Skoltech

Machine Learning Project Report

Who returns to hospital?

Authors:

Marina Dudina

Evgeny Marshakov

Leyla Mirvakhabova

Skoltech

Skolkovo Institute of Science and Technology

In collaboration with
Massachusetts Institute of Technology

Contents

1 Introduction

1.1 Data Description

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The table with comprehensive description of all features you may find in Appendix.

1.2 Data Preprocessing

The database contains incomplete and redundant information.

First of all, we removed the data with high percentage of missing, such as weight (97% of missing values), medical specialty (53% of missing values) and payer code (52% of missing values).

After that, we considered three data partitions: we took all features, only medical features (for example, medications, diagnosis, number of lab procedures etcetera) and only features, which correspond to medication treatment.

Separately, we also excluded non-unique patient results, since in this case the data would not be independent. After that also performed data partitioning. Moreover, we excluded the patients, who died during the treatment or were transferred to hospice. The number of samples reduced from 101766 to 69973. We use notation "Preprocessed data" for this case.

The results for data corresponding to the medication features only were remarkably worse than for other partitions, hence hereinafter we do not consider this case.

1.3 Problem Statement

In our project we want to predict the probability of patient readmission based on the known features. We reduce this problem to three different classification problems

- **Multiclass classification**. We have three labels: the patient was readmitted within 30 days, was readmitted in more that 30 days and was not readmitted $[< 30, > 30, \text{None}]$.
- **Binary classification**.
  - The first class is readmission within 30 days, the second one – in more than 30 days and no readmission
  - The first class is readmission within 30 days and in more than 30 days, the second one – no readmission.

2 Solution

2.1 Data visualization

First of all, we plotted the data to find out whether there are clusters, which correspond to different classes. Initially, we colored the data with respect to three readmission labels: if patient returns within 30 days, if patient returns in more than 30 days and does not return at all. We used several methods in order to perform dimensionality reduction to two- or three- dimensional spaces for plotting the data in the corresponding dimension. The methods we adopted for the purposes for our study were Principal Component Analysis (PCA), Kernel Principal Component Analysis (kernel PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).

Moreover, we also colored the data with respect to different classes (for example, race, gender, age, medications etcetera) to ascertain whether there exist patterns corresponding to these labels.

2.1.1 Principal component analysis

Principal component analysis (PCA) is a preprocessing technique, which is used to emphasize variation and bring out strong patterns in a dataset. It is designed in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under

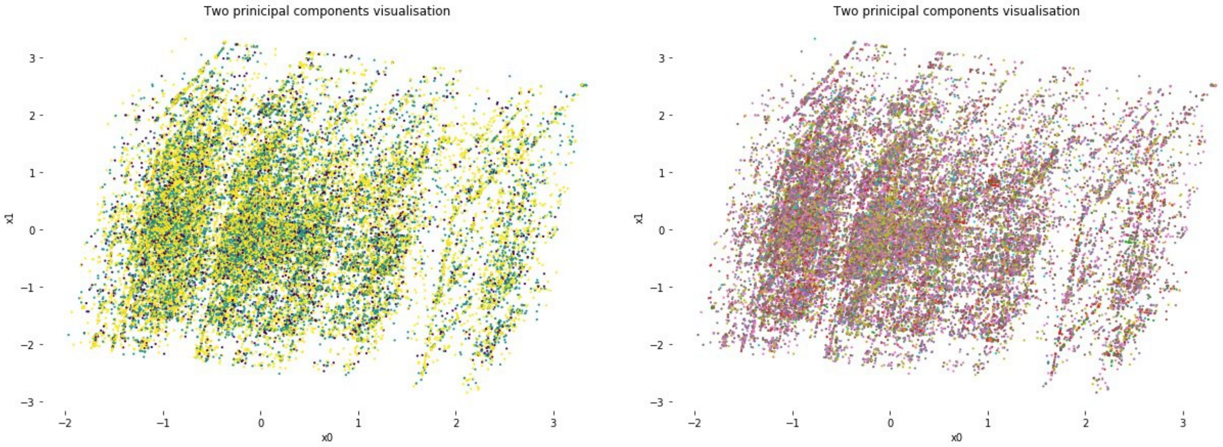

Figure 1: Transformed data via PCA to two-dimensional space, colored with respect to the readmission labels (on the left) and with respect to age labels (on the right).

the constraint that it is orthogonal to the preceding components [1], [2].

The visualizations obtained by PCA method are presented on the Figure 1.

2.1.2 Kernel Principal Component Analysis

Kernel Principal Component Analysis (kernel PCA) is based on the use of integral operator kernel functions in order to compute principal components in high dimensional feature spaces related to input space by some nonlinear map [3].

In our project we chose the following kernel functions: cosine and polynomial. The first kernel, cosine kernel, measures the cosine angle distance and it is more robust to outliers and has better performance in many cases. The second one is polynomial kernel with degree 3. The choice of the polynomial degree was based on the research in the article on Nonlinear PCA with Polynomial Kernels [4]. In this article it was mentioned that high-order polynomials in kernel PCA tend to be very sensitive to outliers.

The visualizations obtained by Kernel PCA method for different kernels are presented on the Figure 2 and Figure 3 respectively.

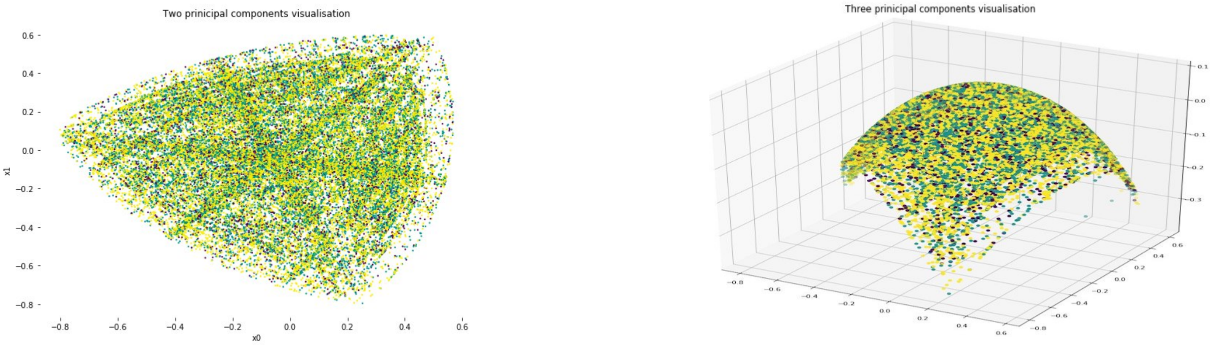

Figure 2: Transformed data via Cosine Kernel PCA to two-dimensional space and three-dimensional spaces respectively with cosine kernel, colored with respect to the readmission labels.

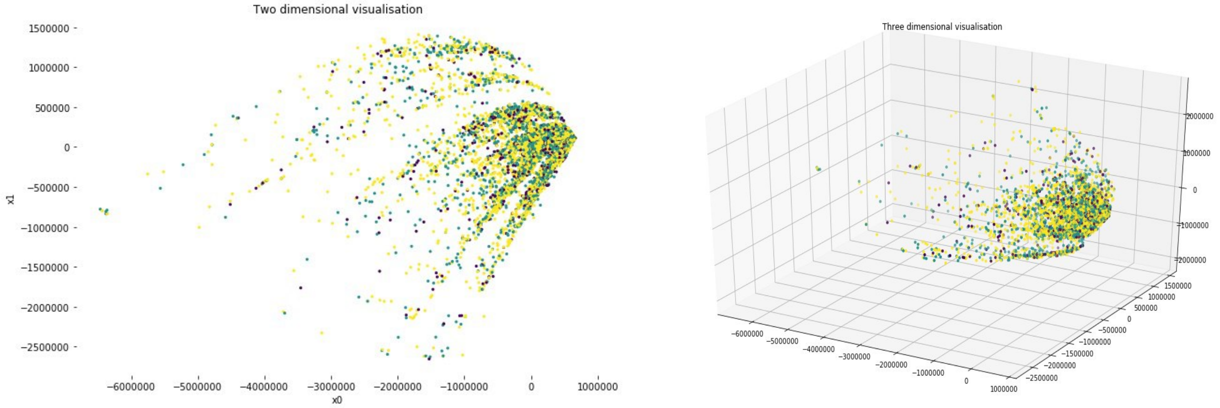

Figure 3: Transformed data via Polynomial Kernel PCA to two-dimensional space and three-dimensional spaces respectively with cosine kernel, colored with respect to the readmission labels.

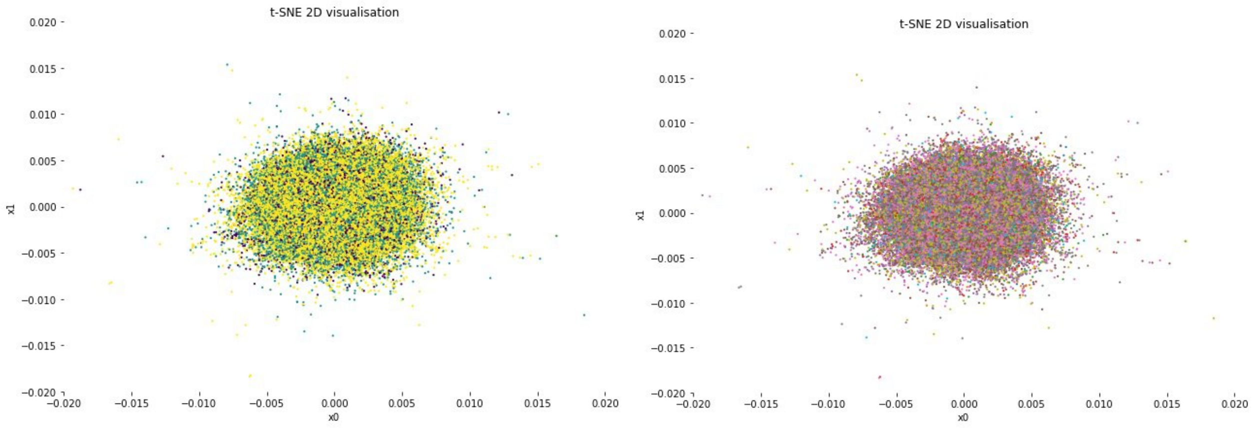

Figure 4: Transformed data via t-SNE to two-dimensional space, colored with respect to the readmission labels and age labels respectively.

2.1.3 t-Distributed Stochastic Neighbor Embedding

TSNE is a nonlinear dimensionality reduction technique that embeds high-dimensional data into a two- or three- dimensional space. It is designed in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points.

The visualizations obtained by t-SNE method for two-dimensional spaces are presented on the Figure 4.

Conslusion on visualization

Based on the plots obtained, one can make a conclusion that there is no occurrence of clusters. So we built our subsequent study on the assumption that data cannot be clustered in a natural way.

2.2 Classification methods

As it was mentioned in Introduction, we implemented two approaches to solve our problem.

The first one, **multiclass classification approach**, is to consider the classification problem with labels $[< 30, > 30, \text{None}]$, which correspond to the case of patient readmission within 30 days, patient readmission in more than 30 days and no patient readmission at all.

The second, **binary classification approach**, is to consider two cases.

The first case has two classes:

- for the first class we take readmission within 30 days
- for the second – readmission in more than 30 days and no patient readmission at all.

For the second case we have the following two classes:

- for the first class we take readmission within 30 days and readmission in more than 30 days
- no patient readmission at all.

2.2.1 Multiclass classification

We used the following methods:

- Random Forest Classifier
- One Vs Rest Classifier, estimator – Logistic Regression
- Output Code Classifier, estimator – Logistic Regression

2.2.1.1 Random Forest Classifier

Random Forest is a composition of decision trees. Each decision tree acts as a weak classifier and combining the responses from multiple decision trees leads to a strong classifier. Each decision tree is trained independently and determines the class of an input by evaluating a series of greedily learned binary questions [8]. The optimal number of estimators of random forest, which is used in our study, was found by grid search on a grid of possible values that vary from 5 to 200 with step length of 5.

We carried out the feature importance analysis for the whole preprocessed data and for preprocessed data with medical features only. The corresponding bar charts are presented on the Figure 5.

The five most important features among all features are the following: admission source id, the number of emergency visits, the number inpatient, the first diagnosis and the number of lab procedures.

The five most important features among medical features are the following: number of lab procedures, the first diagnosis, the second diagnosis, the third diagnosis and the number of medications.

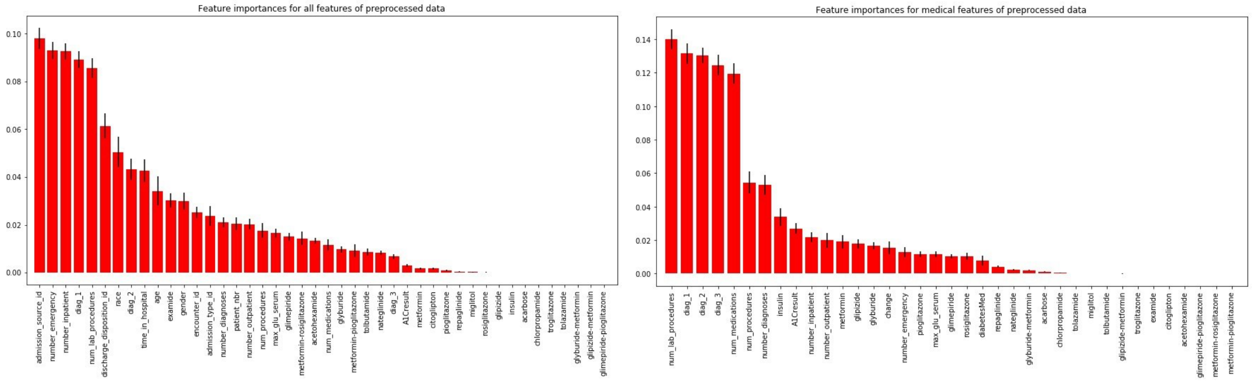

Figure 5: Feature importance bar chart corresponding to the all features and medical features.

2.2.1.2 One vs Rest Classifier

One vs Rest classifier strategy consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes. We chose Logistic Regression as estimator parameter [5].

2.2.1.3 Error-Correcting Output Code Classifier

Output-code classifier is based on the strategies, which consists in representing each class with a binary code [7]. We chose the code length equal 50 and Logistic Regression as estimator parameter.

The ROC-AUC scores obtained on the different preprocessed data for multilabel classification you may find in the Table 1.

Table 1: ROC-AUC scores for multiclass classification

	All data		Preprocessed data	
	All features	Medical features	All features	Medical features
Random Forest Classifier	0.59	0.57	0.61	0.60
OneVsRestClassifier	0.57	0.57	0.60	0.60
OutputCodeClassifier	0.57	0.57	0.60	0.60

2.2.2 Binary classification

We used the following methods

- AdaBoost, estimator - Logistic Regression
- Multi-layer Perception Classifier
- Naive Bayes Classifier
- Linear Discriminant Analysis

2.2.2.1 AdaBoost

Adaboost constructs a strong classifier by sequentially combining a set of weak classifiers. At first iteration, a single classifier is learnt to minimize the classification error. At each consequent iteration, a new classifier is learnt which seeks to minimize the error of the classifier composed of the set of classifiers learnt until the previous iteration. In all our experiments, decision trees were used as weak classifiers [8].

2.2.2.2 Multi-layer Perception Classifier

Multi-layer Perception Classifier consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one [8]. We used the logistic activation function.

2.2.2.3 Naive Bayes Classifier

Naive Bayes: Naive Bayes algorithm is a probabilistic model for classification. It assumes that given the class, features are statistically independent of each other [8].

2.2.2.4 Linear Discriminant Analysis

Linear Discriminant Analysis is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule [6].

Below you may find ROC-AUC curves for two cases:

- data corresponding to all features and first binarization
- data corresponding to medical features and second binarization

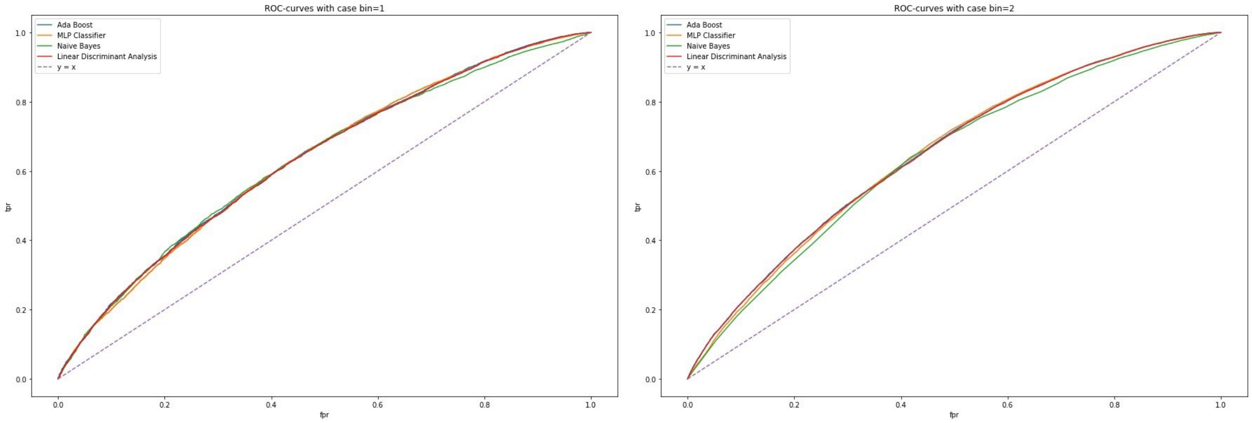

Figure 6: ROC-AUC curves for two cases

The ROC-AUC scores obtained on the different preprocessed data for binary classification you may find in the Table 2. The bin parameter represents the type of the binarization (1 – the first binary problem, 2 – the second one).

Table 2: ROC-AUC scores for binary classification problems

	All data				Preprocessed data			
	All features		Medical features		All features		Medical features	
	bin=1	bin=2	bin=1	bin=2	bin=1	bin=2	bin=1	bin=2
AdaBoost	0.60	0.61	0.59	0.61	0.58	0.59	0.56	0.58
MLPClassifier	0.60	0.61	0.54	0.61	0.50	0.58	0.50	0.57
Naive Bayes	0.54	0.57	0.58	0.60	0.57	0.54	0.55	0.54
LDA	0.60	0.61	0.59	0.61	0.58	0.59	0.56	0.58

3 Comparison to existing studies

The article “Impact of HbA1c Measurement on Hospital Readmission Rates:Analysis of 70,000 Clinical Database Patient Records” [9] studied the same dataset as we were given. The goal of the research was to determine the effect of HbA1 feature on the readmission of a patient. Preprocessing the dataset, the authors determined the features with the least number of non-empty entries and deleted them. Then they identified the unique patients, using patient id feature. There appeared to be 70,000 encounters. Then, the people who died or were discharged to hospices were eliminated. To analyse the preprocessed data, they performed Multivariable logistic regression to fit relationship between measurement of HbA1c and early readmission, within 30 days. Then various groups of features were examined for correlation. They concluded that the probability of readmission depends on the the HbA1 test “quite strongly”, with the measure of HbA1 being strongly correlated with the primary diagnosis. That should be noted, the study did not develop a predictive model for readmission.

Another article [10] based on the same dataset, aimed to identify the major features affecting readmission. In the preprocessing step they reduces the values of readmission to 2 options: readmission before 30 days and other. Then they designed a sample with 1:1 ratio of readmitted and non-readmitted patients. The sample size was reduces to 23,150. They also deleted “irrelevant features”, very unbalanced features and the ones with big percent of empty entries. That resulted in 14 features left. The training to set ratio was 70:30. While developing the predictive model the following methods were used: neural networks with 5-8 hidden layers, Bayesian dependency, Decision Tree, Decision Tree with Quest(Quick Unbiased and Efficient Statistical Tree).Tat should be noted the authors used SAS® Enterprise Miner^{TM} 12.3. Compared to the first study, the authors do not focus on the patients being different in the preprocession step, but rather about the balance in readmission values. The problem is reduced to binary, when in the first case the authors also discuss 3 valued readmission label.

Then there was a group with the same research question and the same dataset [11]. The authors used two-staged predictive model. They firstly predicted the binary label(readmitted or not). On the second they predicted the length of readmission. They used Logistic regression, decision trees, Gradient boost, ensemble, rule induction and neural networks for both stages. Compared to the previous authors, this group used similar pre-procession as in the second study, but they chose two-stage reduction of readmission problem, versus binary.

4 Conclusion

5 References

References

[1] Aishwarya. R et.al, *A Method for Classification Using Machine Learning Technique for Diabetes*, International Journal of Engineering and Technology (IJET) **3** (2013).

[2] *Principal component analysis.* Available at https://en.wikipedia.org/wiki/Principal_component_analysis.

[3] Bernhard Scholkopf et.al, *Kernel Principal Component Analysis.*

[4] Yoonkyung Lee Zhiyu Liang, *Eigen-Analysis of Nonlinear PCA with Polynomial Kernels.*

[5] *One vs rest.* Available at http://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html.

[6] *Output code classifier.* Available at http://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html.

[7] *Output code classifier.* Available at http://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OutputCodeClassifier.html.

[8] *Identifying Diabetic Patients with High Risk of Readmission.* Available at https://www.slidetemplate.org/machine-learning-and-operations-research-to-find-diabetics-at-risk-fo.

[9] Jonathan P. DeShazo Beata Strack Chris Gennings, *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records*, BioMed Research International (2014).

[10] Xiayu Zeng Yi Chun Chien Hong Zhang, *Data Mining for Diabetes Readmission Prediction Team Evolution.*

[11] *Improving the performance of two stage modeling by using Association node of SAS® Enterprise MinerTM 12.3.* Available at https://cepd.okstate.edu/files/11-Improving-the-Performance-of-two-stage-modeling.pdf..

6 Appendix

Table 3: List of features and their descriptions in the initial dataset (Part 1)

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: 0, 10), 10, 20), . . . , 90, 100)	0%
Weight	Numeric	Weight in pounds	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%

Table 4: List of features and their descriptions in the initial dataset (Part 2)

Feature name	Type	Description and values	% missing
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “> 200,” “> 300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “> 8” if the result was greater than 8%, “> 7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Nominal	metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: “¡30” if the patient was readmitted in less than 30 days, “¿30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.	0%