Length of Stay in Hospital for Diabetic Patients

***Abstract*—Length of stay in hospital is a one of the common issues in health care and also it is contributor in total medical expenditure. Most of the unplanned readmission can be avoided. Due to the complexity of healthcare data and process, decision making task becomes an uneasy objective.The goal of this study is to identify the factors that lead to the duration of hospitalization of patient. With this study, hospital management will easily identify the availability of beds in future. In this study statistical model helps to identify the factors that have a direct impact on duration of hospitalization and data mining techniques can help to predict the duration of hospitalization.For this research, we have taken inspiration from the research article by Beata Strack, Jonathan P. DeShazo, Chris Gennings,Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios and John N.Clore (2014) called Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records [1] . Our research will be based on the part of their future work. For this study, the Novel Intention is to propose a supervised machine learning technique for predicting the duration of hospitalization which is yet to be implemented on the sourced data. Our findings show that a simpler C5.0 model performs as well as more complex models, like SVM. Furthermore, a binary classification works best for this data.**

a) Keywords: Data Mining; Length of Stay; Hospital Readmission, Decision Tree; Statistical Methods; Machine Learning;

I. Introduction

Hospitals can attract large crowds at times. While other businesses are usually able to estimate the traffic of customers, hospitals are confined to other influencing factors. For instance, restaurants are busier during the usual lunch or dinner times and therefore, they can adjust to this. Hospitals do not have this luxury. Hospitalisation may depend on the severity of the illness or the impact on a patients health. Unsurprisingly, they may expect certain diseases to be more prominent at certain times in the year and hence, estimate the influx of cases of such disease. However, there are many diseases that may not be as predictable.

Not all patients have to be hospitalised and many may leave the hospital the same day. However, whether some have to stay a night or more depends on the illness and severity of this illness. One such disease, is diabetes. It is usually classified as Type 1 and Type 2. Depending on how much the disease has progressed without treatment, it can have long term effects on the person's health. Since it is a blood disease, it can cause many issues with organs in a human body. Thus, it is likely that patients may have to stay a couple of nights, if not more, to be treated. For the hospital's sake, it is important for them to know how long a patient is going to be stay in the hospital.

That said, it is likely that a person diagnosed with diabetes may have to revisit the hospital several times throughout his/her life time, because the disease may not be fully curable. Re-admissions leading to extension of the length of stay that are unintentional, may be quite a hassle for a patient and their family. Such sudden admissions may also become messy for the hospital management, which could lead to panic, disorganized treatment, and an extra burden on the staff of the hospital. In this study, identifying the parameters that have major effects on the duration of hospitalization of a patient will not only help the hospital management in better resource management, but it may also allow patients to benefit from saving money from their reduced, unnecessary stay at the hospital. The number of beds in a hospital is heavily dependent on the length of stay of a patient, which plays an important role in deciding the performance and efficiency of the hospital. Therefore, it is important to know the parameters affecting the length of stay of a patient [2].

A. Motivation

Due to the existence of many data classification algorithms, such as decision trees, support vector machines, neural networks, they can be implemented for techniques to estimate the length of stay of a patient [3]. As mentioned, it is beneficial for many parties to have an idea of how long the individual will be staying in the hospital. As will be showcased in greater detail in the related work section, research has been done on the length of stay for several diseases. However, no research has been performed on length of stay for diabetic patients. It is exceedingly difficult to build prediction models that will cover all possible scenarios as there many different perspectives to consider, such as patient history, diagnosis, country, medication severity of the disease etc. This is why, research with focus on patients with one specific disease will perform better than trying to fit all patients and/or diseases in the same basket. The research question that will be answered in this study is **Using classification techniques, how can the duration of hospitalisation for diabetic patients be predicted?**

This paper is outlined in the following way. First, we discuss the related work. In this section we cover the research that has been done so far when it comes to length of stay and why our research is relevant and original, then moving to the statistical models and machine learning techniques applied to the calculate the length of stay. Followed by the proposed methodology for our model, and finally the evaluation and results achieved by applying various models on the dataset.

II. Related Work

Length of stay, otherwise known as simply LoS, is the duration of a hospital visit expressed in days or nights and usually does not refer to short, less than a day visit. Predicting LoS of

a patient in a hospital is nothing new. It has been studied since the 1960s [4]. LoS predictions helps to improve the hospital services. Nowadays, researchers have been using statistical and supervised/unsupervised machine learning algorithms to tackle LoS problem. For this section, we have divided the literature into two parts.

A. Based on Statistical Models

A statistical analysis is an effective way to consider the correlation with independent and dependent variables. A statistical model is widely used in the medical field. In 2014, researchers used a statistical model called multiple linear regression to analyse the coefficients using least-squares models, that used to find the best-fitting line against the dependent variable and independent variables (observed data) [5]. That is achieved by the sum of the squares of the vertical deviations from each data point to the line [5, p.430]. In 2012, a different study also used a statistical model for features extraction. They analysed the administrative claims data that is obtained from BNHI [6]. Initially, they tested continuous variables by one-way analysis of variance, and categorical variables were tested by Fisher exact analysis. Using this statistical method, they identified the relationship between the predictors and selected significant variables (p¡0.5). Their work is applicable to ours, but a reason to use one-way ANOVA is not provided by in the paper [6].

Recently in a study of healthcare in Saudi Arabia, [7]. They used statistical model for healthcare predictive analysis, they have used logistic regression, they provided the reason behind to use logistic regression. Through logistic regression, they know the odds ratio and a confidence interval (CI) for each predictor. They have also mentioned logistic regression constraints, such as in logistic regression, non-linear relationships cannot be used directly and requires converting those variables to linear scale. They also stated that linear regression is not as accurate as data mining techniques [7], which we will find in our study to be true as well; more on this in the methodology and evaluation sections.

In 2009, another study used a statistical method called quantile regression [8]. This statistical study was aimed at emergency services. This means that it only looks at short term treatment. Our study, on the other hand, will consider both short term and long term stays in hospitals. Emergency services are mainly short term. The study also covered all medical issues, which may be argued that it is too broad and not applicable to most diseases.

As mentioned, we found most research papers to cover LoS problems for a particular disease. Compiling too many distinct diseases may not work out, because you would be comparing apples and oranges. They also investigated the relationship between each predictor. Since quantile regression considers looking at only parts or percentiles of the data, the researchers examined LoS at the 10th, 50th and 90th percentiles [8]. Finally, we would also like to discuss a similar research to our question, which was performed in India and published in 2016 [9]. Five years' worth of data was used. The analysis, however, was limited in the sense of predictive modelling. It seems to be more of a data exploration piece than using the data for prediction. Chi-square, independent t-tests and ANOVA are the quantitative methods that were used to analyse the diabetic data for length of time taken for treatment. Throughout their research they found that age and gender played a significant role with dependent variable survival time. According to their research, men face more problems with diabetes compared to women, but women who do admit to a hospital had longer length of stay compared to men [9]. Nonetheless, the significance of this study is questionable, as the dataset only contained 195 individuals.

B. Machine Learning

Machine learning techniques are an effective way to find a pattern in hospitalization data, as research shows. Data mining techniques have widely been used in the field of healthcare. Due to the importance, there is a lot of research that focuses on finding the length of stay of patients in hospitals.

In 2014, researchers used DBSCAN (Density Based Spatial Clustering for Applications with Noise), a clustering method, to predict whether the LoS of a patient in a hospital is greater than one week or not [4]. While unsupervised learning, like clustering, may provide great insight in the data; it can be tricky to know what the model actually learned. In other words, it is possible that the machine finds clusters that are not formed for LoS, but are clustered together for other reasons. These researchers bring forth the reason for choosing DBSCAN. "DBSCAN handles outliers and takes care of clusters of different sizes, shapes and densities" [4, p.2] . They formed three training sets and performed four classification algorithms. They also implemented different prediction models based on logistic regression, neural networks, SVM, and nave bayes [4]. They compared these four and found that SVM provided better accuracy than others in all of three datasets. They have not provided justification for choosing only those four methods and there is no mention of future work [4].

Similarly to [4] as discussed in statistical methods subsection, another study analysed daily patient arrivals and quality of service in the emergency department based on waiting time and length of stay [10]. They have applied both statistical and machine learning techniques. Through multiple regression, they found which variables are correlated to each other and on top of that they used artificial neural networks to predict the daily patient arrivals [10].

Also, predicting the inpatient length of stay, Chinese researchers investigated Zhejiang Huzhoc Central Hospital data to predict the inpatient length of stay [11]. They also used linear regression and ANN techniques to analyse the data. With the help of linear regression, they found the relationship between the variables. Then, they proceeded with ANN. Their model, however, can be argued to be lacking in complexity, as they stuck with an already built model using the Weka tool. Thus, it offers limited interpretation of the results. Using average RMSE, they compared their ANN model with a traditional edit distance approach [11]. They have also used

a derivative of KNN and they got best results when k was around 15. Their work was appreciable as they have clearly provided the evidence of each algorithm that had been used [11].

In Europe, on the other hand, UK hospital data was analysed [12] using a simple Poisson regression modelling technique to analyse datasets of knee operations from January 2007 to December 2011 to predict length of stay in the hospital. Factors found to have a significant effect on length of stay were age, gender, consultant, discharge, destination, deprivation and ethnicity [12]. Those features are to be kept in mind for our own study's feature selection; more on this in the methodology section.

Inspired from the previous attempts to predict length of stay, researchers introduced the length of stay prediction algorithm that is based on the precise support and associate rules measure, to predict the inpatient length of stay of a new incoming patient [13]. They divided the length of stay variable in terms of days, e.g. short time (greater than 3 days), medium (greater than 10 days) and long (greater than 20 days). As will become evident in the methodology section, we have adopted the same approach. They justified why they used that algorithm, but they failed to justify the models accuracy, for example, against another algorithm or base model. Hence, it is not fully clear that with the accuracy the proposed algorithm would predict the LoS well [13].

In the following year, healthcare data was analysed by using decision trees based on correlation ratio [14]. Correlation ratio-based decision trees are complement of information gain-based technique that means when information gains fails, correlation ratio based techniques succeeds. Furthermore, they compared the proposed method to gain ratio-based approach, and they concluded their proposed algorithm was better than gain based [14].

In a more recent study, researchers used machine learning to model the LoS at the time of admission of the patient [15]. As [14], they also used a decision tree classification combined with SVM for that. They analysed 20,321 patient records who were suffering from heart failure. They provided the advantages using cubist tree. Their work is applicable to ours and is also more relevant in terms of contemporary.

On the contrary, an Iranian hospital data was analysed in 2017 [16], where the algorithm used was Bayesian boosting method for identifying the factors affecting LoS. They reported an accuracy of 95 percent. Unlike [13], [16] divided LoS into five classes. However, we believe this to be too many and prefer to see clear distinctions in between the classes. Therefore, we opted for fewer classes in our dependent variable. The study was well put together in terms of intensive reliance of literature. There appears to be a common trend of applying types of decision trees on LoS data, as seen in the literature. Another case where this occurs is in research done by [17]. The authors used a Bayesian network and J48 decision trees to predict the length of stay for stroke patients and compared the performance of the two models. According to their research, Bayesian network provided better accuracy (81%) than J48 decision trees (77%). Unfortunately, they had issues with data quality, such as many fields with incomplete data. Furthermore, the report seems rather simplistic and arguably not up to publishable standards.

Research already being carried out using ANN, DBSCAN, SVM, regression modelling for length of stay, all of them had several shortcomings. So, the researchers worked on diabetic patients' data and predicted the short-term in hospital length of stay [5]. They used a wide variety of supervised learning algorithms, such as SVM, Random Forest, Multi-Task Learning and Multiple Linear Regression in diabetes data. They compared the entire applied model through AUC, ACC, and FS measures, whereby the SVM came out best [5]. While this paper is similar to our proposed work, our study covers the use of data for both short and long-term stay. In their future work, they mentioned to increase the dataset size to include more features and furthermore, to apply decision trees. As our dataset is comparatively large and consists of many variables, it inspired us to attempt this method for our research as well.

III. Methodology

We used the CRISP-DM approach for this project. The reason behind using this approach is the fact it is considered one of the most preferred methodology among professionals [18] and because it can be modified and altered to suit the business requirements. Looking at the CRISP-DM methodology we have come up with our modified version of the CRISP-DM as shown in the following figure:

Figure 1: Methodology

A. Business understanding

Medical data has a huge number of variables and problems in different fields (like cardiology, radiology, pathology, neurology). Our research focused on one area in the medical sector, which is the prediction of the duration of a stay in a hospital. This kind of research is of great help to the medical sector and people in general as hospitals can use such systems to organize their rosters and schedules in a better way and therefore they will be able to help more patients by organizing and speeding up the admission process. At first, we looked at the business scope. It is important to be aware of the overall goal of the analysis. As we are predicting the time spent by a diabetes patient in a hospital, the underlying knowledge of diabetes was researched. As we are not medical professionals, our understanding of the admission process and its length came entirely from our research and from the description of the dataset.

B. Data Understanding

We followed an iterative process between understanding the data and understanding the business, as the two are naturally correlated. For this research, data has been obtained from Hindawi Publishing Corporation, which was delivered to them by Cerner Corp. and the VCU Center for Clinical and Translational Research (CTSA Grant no. UL1TR000058).

Our data has been downloaded from this website.[1]. Data from this website had been used by various hospitals for determining the re-admission rate of the patients in a specific time period. But for our research we omitted the readmission variable and used the data for finding the length of stay of the patients. Information has been gathered from 130 hospitals in the United States over a period of ten years with over fifty attributes. The dataset has information related to the inpatients in a hospital. It has details about the medical history of the patient like the number of procedures the patient has previously gone through, the number of medicines he/she is taking, whether they are taking diabetes medicines or not, sex, race and length of stay in days which we will adopt as a dependent variable. The dataset had a lot of missing values and strange characters that needed to be cleaned. Also, our dependent variables (time in hospital) had imbalanced classes that could Jeopardize the accuracy of our model, and we could fix this imbalance using a method we will talk about in later steps.

[1]Data: http://downloads.hindawi.com/journals/bmri/2014/781670.f1.zip/

C. Data Preparation

The next phase was data preparation. Data preparation is an important phase of any machine learning or data mining project because it is usually the longest stage in every modelling project. Data needs to be cleaned from any impurities like missing values or strange characters.

Cleaning of data is done with the help of R, for eliminating null values, special characters and white spaces. Data is transformed according to the needs of the model as their might be instances where variable type is changed from integer to real etc. We have removed some of the variables like weight, payercode, examide and citoglipton using R and we factorised some other features to suite our business requirement During our data preparation phase, the data needed to be studied to fully understand it which is essential in any model building process.

The variable to be predicted (timeInHospital) had values between 1 and 14 days. After checking the distribution of the data, we discovered the data was highly imbalanced when the number of days was between 7 and 14 days. This is understandable because the majority of people don't spend more than a week in the hospital unless they are extremely sick. The best option was to omit the rows of all these cases which left us with more than 80,000 rows and values ranging between 1 and 6. We divided then these cases into short term, medium term and long term (length of stay). Now, this is based on what we had seen in the literature and therefore, it made sense to us to also follow that approach.

Furthermore, we removed columns from the dataset that had only zero values in them, or the ones that carried the patients ID. Also, some of the columns had values that made no sense like the weight column which had some zero values therefore the column needed to be taken out. As our data set consists of over 100,000 records with many distinct diagnoses that may be recorded, we made the decision to remove the type of diagnosis feature. These features are labelled by using the ICD-9 codes. There are over 1,000 different codes, each corresponding to a different diagnosis. The model performance will be hindered by such feature and complexity, as it would try to distinguish between all these different classes. This kind of confusion was avoided by simply removing this feature.

D. Modelling and Evaluation

In this phase we divide our dataset into training and testing data for the development of our model. We worked with many different algorithms, ranging from C5.0 to SVM. We initially had our data divided into short term, medium term and long

term LoS. Now, this is based on what we had seen in the literature and therefore, it made sense to us to also follow that approach. Using this multi-classification approach of the dependent variable, we ran many different algorithms. Again, the choice of algorithms is from what we had researched in the literature. We chose the C5.0 algorithm because it is the industry standard for producing decision tree [19]. Literature [14], [17], [5] have supported our data mining technique. We started with more simple algorithms, like C5.0, to more sophisticated ones, such as random forest and SVM.

However, the problem that we kept facing was that the model would not improve. For example, the accuracy would stay around 48 to 51 %, depending on the algorithm. But, in the end, it would not improve by much; in our mind nothing worth mentioning anyway. We spent a good amount of time on feature selection, because we figured that the issues were probably related to using too many variables, because our variable set is quite large. This meant that we went back and forth between data modelling and data preparation.

On the other hand, we also feared not using enough variables or the right variables for the model to make sense. So, we spent a lot of time on choosing the right features.

We started building our model with exactly 34 variables that remained in the dataset. Recursive Feature Elimination or RFE provided by (caret) package was used to determine the optimal number of features to be used in the model along with (Varimp) function that determines which factors made the biggest contribution. The low contributing columns were omitted to improve the model. It is worth mentioning we depended on accuracy (because our data is fairly balanced now) and Kappa statistic to judge the performance of our model. In order to evaluate the results of the models, we, first of all, applied cross validation, where we set k-fold to 10 times. To interpret the results, we analysed the confusion matrix on the training and test sets. Ultimately, the results from the validation set are most important. Furthermore, we also visualised the results using AUC and ROC plots. For the decision trees, we also plotted the those. However, these plots are not included in the evaluation section.

Reducing the number of variables did not improve the performance and the accuracy along with our kappa score were still low (accuracy 50 % and kappa 20 %) and that is using different multiple classifiers, which made us realize that we were approaching the problem in the wrong way. Therefore, we went back to the drawing board and thought about what other solutions may be possible to help our model. At this stage, we reflected on the work so far and noticed that there may have been some of our own human bias that may have affected the results. As we followed the classification approach of the literature, we built our model by dividing our data into three classes (long, short and medium) which was the wrong approach judging our results, so we had to check for alternatives. Dividing the data into two classes was an option, as binary classifiers should be more comfortable classifying binary variables rather than classifying variables with three classes. In other words, the observation is either 1 or 2; yes or no; long term or short term in our case. A multi classification, on the other hand, makes it more difficult for the model to do the same.

E. Deployment

As this analysis was merely done for research, there is no real deployment now. Nonetheless, producing this report could be considered to be the deployment in this case.

IV. Results and Evaluation

Since most of our data is categorical and thus, not numeric, we did not have to perform Principal Component Analysis (PCA). For PCA, you need numeric data, but as our data does not contain many numeric variables, PCA would not have been effective in our case as we have a large number of variables that could be reduced by such analysis.

The dataset is extremely unbalanced in terms of the dependent variables; length of stay. There are considerably more patients for short term hospitalisation than long term. This makes sense, because you can expect it be rarer for people to stay in a hospital for a longer period than for short. This is because serious conditions are rarer in general anyway.

Once we implemented the binary classification, we saw that the results improved by almost 20 % in accuracy, which is clearly a noticeable improvement. The same procedure as with the multi class models was followed. This means that we attempted to further improve the model by using more complex algorithms, such as SVM, as this is what the literature has informed us [15] [5]. However, we discovered that the results did not improve significantly; they did not change much when we increased the complexity of the models. It turns out that a simple model with C5.0 will produce similar results as a more sophisticated algorithm. Furthermore, this means that there is no need for complicated algorithms and that sometimes a simple algorithm does just adequately.

Caret package offers many evaluation techniques that are very handy when building models. These techniques can help us getting the most of our model. The techniques that we followed along with the results are shown below:

```
> c50Tpred <- predict(c50Train, testing)
> confusionMatrix(c50Tpred, testing$time_in_hospital)
Confusion Matrix and Statistics

          Reference
Prediction long short
     long  2923  1762
     short 3223  7850

               Accuracy : 0.6837
                 95% CI : (0.6763, 0.6909)
    No Information Rate : 0.61
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3054
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4756
            Specificity : 0.8167
         Pos Pred Value : 0.6239
         Neg Pred Value : 0.7089
             Prevalence : 0.3900
         Detection Rate : 0.1855
   Detection Prevalence : 0.2973
      Balanced Accuracy : 0.6461

       'Positive' Class : long
```

Figure 2: Confusion Matrix

The results shown above belong to the confusion matrix that we ran against our testing sample after splitting the data into training and testing with 80/20 ratio. The accuracy here is around 69 % with Kappa statistics around 31 %. Accuracy does not usually tell the whole story about the model. In our case with a dataset that is fairly balanced as mentioned before, the accuracy here is relatively good. Kappa Statistic can take a value between 0 and 1 depending on the model knowing the higher the better. It is the level of agreement with the power of prediction of the model. In this model, Kappa is not strong enough (above 40 % to be considered good) but it is a big improvement from other models we tried where it was somewhere around 20 %. Sensitivity in our case is around 48 % and specificity is around 81 % and these are two pointers of how the model classified the true positive and false negative respectively. Due to our sensitivity being on the lower end, this means that the model predicts more false negatives. So, for example, it may predict that a patient will stay for short term but turns out to be long term.

```
#Tuning Our model
tuneParams <- trainControl(method = "cv", number = 10, savePredictions = 'final')
c50Train <- train(training[,-3], training$time_in_hospital, method="C5.0", trControl=tuneParams, tuneLength=3)
```

Figure 3: K-Fold-10

We tried changing the parameters in the train and train control functions to get the best results out of our c5.0 and we found that the number of iterations (k=10) gave the best results in terms of accuracy and Kappa statistics. A screenshot of our approach is shown below. The above figure is called

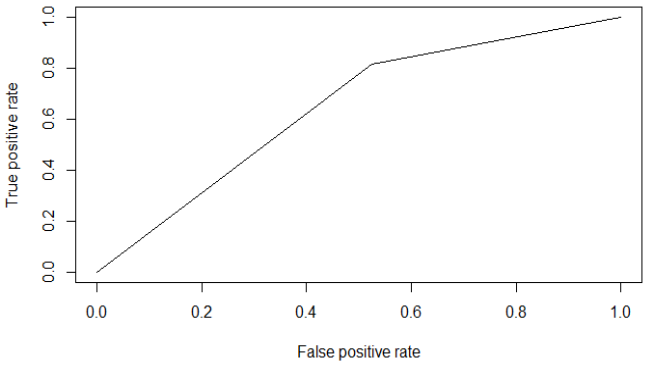

Figure 4: ROC

ROC curve or Receiver Operating Characteristic curve and it shows a graph of true positive rate -vs- false negative one. This curve the trade-off between sensitivity and specificity and usually the higher the graph is and leaning towards the left, the more accurate our model will be. Our AUC which is the area under this curve is around 65 % which is considered a fair value. VarImp function was used to check features that contributed the most to our model. It was used as part of our feature selection process.

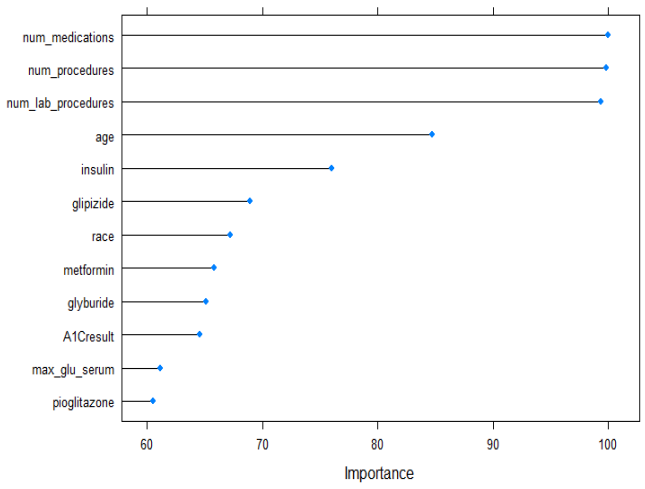

Figure 5: Important Features

V. Conclusion and Future Work

What we can conclude from the data mining piece is that there has been much research that has covered length of stay, but it is important to note that each illness much be treated separately. There are many factors to why someone might have to stay in a hospital depending on the type of severity of the disease. Secondly, it appears that decision trees appear to work best on this kind of data classification. This can be backed up by both the literature and our own findings of the work. Furthermore, we discovered that binary classification outperforms multi classification models. To conclude, over-complication with the multi classification instead of binary throws off the model and it did not work well with our data. Then, what we discovered is that the binary classification is the right approach and sticking with a simpler model turns out best, as there is no need for more complicated algorithms. Such algorithms would take significantly longer to run and produced similar results anyway.

For future work, we propose the following things. We only focused on classifying the LoS in initially three classes and then finally two classes. However, the number of days in the hospital could also be used for a regression model. Future work could rely on regression modelling. Furthermore, we excluded most of the medication variables, because they were deemed not useful due to many missing values. If it were possible to have medication variables with more useful information, it may be relevant.

References

[1] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records," *BioMed research international*, vol. 2014, 2014.

[2] E. Kulinskaya, D. Kornbrot, and H. Gao, "Length of stay as a performance indicator: robust statistical methodo-

logy," *IMA Journal of Management Mathematics*, vol. 16, no. 4, pp. 369–381, 2005.

[3] R. Sharma, S. N. Singh, and S. Khatri, "Medical data mining using different classification and clustering techniques: a critical survey," in *Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on*. IEEE, 2016, pp. 687–691.

[4] V. Panchami and N. Radhika, "A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms," in *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on The*. IEEE, 2014, pp. 207–212.

[5] A. Morton, E. Marzban, G. Giannoulis, A. Patel, R. Aparasu, and I. A. Kakadiaris, "A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014, pp. 428–431.

[6] W.-T. Hung, K.-T. Lee, S.-C. Wang, W.-H. Ho, S.-C. Chang, J.-J. Wang, D.-P. Sun, H.-H. Lee, C.-C. Chiu, and H.-Y. Shi, "Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma and performance comparison with logistic regression model: A nationwide taiwan database study," in *Innovations in Bio-Inspired Computing and Applications (IBICA), 2012 Third International Conference on*. IEEE, 2012, pp. 241–245.

[7] H. Alharthi, "Healthcare predictive analytics: An overview with a focus on saudi arabia," *Journal of infection and public health*, 2018.

[8] R. Ding, M. L. McCarthy, J. Lee, J. S. Desmond, S. L. Zeger, and D. Aronsky, "Predicting emergency department length of stay using quantile regression," in *Management and Service Science, 2009. MASS'09. International Conference on*. IEEE, 2009, pp. 1–4.

[9] V. Sujatha, S. P. Devi, S. V. Kiran, and S. Manivannan, "Bigdata analytics on diabetic retinopathy study (drs) on real-time data set identifying survival time and length of stay," *Procedia Computer Science*, vol. 87, pp. 227–232, 2016.

[10] M. Xu, T. C. Wong, and K. S. Chin, "Modeling daily patient arrivals at emergency department and quantifying the relative importance of contributing variables using artificial neural network," *Decision Support Systems*, vol. 54, no. 3, pp. 1488–1498, 2013.

[11] Z. Huang, J. M. Juarez, H. Duan, and H. Li, "Reprint of length of stay prediction for clinical treatment process using temporal similarity," *Expert Systems with Applications*, vol. 41, no. 2, pp. 274–283, 2014.

[12] E. M. Carter and H. W. Potts, "Predicting length of stay from an electronic patient record system: a primary total knee replacement example," *BMC medical informatics and decision making*, vol. 14, no. 1, p. 26, 2014.

[13] I. Nouaouri, A. Samet, and H. Allaoui, "Evidential data mining for length of stay (los) prediction problem," in *Automation Science and Engineering (CASE), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1415–1420.

[14] S. Roy, S. Mondal, A. Ekbal, and M. S. Desarkar, "Crdt: Correlation ratio based decision tree model for healthcare data mining," in *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 36–43.

[15] L. Turgeman, J. H. May, and R. Sciulli, "Insights from a machine learning model for predicting the hospital length of stay (los) at the time of admission," *Expert Systems with Applications*, vol. 78, pp. 376–385, 2017.

[16] N. Khajehali and S. Alizadeh, "Extract critical factors affecting the length of hospital stay of pneumonia patient by data mining (case study: an iranian hospital)," *Artificial intelligence in medicine*, vol. 83, pp. 2–13, 2017.

[17] A. R. Al Taleb, M. Hoque, A. Hasanat, and M. B. Khan, "Application of data mining techniques to predict length of stay of stroke patients," in *Informatics, Health & Technology (ICIHT), International Conference on*. IEEE, 2017, pp. 1–5.

[18] G. Piatetsky. Kdnuggets. [Online]. Available: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-proje html

[19] D. Caton and D. C. Rusu, "Advanced data mining," *National College of Ireland*, 2018.