Data mining for Prediction of Length of Stay of Cardiovascular Accident Inpatients

Cristiana Silva[1], Daniela Oliveira[2], Hugo Peixoto[2], José Machado*[2], and António Abelha[2]

[1] Department of Information, University of Minho, Braga, Portugal.
[2]Algoritmi Research Center, Department of Information, University of Minho, Braga, Portugal.

{a71665, id7220}@alunos.uminho.pt, {hpeixoto, jmac, abelha}@di.uminho.pt

*Corresponding author

Abstract. The healthcare sector generates large amounts of data on a daily basis. This data holds valuable knowledge that, beyond supporting a wide range of medical and healthcare functions such as clinical decision support, can be used for improving profits and cutting down on wasted overhead. The evaluation and analysis of stored clinical data may lead to the discovery of trends and patterns that can significantly enhance overall understanding of disease progression and clinical management. Data mining techniques aim precisely at the extraction of useful knowledge from raw data. This work describes an implementation of a data mining project approach to predict the hospitalization period of cardiovascular accident patients. This provides an effective tool for the hospital cost containment and management efficiency. The data used for this project contains information about patients hospitalized in Cardiovascular Accident's unit in 2016 for having suffered a stroke. The Weka software was used as the machine learning toolkit.

Keywords: data mining, weka, prediction, cardiovascular accident.

1 Introduction

We live in a world where vast amounts of data are collected daily. This explosively growing, widely available, and gigantic body of data makes our time no longer the "information age" but the "data age" [1]. Hospitals itself are nowadays collecting vast amounts of data related to patient records [2]. All this data holds valuable knowledge that can be used to improve hospital decision making [3][4]. Therefore, analyzing such data in order to extract useful knowledge from it has become an important need. This is possible through powerful and adaptable data mining tools which aim precisely at the extraction of useful knowledge from raw data.

The project of this work primarily consists in the implementation of data mining techniques to predict the hospital Length Of Stay (LOS) of cardiovascular accident (CVA) patients based on indicators that are commonly available at the hospitalization process (e.g., age, gender, risk factors, stroke subtypes). For this purpose, it was developed two predictive models through classification learning techniques.
LOS is used to describe the duration of a single episode of hospitalization, that is, the time between the admission and discharge dates. It is useful to predict a patient's expected LOS or to model LOS in order to determine the factors that affect it [5][6].
This model can be an effective tool for hospitals to forecast the discharge dates of admitted patients with a high level of certainty and therefore improve the scheduling of elective admissions, leading to a reduction in the variance of hospital bed occupancy. These fluctuations prevent the hospital from having an efficient scheduling of resource allocation and management, resulting in short supply for the required resources or in the opposite scenario, that is, the supply being over the demand. The prediction of a patient's LOS can therefore enable more efficient utilization of manpower and facilities in the hospital, resulting in a higher average bed occupancy and, consequently, in cutting down on wasted overhead and improving profits [3][7].
The clinical data used for this matter was obtained from one single hospital and contains information about patients who were hospitalized in CVA's unit in 2016 for having suffered a stroke.
For the purpose of this work, the Waikato Environment for Knowledge Analysis (Weka) was utilized as the machine learning toolkit.

2 Background

Today's data flood has outpaced human's capability to process, analyze, store and understand all the datasets. Powerful and versatile tools are increasingly needed to automatically uncover valuable information from the tremendous amounts of data generated from trillions of connected components (people and devices) and to transform such data into organized knowledge than can help improve quality of life and make the world a better place [1][8]. Many forward-looking companies are using machine learning and data mining tools to analyze their databases for interesting and useful patterns. Products and services are recommended based on our habits [9]. Several banks, using patterns discovered in loan and credit histories, have derived better loan approval and bankruptcy prediction methods [10][11].
The healthcare industry itself generates large amounts of data on a daily basis for various reasons, from simple record keeping to improving patient care with foreknowledge of the subject's own medical history, not to mention the information required for the organization's day-to-day management operations. Each person's data is compared and analyzed alongside thousands of others, highlighting specific threats and issues through patterns that emerge along the process. This enables sophisticated predictive modelling to take place [12][13][14].

2.1 Data mining: the heart of KDD

Knowledge Discovery in Databases (KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [10]. Its main goal is to turn a large collection of data into knowledge through the discovery of interesting

patterns [15] [16]. Given a set of facts (data), a pattern is a collection or class of facts sharing something in common, describing relationships among a subset of that data with some level of certainty. A pattern that is interesting and certain enough, according to user's criteria, is recognized as knowledge [10].

This being said, KDD shares the same ultimate goal as the data mining process, since the second is an essential element of the first. The typical data mining process requires the previous transference of data originally collected in production systems into a data warehouse, data cleaning and consistency check. While KDD consists of the whole process from data preprocessing to the pattern discovery and evaluation, data mining, an essential step in the process of KDD, is the search itself for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and its hospitalization period (See Figure 1) [1] [17] [16] [18].

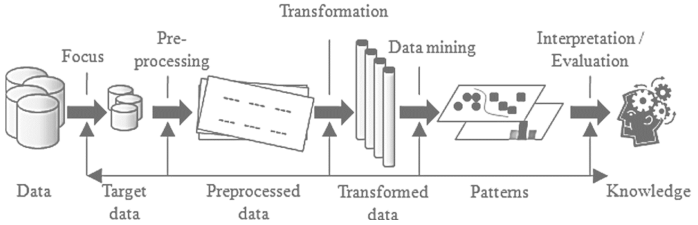

Fig. 1. Step of the KDD process. Adapted from [18].

When it comes to discovering pattern classes in the data mining process, in practice, the two primary goals consist of prediction and description. While the first consists of pattern identification and involves the usage of a certain number of variables/fields in the dataset to predict unknown or future values of other variables of interest, the second consists of class identification (clustering) and is focused on grouping individuals that share certain characteristics together, finding patterns that describe the data to be interpreted by humans [10]. These types of learning are also called supervised and un-supervised learning, respectively [19].

After a predictive model is built and validated, it is deemed able to generalize the knowledge it learned from historical data to predict the future [9]. In this way, for example, it can be used to predict the diagnosis for a certain patient based on existing clinical data from other previous patients with similar features. Models like these implement a classification function, in which the result is a class or a categorical label. Predictive models can also be used to predict numeric or continuous values by implementing a regression function [20].

2.2 Classification

Classification is probably the oldest and most widely-used of all the KDD approaches. In a classification problem, typically there are labeled examples (historical data) which consist of the predictor attributes and the target attribute (dependent variable which value is a class label). The unlabeled examples consist of the predictors attributes only. As mentioned above, classification is learning a function that maps the unlabeled examples into one of several predefined categorical class labels [19] [21].

It is a two-step process consisting of training and testing. The training step is where the classification model is build, by analyzing training data (usually a large portion of the

dataset). A classification model consists of classification rules that are created through a classification algorithm (classifier) that, in turn, entails a set of heuristics and calculations. In the testing step is where the classifier is examined for accuracy or by its ability to classify unknown individuals, by using testing data. Its accuracy depends on the degree to which classifying rules are true, being that classification rules with over 90% accuracy are regarded as solid rules [22].

3 Related Work

The matter of this work has been broadly studied since the advantages of knowing how long patients will stay in a hospital are overall recognized. Thus, there are several studies trying to address this problem by building prediction models. Even though many studies have been developed towards the predictions of LOS related other health problems (e.g., congestive heart failure [3], end stage renal disease [23], burn [24]), or not related to any specific health issue [7] [25], only a few are directly related to the prediction of LOS for stroke patients.

In [5], a group of 330 patients who suffered a first-ever ischemic stroke of this type and were consecutively admitted to a medical center in southern Taiwan were followed, prospectively. The purpose of this study was to identify the major predictors of LOS from the information available at the time of admission. Univariate and multiple regression analysis were used for this purpose. The median LOS was 7 days (mean, 11 days; range, 1 to 122 days). The main explanatory factors for LOS were identified as being the NIHSS score, modified Barthel Index score at admission, small-vessel occlusion stroke, gender and smoking. The main conclusion was that the severity of stroke, as rated by the total score on NIHSS, is an important factor that influences LOS after stroke hospitalization.

A similar study was presented in [26] where a group of 295 first-ever stroke patients were subjects of assessment in order to identify the factors that influence both acute and total LOS. Once again, a multiple regression analysis was performed for this purpose. The mean LOS was 12 days and the mean total was 29 days. Stroke severity measured with NIHSS was identified as being a strong predictor of both acute and total LOS. Also, while prestroke dementia and smoking revealed to have a negative impact in acute LOS, prestroke activities of daily living dependency was identified as a predictor of shorter total LOS.

4 Methods

The available clinical data for this project included 477 cardiovascular accident cases consecutively admitted at a CVA's unit in 2016. The dataset was obtained from a data warehouse in a comma separated value (csv) format and contained several attributes such as patient's gender, age, risk factors (presence or absence of history of hypertension, hypocoagulation, diabetes, atrial fibrillation, previous antiagregation, previous stroke, and smoking), provenance (whether the patient arrived to the hospital on its own, in an ambulance, through another hospital, or though Urgent Patient Orientation Centers), stroke's subtypes, clinical classification, previous and exit ranking (degree of disability), treatments, procedures, complications, and destinations. It also contained the time symptom-door, that is, the time between the moment the

patient has the first symptom and the moment he enters the hospital, time door-neurology and time door-CT, that is, the time between the moment he enters the hospital and the moment he enters the neurology department and the moment that he develops a CT exam, respectively.
Since the purpose of this work was to predict LOS for a certain patient at the CVA unit's time of admission, it was only taken into consideration information available at that moment, that is, factors that can be assessed the moment the patient enters hospitalization. Even though some factors during hospitalization may have a major impact in its duration, the goal of this study is to provide a way for clinic professionals to make an estimation right away, being this information extremely useful for the hospital administration as well as the patient's relatives. In this sense, based on knowledge acquired from the previous research, the predictor variables, that is, the possible explanatory factors for LOS available at the time of admission, were prospectively selected.

Data preprocessing

In the data cleaning process, all the missing values were removed. These unknown values were represented by either the value NULL in some classes or the value 0 in others. Since there was a vast amount of unknown values, especially for the time symptom-door and time door-neurology, numerous cases were eliminated from the dataset. This led to a final number of 211 cases used in this study. Consequently, it was necessary a relevance analysis since some of the classes contained the same value for all the cases or the majority of them. In this sense, all the valueless factors were removed from the dataset such as the different stroke subtypes and some risk factors.
The data transformation process was performed using normalization, which involved scaling all values to make them fall within a small specified range ([0-1]). This was performed at a stage where it wasn't established whether the final purpose would be a classification or regression prediction, for which normalization, is strictly necessary.

Modeling

WEKA software was used for the modeling process, has the capacity to read ".csv" files, change the classes' data type and then store these files in attribute-relation file format (arff) which is Weka's own format. However, in this project, the data was converted to arff format before it was loaded into Weka software, so the various attributes could be easily classified as being real (numeric) or nominal (categorical). Initially, the only numeric attributes were age, time symptom-door, time door-neurology, and LOS.
At the modeling stage, after a few attempts of adopting a regression approach in Weka, which was not giving satisfying results, the target attribute was converted into categorical classes in order to obtain better outcomes. Instead of predicting a numeric value (LOS in days), the goal became the prediction of a class (LOS in period of days). Although LOS had a range of 0 to 116 days, it had a mean value of 13 days. This was taken into consideration for the definition of datasets with different intervals of days. Numerous possibilities were tested by comparing several learning methods such as ZeroR, IBk, and Random Forest. Since accuracy is not the ideal metric to use when working with an imbalanced dataset [27], which is the case, the evaluation was made with four performance measures based on the values of the confusion table: true

positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The mentioned measures are [7] [27]: **accuracy** – correctly classified instances ((TP+TN)/(TP+TN+FP+FN)); **kappa statistic** - accuracy normalized by the imbalance of the classes in the data to see if the result is indeed a true outcome or occurring by chance; **precision** - measure of a classifier exactness (TP/(TP+FP)); **recall/sensitivity** - measure of a classifier completeness (TP/(TP+FN)).
The error rates were not taken into consideration since these are used for numeric prediction rather than classification.
It was also necessary to check which factors had a negative influence in the final result in order to determine what would be the predictor variables for LOS. Each one of the various attributes were removed from the datasets, one at a time, and the best classifiers determined in the previous step were applied each time. If the removal of a certain attribute resulted in better metrics, it wouldn't be retrieved to the dataset.
After the final datasets and the corresponding classifiers had been selected, the sampling methods, more specifically cross-validation, percentage split, and supplied test set, were evaluated.

5 Results and Discussion

In this section, it is presented the results for the practical steps enumerated in the previous section, as well as its respective discussion.
Some of the results for the best datasets mentioned in the previous section where different category labels were applied to the target attribute, are presented in Table 1. In this table, the selected measures are displayed for each dataset and classifier.

Table 1. Results for prediction of LOS with different datasets.

Datasets	Classifier	Accuracy	Kappa statistic	Precision	Sensitivity
A. 4 intervals (0-20-40-60-116 days)	ZeroR	68,72%	0	0,472	0,687
	IBk	91,00%	0,8161	0,912	0,910
	RandomForest	90,05%	0,7851	0,9	0,9
B. 2 intervals (0-7-116 days)	ZeroR	53,55%	0	0,287	0,536
	IBk	81,99%	0,6370	0,82	0,82
	RandomForest	81,52%	0,6241	0,822	0,815
C. 3 intervals (0-7-30-116 days)	ZeroR	46,45%	0	0,216	0,464
	IBk	82,46%	0,7236	0,825	0,825
	RandomForest	80,57%	0,6911	0,807	0,806
D. 4 intervals (0-10-20-60-116 days)	ZeroR	58,29%	0	0,340	0,583
	IBk	81,99%	0,6892	0,825	0,820
	RandomForest	84,83%	0,7278	0,849	0,848
E. 3 intervals (0-20-40-60 days)	ZeroR	70,73%	0	0,5	0,707
	IBk	89,27%	0,7637	0,893	0,893
	RandomForest	89,76%	0,7548	0,899	0,898

By analyzing Table 1, it's clear that the best dataset is A since it presents the best overall values for the present measures. However, in reality, it would be more useful to be able to predict LOS for a period whose limit was shorter than 20 days. Dataset D also presents decent overall results and allows the prediction of LOS for a period limit of 10 days. This being said, datasets A and E were both selected to further assessment.
It's important to mention that, since only a rare number of cases were within the range of 60 to 116 hospitalization days, the dataset E was created to evaluate whether the removal of these cases from the first dataset would improve the results. Since the opposite occurred and, in reality, it's actually useful to know if a patient is expected to stay for that long, dataset E was discarded.
In Table 2, some of the results for the selected datasets when comparing several learning methods are presented.

Table 2. Results for prediction of LOS with different learning methods.

Dataset	**Classifier / Measure**	**IBk**	**KStar**	**J48**	**LMT**	**Random Forest**
A	Accuracy	91,00 %	86,26 %	81,99 %	81,52 %	90,05%
	Kappa statistic	0,8161	0,7137	0,6203	0,6256	0,7851
	Precision	0,912	0,862	0,815	0,823	0,900
	Sensitivity	0,910	0,863	0,820	0,815	0,900
D	Accuracy	81,99 %	81,99 %	66,35 %	76,77 %	84,83%
	Kappa statistic	0,6892	0,6827	0,3822	0,5848	0,7278
	Precision	0,825	0,822	0,661	0,765	0,849
	Sensitivity	0,825	0,820	0,664	0,768	0,848

As it can be seen in the table above, the best classifiers were IBk for dataset A with an accuracy of 91% and Random Forest for dataset D with an accuracy of 84,83%.
The accuracy results for the attributes assessment as mentioned in the previous section, are presented in Table 3. All measurements were taken into consideration.

Table 3. Results for predictions of LOS with different datasets.

	Accuracy			**Accuracy**	
Atributte	**A**	**D**	**Attribute**	**A**	**D**
Gender	87,20%	83,41%	Atrial fibrilation	90,52%	84,36%

Age	88,15%	86,26%	Prev. antiagregation	88,15%	85,78%
Provenance	90,52%	87,68%	Smoking	91,47%	88,15%
Previous ranking	88,15%	85,30%	Previous stroke	91,00%	86,26%
Clinical classif.	88,62%	82,94%	Time symptom-door	90,52%	77,25%
Diabetes	85,78%	86,26%	Time door-neurology	89,43%	84,36%

The final predictor variables for LOS in each dataset were then determined as well as new values for accuracy. The determined attributes for the new dataset A2 were all the factors except for smoking, which resulted in a accuracy of 91,47%. On the other hand, the determined attributes for dataset D2 were all the factor except for age, provenance, and smoking, which resulted in a accuracy of 88,15%.
In table 4, some of the accuracy results for different sampling methods are displayed. By its analyzation, it is visible that 10-fold cross validation was the best sampling method for both datasets, by maintaining the same accuracy results as before.

Table 4. Results for prediction of LOS with different sampling methods.

Accuracy	**Cross-validation**			**Percentage split**		
Dataset	**6-fold**	**8-fold**	**10-fold**	**66%**	**75%**	**80%**
A2	86,73 %	84,36 %	91,47 %	77,78 %	83,02 %	85,71 %
D2	83,89 %	84,36 %	88,15 %	76,39 %	79,25 %	83,33 %

In order to evaluate the supplied test set sampling method and verify if there was overfitting or not, datasets A and D were distributed into two sets: training set (70% of the data) and test set (30% of the data). The best obtained accuracy results were 90,48% with KStar classifier for dataset A2 and 82,54% with Random Forest classifier for dataset E2. Even though accuracy values decreased in a visible way, it is due to the natural data variance. It wasn't a substancial decrease that could raise any concerns.

Table 5. Results of the best models for datasets A2 and E2.

Dataset	**Accuracy**	**Precision**	**Recall**
A2	91,47%	0,915	0,915
E2	88,15%	0,883	0,882

Table 6. Confusion matrix of the best models for datasets A2 and E2.

Dataset A2				**Dataset E2**			
137	4	4	0	**113**	1	9	0
6	**34**	0	0	4	**17**	1	0
2	0	**17**	1	8	0	**51**	1
1	0	0	**5**	1	0	0	**5**

In table 5, the accuracy values tell us that the first model correctly identifies 91,47% of the cases while the second model correctly identifies 88,16% of them. The value for both precision and sensitivity is 0,915 in dataset A2, which means that the first model is 91,5% exact and complete, presenting low false positives and negatives. The same stands for dataset E2, for which the precision and sensitivity values are 0,883 and 0,882. Even though these values are slightly lower, it's still a solid outcome for the model's positive predictive value and true positive rate. In this classification problem, the stastistic, precision and sensitivity values were, in general, proporcionally equivalent to the accuracy values, which facilitated the classifiers and sampling methods assessment and selection for each dataset.

By analyzing the confusion matrix illustrated in Table 6, it can be seen that the majority of the instances were well classified in both models since they are mostly found in the diagonal elements. There was a small number of false positives and negatives. In both datasets, the class which was the least well classified was naturally the first class since it is the most popular class. It presents 9 FP and 8 FN for dataset A2 (interval of days from 0 to 20), and 13 FP and 10 FN for dataset E2 (interval of days from 0 to 10). False positives are slightly more serious in the context of this work, since they mean that the hospital will be expecting a less hospitalization period than it actually is predominated to be, which can result in lack of resources. On the other hand, false negatives mean that the hospital will prepare itself for a longer hospitalization period that will not happen, resulting in wasted overhead.

It was not made an attempt of deleting random instances since the quantity of data was already very reduced. However, it would be a legimitate method for possibly getting better accuracy and overall measures.

From the above-mentioned results, it can be concluded that it isn't truthful to define a certain classifier as the best one for any predictive classification of data because each problem has an adequate classifier that will perform better than others, even though it might not be the case for other datasets. It was also possible to conclude that classifiers of a particular group don't necessarily give similar accuracies. Additionally, it became clear that measures and more importantly, the appropriate classifier, vary according to the dataset being used, specifically the number of attributes, number of instances, and the categorical classes defined for the target attribute. It was also possible to realize that a categorical prediction allowed the obtainment of better results than a numeric one.

Finally, the selected predictor variables didn't corroborate what was theoretically expected from the state of the art research. In [5] and [26], smoking was one of the variables defined as being the explanatory factors for LOS, which was the only variable excluded from both datasets in this study. However, gender and the severity of the stroke were declared as being important factors, which happened in this case also, since previous ranking stands for the degree of disability the patient presents when he initiates hospitalization.

6 Conclusions

This project primarily consisted in the implementation of data mining techniques to predict the hospital Length Of Stay (LOS) of cardiovascular accident (CVA) patients based on indicators that are commonly available at the hospitalization process (e.g., age, gender, risk factors, CVA's type). For this purpose, it was developed two predictive models, through classification learning techniques.
The best learning models were obtained by the IBk and Random Forest methods, which presented high accuracy values for two datasets with different categorical classes (91.47% and 88.16%, respectively) and overall measures such as precision and sensitivity. The number of false positives and negatives was quite acceptable, which is essential to determine how much faith the system or user should put into this model. Since the goal of this predictive model is not directly related to the patient health and more related to the hospital management, false positives or negatives are not so serious as they usually would be in the medical field, specially the second ones. However, the lack of clinic available resources can represent a serious threat for patient health. In this case, either the hospital will prepare itself for a longer hospitalization period that will not happen, resulting in wasted overhead, or it will be expecting a less hospitalization period than it actually is predominated to be, which can result in lack of resources.
These models were obtained through an extensive analysis procedure that revealed the following influential input attributes: gender, previous ranking, clinical classification, diabetes, atrial fibrillation, previous antiagregation, previous stroke, time symptom-door and time door-neurology. For one of the datasets, it was also age and provenance. This showed that these predictor variables are not certain for every problem similar to this.
All the extracted knowledge confirmed that the obtained predictive model is credible and with potential value for supporting decisions of hospital managers. These models can be used by other researchers in order to improve their work, possibly in other fields of study. However, it has to be taken into consideration that each problem needs its individual assessment and intensive analysis of different methods.

Acknowledgments

This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

[1] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*., Morgan Kaufman, 2011.

[2] D. Oliveira, J. Duarte, A. Abelha, and J. Machado, "Improving Nursing Practice through Interoperability and Intelligence," in *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 2017, pp. 194–199.

[3] 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 Syst. Appl., vol. 78, pp. 376–385, Jul. 2017.

[4] M. Miranda, A. Abelha, M. Santos, J. Machado, and J. Neves, "A Group Decision Support System for Staging of Cancer," in *Electronic Healthcare*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 114–121.

[5] K.-C. Chang, M.-C. Tseng, H.-H. Weng, Y.-H. Lin, C.-W. Liou, and T.-Y. Tan, "Prediction of length of stay of first-ever ischemic stroke.," *Stroke*, vol. 33, no. 11, pp. 2670–4, Nov. 2002.

[6] F. Portela, R. Veloso, S. Oliveira, M. Santos, A. Abelha, J. Machado, A. Silva and F. Rua., "Predict hourly patient discharge probability in Intensive Care Units using Data Mining," Indian J. Sci. Technol., vol. 8, no. 32, Nov. 2015.

[7] A. Azari, V. P. Janeja, and A. Mohseni, "Healthcare Data Mining," *Int. J. Knowl. Discov. Bioinforma.*, vol. 3, no. 3, pp. 44–66, Jul. 2012.

[8] W. Fan and A. Bifet, "Mining big data," *ACM SIGKDD Explor. Newsl.*, vol. 14, no. 2, p. 1, Apr. 2013.

[9] A. Guazzelli, "Predicting the future, part 2: Predictive modeling techniques," 2012.

[10] M. Kantardzic, *Data Mining*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011.

[11] D. P. Foster and R. A. Stine, "Variable Selection in Data Mining," *J. Am. Stat. Assoc.*, vol. 99, no. 466, pp. 303–313, Jun. 2004.

[12] B. Marr, "How Big Data Is Changing Healthcare," 2015. .

[13] J. Machado, A. Abelha, J. Neves, and M. Santos, "Ambient intelligence in medicine," in *2006 IEEE Biomedical Circuits and Systems Conference*, 2006, pp. 94–97.

[14] J. Duarte, C. F. Portela, A. Abelha, J. Machado, and M. F. Santos, "Electronic Health Record in Dermatology Service", Communications in Computer and Information Science, 221, Springer, 2011.

[15] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse.," *Proc. of Conf. Am. Med. Informatics Assoc. AMIA Fall Symp.*, vol. 89, no. 10, pp. 101–5, Oct. 1997.

[16] M. Holsheimer and A. Siebes, "Data Mining - The Search for Knowledge in Databases," 1991.

[17] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse.," *Proc. a Conf. Am. Med. Informatics Assoc. AMIA Fall Symp.*, pp. 101–5, 1997.

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework.,” *Int Conf Knowl. Discov. Data Min.*, pp. 82–88, 1996.

[19] A. G. Eapen, “Application of Data mining in Medical Applications,” pp. 1–117, 2004.

[20] M. Chapple, “Defining the Regression Statistical Model,” 2016. .

[21] Oracle, “3 Predictive Data Mining Models.” [Online]. Available: https://docs.oracle.com/cd/B13789_01/datamine.101/b10698/3predict.htm.

[22] I. Yoo *et al.*, “Data Mining in Healthcare and Biomedicine: A Survey of the Literature,” *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, Aug. 2012.

[23] J. Y. Yeh, T. H. Wu, and C. W. Tsao, “Using data mining techniques to predict hospitalization of hemodialysis patients,” *Decis. Support Syst.*, vol. 50, no. 2, pp. 439–448, 2011.

[24] C.-S. Yang, C.-P. Wei, C.-C. Yuan, and J.-Y. Schoung, “Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages,” *Decis. Support Syst.*, vol. 50, no. 1, pp. 325–335, Dec. 2010.

[25] S. Tanuja, D. U. Acharya, and K. R. Shailesh, “Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay,” *J. Pharm. Biomed. Sci.*, vol. 7, no. 7, pp. 1–4, 2011.

[26] Appelros, P. (2007). Prediction of length of stay for stroke patients. Acta Neurologica Scandinavica, 116(1), 15-19.

[27] J. Brownlee, “8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset,” 2015.