DIABETES PATIENTS HOSPITAL READMISSION PREDICTION USING MACHINE LEARNING ALGORITHMS

Sneha Grampurohit[1]

[1] Department of Electronics and Communication
[1] K.L.E Institute of Technology
Hubballi-580027, India.
[1] snehagrampurohit5@gmail.com

Abstract The excess amount of blood glucose in the body leads to a chronic disease called Diabetes. It causes severe damage to the eyes, kidneys, nerves, and other parts of body. Hospital Re-admission is a scenario in which the patient gets readmitted to the hospital after a certain duration of time. Diabetes patient's hospital Re-admission majorly impacts on the health care cost reduction as diabetic patients are more likely to get readmitted than those without diabetes. The proposed work aims to predict the Re-admission of diabetic patients and highlight the factors that lead to Re-admission within 30 days of their discharge considering the database of 10-year administrative patients record using Decision tree and Adaboost Classifiers. With all the preprocessing and feature selection techniques, the proposed approach has obtained an accuracy of 95 percent.

Keywords: Machine learning, Hospital Re-admission, feature engineering, data preprocessing, Hyper parameter tuning.

1. Introduction

When the blood glucose also called the blood sugar gets high in the body, it leads to a chronic disease called Diabetes. In India, as a result of less attention paid to diabetic patients, diabetes is listed as the primary diagnosis by less that 2 percent of hospital discharges annually [14]. The global health expenditure on diabetes is expected to total at least 376 USD billion in 2010 and is estimated to go up to 490 USD in 2030[22]. About 9 percent of the current US population is represented by diabetic patients[15], but they account to approximately 25 percent of hospitalization[14,16,17].The re-admission rates of a diabetic patient within 30 days is found to be around 14.4 percent to 22.7 percent [16,18,19,20,21]. In USA [2012], $124 billion was spent of hospitalization of diabetic patients out of which $25 billion was attributable to 30-day Re-admission with the assumption of 20 percent re-admission rate [16,22].

A scenario in which a patient gets readmitted again to the hospital within a particular duration of time after his/her discharge refers to the hospital re-admission [RA]. Hospital RA can be one of the strong measures to judge the quality of service provided by the hospitals.

The tremendous growth in big data has generated opportunities for greater patient insights that can help the healthcare department to reduce its cost while providing a better health care service. The main objective of the proposed solution is to build a predictive model to identify diabetic patients who are more likely to get readmitted to the hospital within 30 days of their discharge based on the administrative data of the patients collected by the hospital and to highlight the important factors leading to the RA. The dataset considered is a real-world dataset and it represents 10 years of clinical care at 130 US hospitals and integrated delivery network. It has been taken from the Kaggle repository[24]. The dataset consists of 101766 diabetic patients records with 50 features/factors which contribute to diabetic hospital RA. Information was extracted from the database for encounters that satisfied the following criteria. (1) It is an inpatient encounter (a hospital admission). (2) Is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis. (3) Laboratory tests were performed and medications were administrated during the encounter. The factors or the independent variables include Number of lab tests performed during the encounter; Number of diagnoses entered to the system, and so on. The dependent variables have the value "<30" if the patient got re-admitted in less than 30 days of his discharge, ">30" if the patient was re-admitted after 30 days of his discharge and "No" if the hospital does not hold any record of re-admission.

The remaining part of the paper consists of four sections: Section 2 highlights the literature survey done prior to the beginning of the work. Section 3 gives the detailed explanation of the methodology which includes the steps such as: data preprocessing, feature engineering, data balancing, normalization and lastly investigation of machine learning algorithms on the considered dataset. The proposed approach makes use of the data mining algorithm such as Decision Tree classifier, boosting algorithm such as AdaBoost algorithm which was further tuned using GridSearchCV function. Section 4 comprises of results and discussions which gives a comparative performance of the algorithms along with the Risk factors. The evaluation methods considered are: The accuracy rate, Precision rate, recall rate and confusion matrix. Section 5 concludes the presented work.

2. Literature Survey

The previous studies which were analysed have highlighted the risk factors that predict the hospital RA rates of diabetic patients[1-12] . Reena et.al [9] has put forth the key factors leading to RA of diabetes as number of inpatient visits and LOS and has also inspected the cost reduction using 5 different machine learning algorithms. However they have obtained an accuracy up to 87 percent. Hanan et.al [10] has addressed the problems of patient RA and has obtained maximum accuracy of 93 percent using SVM algorithm. Ahmed et.al [12] has proposed an approach using deep learning models to predict the hospital RA of diabetic patients. The work yields an accuracy of 93 percent and does not focus on the key features that correspond to hospital RA. Jiang [1] has explained about the demographic and socio-economic factors that has major impact of the hospital RA rates. More than 52,000 patients' records were examined by Eby [4] to predict the RA risks. Rubin et.al [6] has reported the strategies that can help to overcome the RA problem along with the barriers to reduce RA risk of patients with diabetes. Bhuvan et.al [8] has investigated different machine learning algorithms on public dataset for short term and long-term diabetic RA. This work mainly focuses on the cost analysis due to hospital RA.

In contrast to any previous works, the presented paper highlights risk factors that majorly impact on the RA of diabetic patients and predict whether the patient with the related factors is likely to get readmitted to the hospital within the period of 30 days with a highest accuracy compared to previous works i.e. up to 95 percent.

3. Proposed Methodology

The Proposed Methodology presents the brief explanation of the work from data preprocessing up to the investigation of algorithms.

3.1 Data Exploration

The distribution of data of some of the important features in the raw dataset collected has been presented below [1-6]. Based on the data distributions the pre-processing (data cleaning), feature engineering and data balancing techniques were applied.

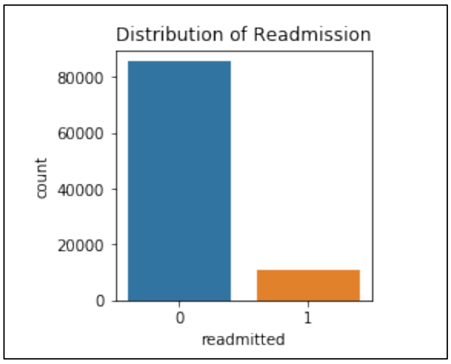

Fig. 1. Distribution of RA

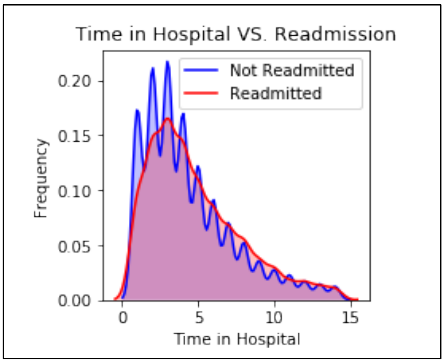

Fig. 2. Time in hospital vs RA

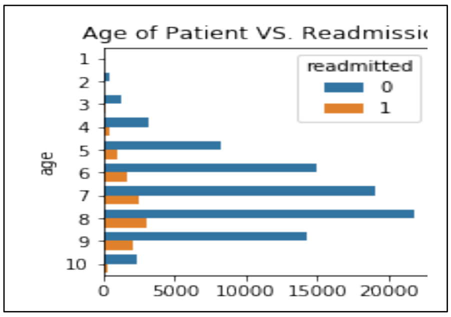

Fig. 3. Age of patient vs RA

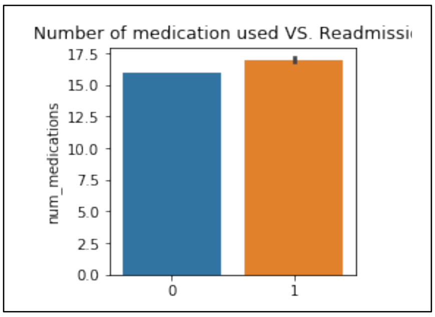

Fig. 4. Num of medications vs RA

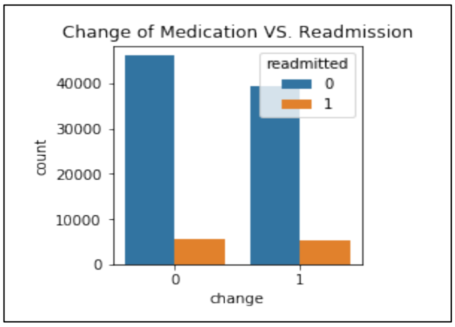

Fig. 5. Change in medication vs RA

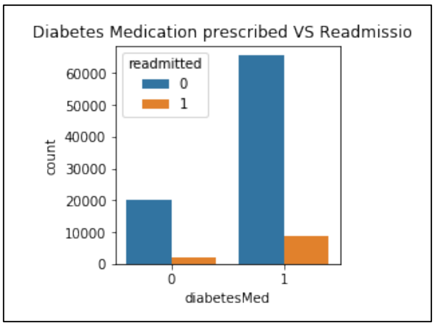

Fig. 6. Diabetes medication prescribed vs RA

3.2 Data Preprocessing and feature engineering

While we are dealing with the real-world data which are often inconsistent, incomplete and noisy, pre-processing them is necessary to convert them to an interpretable form. The pre-processing techniques that were implemented are

- Removal of factors with large missing values.
- In order to measure how much of hospital services a person has utilized in a year, a feature column was created called "Service_utilization" which is the resultant column of addition of two columns namely: Number of inpatient(admissions) and outpatient visits for a particular patient in a year.
- The dataset considered consists of 23 features for 23 drugs which indicate for each of these, whether the hospital had made the change in the medications or not during the entire course of treatment, as some of the research works suggest that change in medication for diabetes leads to low admission rates[13]. Hence, another column "num_of_changes" was added which consists of the count of how many medication changes were made for a particular person during the entire treatment.
- Label encoding was performed on the dataset for columns such as Gender, Categorical encoding was performed on attributes such as tests, age etc, Collapsing of multiple encounters for same patient.
- Categorization of diagnosis: Three diagnosis levels were present in the considered dataset i.e. primary, secondary and addition, each of these diagnoses had 700-900 ICD codes. Henceforth, the above three diagnosis were clubbed into 9 disease categories namely circulatory, respiratory, digestive, diabetes, injury, etc. as clubbed in [5].
- The confidence internals cannot be reliably calculated if the data is not normally distributed, hence moment-based measures such as skewness and kurtosis is used on data to investigate how much the data is deviated from the normality. Further, log transformations has been applied to normalize the data.
- The number of non-re-admitted was found to be very large compared to the number of re-admitted. Hence, the model will not be able to effectively learn th e decision boundary, in case of imbalanced data To prevent this, the SMOTE(Synthetic Minority Oversampling Technique) technique has been applied[11].

3.3 Decision Tree Classifier

Decision tree classifiers are well known for their classification techniques they possess in character recognition, image classification etc. They can efficiently solve the classification problems due to their high adaptability and computationally effective features. One of the most important features of decision tree is the capability of capturing descriptive decision-making knowledge from the supplied data. Decision Tree classifiers comprises of different decision rules or learning models. There are different learning models, one of the most notable models is the CART model, which is a binary recursive practioning technique [23]. It involves constructing a suitable tree by selecting the input variables and split points on those variables based on the algorithm used. The selection of the specific split or cut point or the input variable to use is decided based on a greedy algorithm we use to minimize the cost function. Gini index has been used as the cost function to evaluate the splits in the dataset. It gives an idea of how good a split is, by how mixed the classes are in the considered two groups (<30 days and >30 days) created by the split.

The target variable is the "re-admission" feature whose values:

- 0→ the patient may not get re-admitted to the hospital within 30 days of his discharge or his record of re-admission is not found.
- 1→ the patient is more likely to get re-admitted to the hospital within 30 days of his discharge.

The Gini index for each feature can be calculated by the below formula:

$$G = \left\{ \left(1 - \left((h_{10})^2 + (h_{11})^2\right)\right) * \left(\frac{n_{h1}}{n}\right) \right\} + \left\{ \left(1 - \left((h_{20})^2 + (h_{21})^2\right)\right) * \left(\frac{n_{h2}}{n}\right) \right\} \quad \textbf{(1)}$$

Where

h_{10} , h_{11} = Proportion of instances in group1 of class '0' and '1'.

h_{20}, h_{21} = Proportion of instances in group2 of class '0' and '1'.

n_{h1}, n_{h2} =Total number of instances in group1 and group 2.

n = Total number of instances we are trying to group from parent node.

3.4 Feature importance

Feature importance is checking how relevant each feature is or how much the feature contributes towards the final output result. It is calculated as the decrease or amount of reduction in node impurity weighted by the probability of reaching that particular node. The node probability can be calculated by the ratio of number of instances that reach the node to the total number of instances. The greater the value, the more important the feature will be.

For each decision tree, the nodes importance is calculated using Gini importance. The importance of each feature in a decision tree with help of node probability is calculated as:

$$Ni_k = G_k P_k - \left(G_{left(k)} * P_{left(k)} + G_{right(k)} * \ P_{right(k)}\right) \quad \textbf{(2)}$$

Where:

Ni_k = The importance of feature k.

G_k= the impurity measure or value of parent node k

P_k= weighted number of instances reaching node k.
$G_{left(k)}$, $G_{right(k)}$ = impurity at left node and right node
$P_{left(k)}$, $P_{right(k)}$ = Number of samples at left and right node

3.5 AdaBoost classifier

A series of low performing week classifiers are combined with the aim to create an improved classifier. This technique is called as ensemble learning. Under ensemble techniques, boosting algorithms are one of the kinds. Boosting works in a sequential manner and doesn't involve bootstrap sampling. Instead, every tree is fitted on a modified version of an original dataset and lastly summed up to build a strong classifier.
Step1: Initialize the sample weights. Every datapoint is initialized with the weight equal to 1/N where N=Total number of instances in the dataset.
Step2: For each feature in the considered dataset, a decision tree(criterion=Gini) is built with a depth 1. Further, the predictions made by each tree is compared with the actual labels in the training set. The feature and the corresponding tree that has performed the best i.e. the one with the smallest incorrect predictions in classifying the training instances becomes the next tree in the forest.
Step3: The significance (Sig) of that tree is calculated as:

$$\text{Sig}=\frac{1}{2}log\left[\frac{1-total\_error}{total\_error}\right] \quad \textbf{(3)}$$

Where total_error = sum of the weights of the incorrectly classified instances
Step4: Update the instance weights: While updating the data weights the main focus is on the datapoints that were incorrectly classified. Hence, the weights for the correctly classified instances will be decreased and the weights for the incorrect classified datapoints will be increased.

- New_instance_weight [incorrectly classified] $= instance_{weight}\ x\ e^{Sig}$
- New_instance_weight [correctly classified] $= instance_{weight}\ x\ e^{-Sig}$

Step 5: Since the instances that were incorrectly classified possess higher weights compared to correctly classified instances, the likelihood that the random number falling into the latter category is greater. Therefore, the new instance or datapoint will have the tendency to contain several copies of the instances that were incorrectly classified by the previous trees. Hence, moving back to the step where the predictions made by each decision trees evaluated, the decision tree with the highest score will have correctly classified the instances which were incorrectly classified previously.
Step6: The steps from 2 to 5 are repeated until n iterations (as n_estimators)
Step7: Use the forest of decision trees to predict the new dataset apart from the training set. The AdaBoost model makes predictions on new instances by making each Decision tree in the forest classify the new instance. Then, the trees are split into groups according to their decisions. For each group, the significance value of every decision tree is added in the group. The final classification made by the AdaBoost classifier as a whole is determined by the group with the largest significance value.

3.6 Feature importance

AdaBoost’s feature importance is derived from the base classifier used in it, as decision tree classifier with “Gini” criterion has been used as the base classifier, the AdaBoost feature importance is calculated by the average feature importance provided by each decision tree used i.e.:

$$ANi_k = \frac{\sum_{k \in all\, trees} normalized\ Ni_k}{T} \quad \textbf{(4)}$$

ANi_k= The importance of feature k calculated from all the decision trees in the AdaBoost model
$normalized\ Ni_k$ = the normalized feature importance for k in tree i.
T= total number of decision trees used.

3.7 Hyperparameter tuning of AdaBoost with GridSearchCV

One of the primary objectives and challenges in the ML process is improving the performance score based on data patterns and observed evidence. To achieve the objective, almost all the ML algorithms have a specific set of parameters that need to estimate from a dataset which will maximize the performance score. These parameters are the knobs that we need to adjust to different values to find the optimal combination of the parameters that gives us the best accuracy. Scikit learn library’s GridsearchCV function facilitate an automatic and reproducible approach for hyperparameter tuning.

For a given model, a set of parameter values can be defined which we would like to try. Further, using the GridCV function of scikit learn, models are built for all possible combinations of a preset list of values of hyperparameter provided by us and the best combination is chosen on the cross-validation score. For the presented work, the GridSearch parameters used are: N_estimators: 100,200,500 with Learning rate: 0.2, 0.5 and 1.0

4. Results and Discussions

4.1 Analysis of classifiers

Figure 7 presents the comparative results of classifiers based on the evaluation measures considered. Table 1 gives the accuracy, precision and recall values of the algorithm based on their performance on test dataset. It can be observed that **AdaBoost classifier algorithm tuned using GridCV function** has given the best results compared to the results without tuning and Decision Tree classifier.

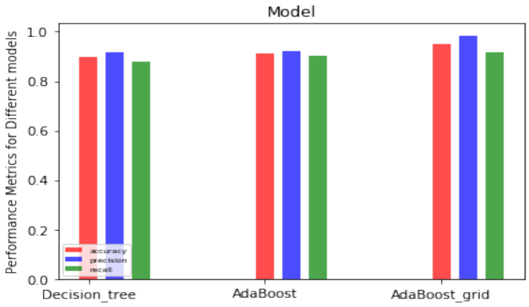

Fig. 7. Comparative results of classifiers

Table 1. Accuracy, precision and recall rates of Investigated Algorithms

Algorithm	Accuracy	Precision	Recall
Decision tree classifier	0.90	0.91	0.88
AdaBoost classifier	0.91	0.92	0.90
AdaBoost classifier with hyperparameter tuning	0.95	0.98	0.91

4.2 Identifying the critical factors

The feature importance function of sklearn library has been used to identify the top 10 risk factors impacting the diabetic RA considered by each algorithm. As it can be observed from figure 8 and 9 risk factors such as Number of diagnosis performed during the encounter, age, number of outpatients, insulin provided, gender, metformin, discharge deposition Id are some of the major risk factors leading to diabetic RA's.

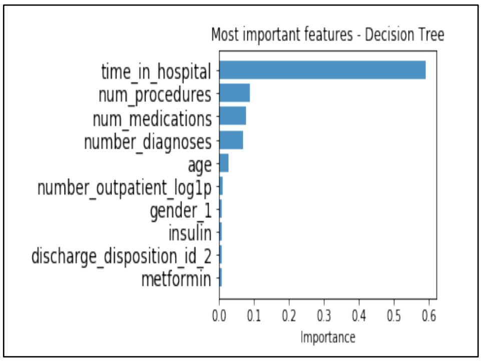

Fig. 8. Risk factors pertaining to decision tree Classifier

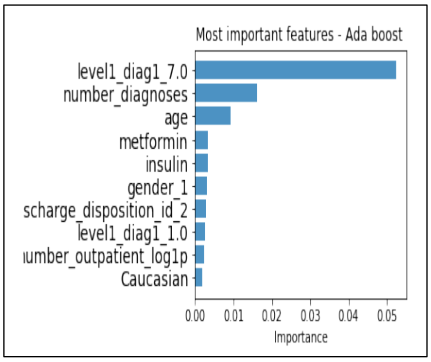

Fig. 9. Risk factors pertaining to AdaBoost classifier.

The results are in sync with some of the research works such as Bhuvan et.al[8] who has also discovered that that discharge disposition id is one of the major risk factor. Reena et.al has identified age and gender as some of the risk factors[9] . Many authors have also used decision tree classifier including feature selection and data pre-processing techniques in their research works[9][12].

The proposed work brings up some of the prominent rules. For instance, if a diabetic patient with the above mentioned risk factors is predicted to get readmitted within 30 days of his discharge by the proposed model, then the physician can provide special services and attention to such patients which instead saves a lot of lives and money. The physician based on the prediction can also relate the ruleset statistically and can have the knowledge of what kind of diabetic patients are more likely to get readmitted within few days of their discharge. As mentioned in Section 1, lack of attention towards diabetes RA can lead to great healthcare losses. Hence the proposed work also contributes towards preventing these losses and improvising the quality of healthcare systems.

The model does not suggest that the non-diabetic patients should be ignored, instead it urges a special attention to diabetic patients. Hence the model is conservative in nature and is safe to be used in health care institutions to assist the physicians to make informed decisions considering the risk factors.

5. Conclusion and future work

Prediction of diabetic RA can widely help in saving huge expenses and lives. The proposed work deals with the real-world dataset i.e. it comprises of 10 years clinical records, upon which two classifiers were investigated. The dataset divided the diabetic patients into two classes of risk group of RAs (yes or no). The objective of the proposed work was to identify the diabetic patients who are more likely to get readmitted within 30 days of their discharge and to recognize the factors leading towards the RA. As seen, AdaBoost classifier tuned with GridCV has given the most optimal results w.r.t. all evaluation measures. Some of the risk factors are Number of diagnosis performed during the encounter, age, number of outpatients etc. The algorithms were able to build a rule set based on mining the hidden patterns between the risk factors. The proposed work can efficiently contribute to reduce the healthcare costs and improve the hospital services. The research focusses only on the diabetic patients Hence, a much-detailed study about hospital RA of other chronic diseases like dengue, heart diseases etc can help reduce RA rates, develop the hospital standards and also bring a reform in the important groups of patients.

References

1. Jiang HJ, Stryer D, Friedman B, Andrews R. Multiple hospitalizations for patients with diabetes. Diabetes Care. 2003;26(5):1421–6.

2. Kim H, Ross JS, Melkus GD, Zhao Z, Boockvar K. Scheduled and unscheduled hospital RAs among diabetes patients. Am J Manag Care. 2010;16(10):760.
3. Dungan KM. The effect of diabetes on hospital RAs. J Diabetes Sci Technol. 2012;6(5):1045–52.
4. Eby E, Hardwick C, Yu M, Gelwicks S, Deschamps K, Xie J, et al. Predictors of 30 day hospital RA in patients with type 2 diabetes: a retrospective, case-control, database study. Curr Med es Opin. 2015;31(1):107–14.
5. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN. Impact of HbA1c measurement on hospital RA rates: analysis of 70,000 clinical database patient records. BioMed research international. 2014;2014:1–11
6. Rubin DJ, Donnell-Jackson K, Jhingan R, Golden SH, Paranjape A. Early RA among patients with diabetes: a qualitative assessment of contributing factors. J Diabetes Complications. 2014;28(6): 869–73.
7. Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnakumar B. Predicting RA risk with institution-specific prediction models. Artificial Intelligence Med. 2015;65(2):89–96.
8. Bhuvan MS, Kumar A, Zafar A, Kishore V. Identifying diabetic patients with high risk of RA. arXiv preprint arXiv: 1602.04257. 2016.
9. Reena Duggal et.al. Predictive risk modelling for early hospital RA of patients with diabetes in India. Springer: DOI 10.1007/s13410-016-0511
10. Hanan et.al.Hospital RA of Patients with Diabetes. IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 4, 2019
11. Nitesh Chawla et,al SMOTE: Synthetic minority oversampling technique, arXiv, 2002.
12. Ahmad Hammoudeh et,al. Predicting Hospital RA among Diabetics using Deep Learning.EICN,2018.
13. Daniel .J.Rubin Hospital RA of Patients with Diabetes.Springer,2018.
14. HCUP Nationwide Inpatient Sample (NIS). Agency for healthcare research and quality (AHRQ). 2014. http://hcupnet.ahrq.gov/ HCUPnet.jsp (2012). Accessed 15 June 2020.
15. Centers for Disease Control and Prevention. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Atlanta: U.S. Department of Health and Human Services; 2014.
16. HCUP Nationwide Inpatient Sample (NIS). Agency for healthcare research and quality (AHRQ); 2011. http://hcupnet.ahrq.gov/ HCUPnet.jsp. Accessed 11 March 2020.
17. Umpierrez GE, Isaacs SD, et.al. Hyperglycemia: an independent marker of in- hospital mortality in patients with undiagnosed diabetes. J Clin Endocrinol Metab. 2002;87(3):978–82.
18. Robbins JM, Webb DA. Diagnosing diabetes and preventing rehospitalizations: the urban diabetes study. Med Care. 2006;44(3):292–6.
19. Bennett KJ, Probst JC, Vyavaharkar M, Glover SH. Lower re-hospitalization rates among rural Medicare beneficiaries with diabetes. J Rural Health. 2012;28(3):227–34.
20. Chen JY, Ma Q, Chen H, Yermilov I. New bundled world: quality of care and RA in diabetes patients. J Diabetes Sci Technol. 2012;6(3):563–71.
21. Rubin D, McDonnell M, Nelson D, Zhao H, Golden SH. Predicting hospital RA risk with a novel tool: the diabetes early read- mission risk index (DERRI). 1508-P. American Diabetes Association 74th Scientific Sessions, 06/2014. San Francisco, CA; 2014. Describes a novel tool to predict RA risk of individual patients with diabetes prior to discharge.
22. ADA: Economic Costs of Diabetes in the U.S. in 2012. Diabetes Care. 2013.
23. Everitt, Brian. (2005). Classification and Regression Trees. 10.1002/0470013192.bsa753.
24. Centre for clinical and translational Research, Virginia Commonwealth University. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008. Retrieved on August 2019.