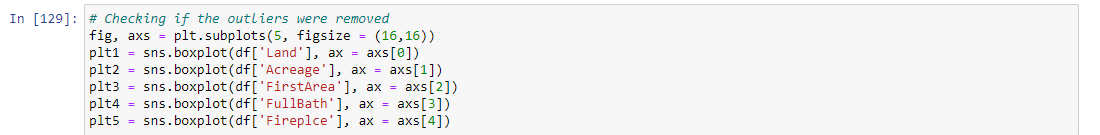
Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated



Chart, box and whisker chart

Description automatically generated

Figure 6

Text

Description automatically generated

Figure 9

Graphical user interface, text

Description automatically generated with medium confidence

Figure 20

Table

Description automatically generated with low confidence

Figure 21

Graphical user interface, text

Description automatically generated with medium confidence

Figure 22

Graphical user interface, text

Description automatically generated

Text

Description automatically generated with medium confidence

A picture containing table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with medium confidence

Chart

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Figure 7 Converting categorical values to numerical

Text

Description automatically generated

Table

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

Figure 8 Dividing data into test and training

Graphical user interface

Description automatically generated with low confidence

Table

Description automatically generated with low confidence

Figure 23

Table

Description automatically generated with medium confidence

Figure 24

Graphical user interface, text

Description automatically generated

Figure 25

Graphical user interface, text

Description automatically generated

Figure 26

Graphical user interface, text

Description automatically generated

Figure 27

Graphical user interface, text, application

Description automatically generated

Figure 28

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, table

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

# **References**

1. Data Science Lectures and Seminars
2. Balan, Shilpa, Tanvi Gawade, and Aakanksha Tasgaonkar. "A Machine Learning Approach for Prediction of Length of Stay for the Kid’s Inpatient Database." *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.
3. T. Gentimis, A. J. Alnaser, A. Durante, K. Cook and R. Steele, "Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data," 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2017, pp. 1194-1201, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.

1. Balan, Shilpa, Tanvi Gawade, and Aakanksha Tasgaonkar. "A Machine Learning Approach for Prediction of Length of Stay for the Kid’s Inpatient Database." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019. [↑](#footnote-ref-1)
2. T. Gentimis, A. J. Alnaser, A. Durante, K. Cook and R. Steele, "Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data," 2017 [↑](#footnote-ref-2)
3. Data Science Seminar, Week 8 [↑](#footnote-ref-3)
4. Data Science Seminar, Week 8 [↑](#footnote-ref-4)
5. Data Science Seminar, Week 6 [↑](#footnote-ref-5)