# Diabetes Patients Early Readmission Prediction

Peter Mačinec and František Šefčík

Faculty of Informatics and Information Technologies,
Slovak University of Technology, Bratislava

**Abstract.** Nowadays, more and more patients suffer from still incurable diabetes disease. Every wrong chosen treatment for patients can harm their health and lead to early readmission that costs more money. Therefore, there is a demand for predicting the readmission of patients to increase quality of health care and also to reduce costs. However, standard methods for identifying the patients with risk of readmission perform poorly (e.g. LACE index). With growing number of patients with diabetes, there is a need for methods that can automatically and more accurate predict the readmission. In this paper, we provide method based on machine learning for predicting early readmission of patient. The results of data analysis already showed that there is a potential of using data-driven approach for this problem.

**Keywords:** diabetes · early readmission prediction · machine learning · data analysis.

## 1 Motivation

Diabetes is a wide spread chronic disease that is related to irregular blood glucose levels caused by problems with insulin. The number of people with diabetes has increased enormously in recent years and costs for health care with each hospital admission are rising simultaneously. The methods of diabetes treatment of patient have high impact on mortality and morbidity. Wrong treatment can endanger patient's health and may lead to early readmission.

A hospital readmission is when a patient who is discharged from the hospital, gets re-admitted again within certain period of time. By predicting readmission, more attention may be given to treatment of patients with high probability of readmission and so increase the quality of care during hospitalization. Because there is no cure for the diabetes yet[1] and diabetic patients can be readmitted in the future, an *early readmission prediction* can help mostly when it comes to selection of best treatment for the patient.

Data of patients clinical encounters are being collected naturally with healthcare systems, thus data-driven approach seem to be appropriate for this problem. Machine learning algorithms used for early readmission prediction provide ability to process the data of a lot of patients and may help to find hidden dependencies in the data to outperform basic methods (e.g. LACE index).

[1] https://www.diabetes.org.uk/diabetes-the-basics/is-there-a-cure

The task of *early readmission prediction* can be represented as binary classification problem into two classes - patient was early readmitted or not. The term *early readmission* is very relative. In our case, we define it as patient being readmitted in less than one month (30 days).

## 2 Related works

Task of predicting diabetic patients readmission is substantive from two points of view - health of patients and saving money because of readmission. Because data of patients are usually available from health records, much research has been done also in data mining area.

Majority of previous works in this area are based on traditional machine learning workflow, including data analysis, data preprocessing and modeling. However, the task defined can vary across researchers - some of them are trying to predict early patients readmission [2, 4, 3], and then there is also work where authors are trying to predict short and long term readmissions [1]. Different but interesting approach that was also the subject of research is to predict readission across age groups [3].

In the previous works, much effort has been spent on data preprocessing. Basic well-known steps of preprocessing like filtering useless columns, normalization, outliers removal and one-hot encoding were usually performed [4, 3, 2, 1]. Also heavy ensemble models have been tried instead of detailed preprocessing [3], but comparison of the achieved results with other works showed that the proper preprocessing always has its place in machine learning tasks.

Because the classes (patient was early readmitted or not) are naturally imbalanced, class balancing was usually performed. Oversampling was preferred over undersampling, and usually SMOTE (Synthetic Minority Over-sampling Technique) was used [4, 2]. From the analyzed works we can deduce that training machine learning algorithms after data oversampling leads to better results.

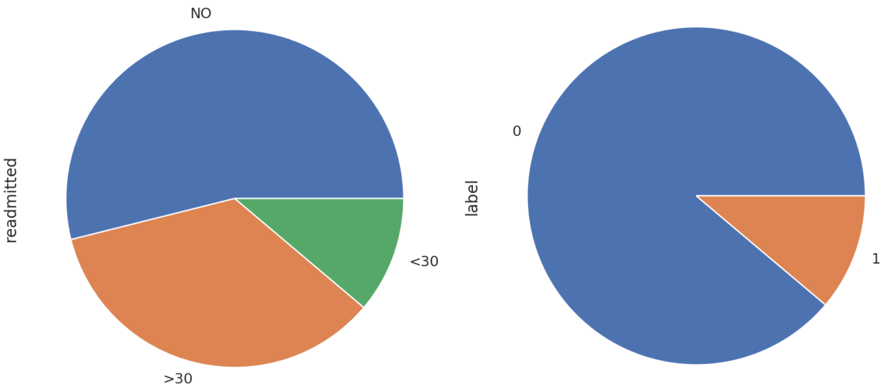
When it comes to modeling, many different approaches and methods were used. Mostly used were tree-based algorithms, like random forest [4, 1], decision tree [4] or xgboost [4]. Also, other algorithms were tried, e.g. logistic regression [4], Adaboost [4, 1], naive Bayes [1]. Neural networks are not so popular in this area, but for example multi-layer perceptron [1] or convolutional neural networks [2] were used. In general, tree-based models (and mostly random forest) alongside with convolutional neural networks achieved the best results - accuracy about 90%-94% after data balancing.

## 3 Dataset

To evaluate our method of diabetes patients early readmission prediction, we have chosen real world dataset of diabetes patients clinical encounters [5]. The dataset was created from Health Facts database with data collected during 10 years (1999-2008) across 130 hospitals in United States. From this large-scale database with millions of records, final dataset with 101 766 records of patients encounters was derived using 5 criteria defined by authors (e.g. it is diabetic and inpatient encounter, or that laboratory tests were performed during the encounter, etc.).

The dataset contains 50 attributes (features), both numerical and categorical. Attributes are of various types - demographics of the patient (like *race*, *gender*, etc.), diagnoses, diabetic medications or number of visits in the preceding year. All of these attributes have been chosen by clinical experts to be potentially associated with patient's diabetic condition.

As shown on Fig. 1a, 3 classes describing whether and when was patient readmitted are provided for prediction. According to the authors of the dataset [5] and our task definition, we transformed the problem to binary classification of *early readmission*. In this scenario, records of patients readmitted in less than 30 days are considered to be in positive class (early readmitted), othervise not early readmitted. Final distribution of the classes after task adjustment is shown on Fig. 1b. As we can see from the figure, classes are highly imbalanced. We provide also detailed analysis of the dataset and its attributes[2].



(a) Original classes distribution

(b) Binary-converted classes distribution

Fig. 1: Distribution of predicted classes - original (whether and when was patient readmitted) and after converting to binary form (only whether was patient readmitted or not).

## 4 Data preprocessing

During analysis of our dataset, some problems with data have been identified. To ensure reproducibility of preprocessing, data were preprocessed via *pipeline*. Basic preprocessing steps based on data analysis results are included in pipeline:

[2] https://github.com/pmacinec/diabetes-patients-readmissions-prediction

1. drop redundant columns (or those with too low diversity, e.g. containing one major value),
2. merge too small classes in categorical attributes into one *other* value,
3. fill missing values (most-frequent value for categorical attributes and median for numerical ones),
4. feature engineering (new features identified in data analysis),
5. map ordinal features values into numbers,
6. one-hot encoding (nominal attributes into numbers).

We have checked numerical attributes carefully to see whether normalization and outliers removal should be performed. According to analysis, there are no such differences in measures so normalization is not needed. We have checked also outliers, but we have not found any extreme value that should be removed.

Some new features have been explored - total number of visits, number of medicaments changes and total number of medicaments used. Also, diagnoses codes had to be mapped into diagnoses categories (according to original paper [5]).

## 5 Methods

Our method to help to solve the task of early readmission prediction is based on basic machine learning workflow, using variety aspects of data analysis and machine learning.

From data analysis, we have found out that most of the attributes are categorical. Depending on concrete attribute type (nominal or ordinal) correct encoding has to be chosen.

There are also attributes with a lot of missing values (4 attributes only), where we will have to choose whether fill the missing values and use the attribute for prediction or not. In other attributes, corrupted values have to be fixed (e.g. unknown gender). In numerical attributes, we do not find any extreme values or outliers (all values looks natural).

Next challenging problem will be to handle imbalanced classes. Minor class has just about 10% of the values. According to chosen algorithm and preprocessing, we will use either undersampling or oversampling.

## 6 Evaluation

DESCRIBE EVALUATION TECHNIQUE

DESCRIBE PRELIMINARY EXPERIMENTS

## 7 Conclusion

ARE IMBALANCED DATA MAIN PROBLEM?

SMALL CLASS IS HARD TO BE LEARNED?

## References

1. Bhuvan, M.S., Kumar, A., Zafar, A., Kishore, V.: Identifying diabetic patients with high risk of readmission. ArXiv **abs/1602.04257** (2016)
2. Hammoudeh, A., Al-Naymat, G., Ghannam, I., Obied, N.: Predicting hospital readmission among diabetics using deep learning. Procedia Computer Science **141**, 484 – 489 (2018). https://doi.org/10.1016/j.procs.2018.10.138, the 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops
3. Mingle, D.: Predicting diabetic readmission rates: Moving beyond hba1c. Current Trends in Biomedical Engineering & Biosciences **7** (01 2017). https://doi.org/10.19080/CTBEB.2017.07.555715
4. Sharma, A., Agrawal, P., Madaan, V., Goyal, S.: Prediction on diabetes patient's hospital readmission rates. In: Proceedings of the Third International Conference on Advanced Informatics for Computing Research. ICAICR '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3339311.3339349
5. Strack, B., Deshazo, J., Gennings, C., Olmo Ortiz, J.L., Ventura, S., Cios, K., Clore, J.: Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed research international **2014** (04 2014). https://doi.org/10.1155/2014/781670