Predicting Important Features for Readmission of Diabetic Patients

Team Members: **Shivam Sharma** **Saurabh Shinde** **Rutwik Patil**

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
1	Rutwik Patil	RSP
2	Saurabh Shinde	SAS
3	Shivam Sharma	SSS

I. Introduction:

Hospital readmissions pose a significant challenge in the healthcare sector, affecting both patient well-being and healthcare expenses. The Centers for Medicare & Medicaid Services (CMS) have launched initiatives aimed at curbing readmissions, underscoring the critical need to elevate the quality of patient care while also reducing healthcare expenditures.

Our initiative focuses on scrutinizing a dataset to unearth pivotal insights related to hospital readmissions among patients with diabetes:

What elements lead to increased readmission rates? We are delving into machine learning techniques to enhance the precision of our predictive capabilities. Through detailed data examination and predictive analytics, our goal is to discover practical recommendations that could refine patient management and the allocation of resources within healthcare facilities.

II. **Executive Summary:**

Our study ventured into the intersection of healthcare analytics and big data, with a specific focus on the predictive modeling of hospital readmissions for diabetic patients using PySpark. We discovered that the imbalance in our dataset significantly affected the performance of our predictive models, as the rarity of readmission events skewed the model's ability to learn effectively. PySpark, while robust in handling large volumes of data, presented limitations in its native capabilities to address this imbalance directly through advanced oversampling techniques.

To counteract this issue, we explored the potential of oversampling methods, which create synthetic samples of the underrepresented class—in this case, readmissions—to better train our models. However, we recognized that while oversampling can improve model performance, it may not fully capture the complexity of patient health dynamics. Therefore, we emphasized the necessity of incorporating domain expertise from healthcare professionals to refine our feature selection and to provide more nuanced insights into the data.

An intriguing finding from our study is the impact of feature selection on model performance. We found that not all features that are statistically significant are equally informative or relevant in a clinical context. This underscores the importance of coupling data-driven approaches with clinical expertise to identify which factors are truly impactful in predicting readmissions.

What makes our study particularly novel is the application of PySpark to healthcare data analytics, a domain where such big data tools have not been traditionally employed. By identifying the limitations within PySpark and proposing a combined approach of oversampling techniques and expert consultation, our research presents a pathway to more accurate, scalable, and clinically relevant predictive models. The significance of our work lies in its potential to transform hospital readmission predictions into a more proactive and patient-centered process, ultimately serving the purpose of enhancing patient care and operational efficiency in healthcare settings.

III. **Data Description :**

Data source : Kaggle, Fairlearn (https://fairlearn.org/)

The dataset comprises 101,766 records, encapsulating a decade of clinical services across 130 hospitals and integrated healthcare networks spanning the Midwest, Northeast, South, and West regions of the United States. The data encompasses a variety of patient demographics including race, gender, age, and weight, along with detailed information on hospital diagnoses and treatments such as the number of lab procedures, medications prescribed, outpatient visits, diagnoses, and medication orders. This dataset is a curated selection from the larger Health fact dataset. It is characterized by its open access, longitudinal and cross-sectional nature, and the comprehensive coverage of attributes (50 in total), making it an excellent resource for our investigation into the factors influencing hospital readmission rates among diabetic patients.

encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	payer_code	medical_specialty
2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	?	Pediatrics-Endocr...
149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	?	?
64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	?	?
500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	?	?
16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	?	?
35754	82637451	Caucasian	Male	[50-60)	?	2	1	2	3	?	?
55842	84259809	Caucasian	Male	[60-70)	?	3	1	2	4	?	?
63768	114882984	Caucasian	Male	[70-80)	?	1	1	7	5	?	?
12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13	?	?
15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12	?	InternalMedicine
28236	89869032	AfricanAmerican	Female	[40-50)	?	1	1	7	9	?	?
36900	77391171	AfricanAmerican	Male	[60-70)	?	2	1	4	7	?	?
40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7	?	Family/GeneralPra...
42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10	?	Family/GeneralPra...
62256	49726791	AfricanAmerican	Female	[60-70)	?	3	1	2	1	?	?
73578	86328819	AfricanAmerican	Male	[60-70)	?	1	3	7	12	?	?
77076	92519352	AfricanAmerican	Male	[50-60)	?	1	1	7	4	?	?
84222	108662661	Caucasian	Female	[50-60)	?	1	1	7	3	?	Cardiology
89682	107389323	AfricanAmerican	Male	[70-80)	?	1	1	7	5	?	?
148530	69422211	?	Male	[70-80)	?	3	6	2	6	?	?

only showing top 20 rows

```
data.shape
```

```
(101766, 50)
```

III. Research Questions:

How can we leverage advanced predictive modeling techniques, specifically within a PySpark environment, to accurately predict hospital readmissions among diabetic patients, and what are the key factors influencing these readmissions?

This overarching question can be broken down into several focused research questions that address different aspects of the problem and the goals of the analysis:

Predictive Modeling and Algorithm Comparison:

1. What are the comparative predictive performances of Logistic Regression and Random Forest classifiers within a PySpark environment for hospital readmission among diabetic patients?
2. How do the different algorithms compare in terms of AUC, and what does this indicate about their respective abilities to discriminate between patients at high and low risk for readmission?

Feature Importance and Disease Severity:

1. What features are the strongest predictors of hospital readmission in diabetic patients?
2. How does the duration of hospital stay (time_in_hospital) and the number of inpatient stays correlate with the likelihood of readmission, and what can this tell us about patient care and hospital practices?

PySpark's Capabilities and Limitations:

1. In what ways does PySpark enable the effective handling and analysis of large-scale healthcare data sets when predicting hospital readmissions?
2. What are the limitations of using PySpark for predictive modeling in healthcare analytics, specifically for readmission prediction, and how can these limitations be addressed or mitigated?

Operationalization of Predictive Insights:

1. How can the predictive insights gained from PySpark modeling be operationalized within hospital settings to prepare and prevent high-risk readmissions?
2. What interventions can be derived from the predictive models to assist in care transition planning and readmission prevention strategies?

Integrating Clinical Expertise:

1. How can clinical expertise be integrated into the PySpark modeling process to enhance the predictive accuracy and clinical relevance of the models?
2. What role do clinical experts play in interpreting the predictive outcomes and in the feature engineering process to refine the models further?

Model Deployment and Real-time Prediction:

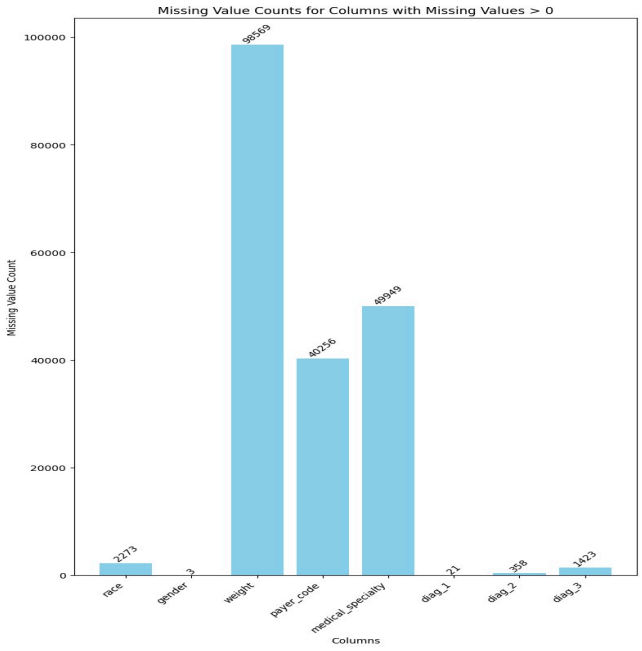
1. How can the developed predictive models be deployed within a real-time hospital information system using PySpark?
2. What are the challenges and considerations in implementing real-time predictive analytics for readmission in a clinical environment?

IV. Methodology:

Prior to performing any analysis, we conducted exploratory analysis to preview the data type, attributes, and overall patterns of the data. We are interested in the class label "Readmitted".

Data Cleaning:

Our dataset contains 101,766 records from 10 years of clinical activities across 130 U.S. healthcare institutions, covering diverse regions. It features comprehensive data on patient demographics, hospital diagnostics, treatments, and more, with a total of 50 attributes. During exploratory data analysis, we identified extensive missing values in several variables. Particularly, we opted to exclude variables like weight, medical specialty, and payer code due to their high percentage of missing data and limited relevance to our research. For variables with a smaller percentage of missing data, such as race, we removed the missing entries while retaining the rest.



Upon initial examination of our dataset, we identified and removed records for patients who died during hospital admission, as they have no chance of readmission (identified by discharge_disposition_id=11). Additionally, we eliminated two variables, citoglipton and examide, due to their uniform value across all entries, rendering them non-contributory to our analysis. Furthermore, we identified two variables, encounter_id and patient_nbr, as irrelevant to the outcome of readmission and opted to remove them from our dataset to streamline our analysis process.

Creation of New Features:

1. patient_service: A composite measure capturing the aggregate of hospital and clinician services utilized by a patient over the preceding year, derived from summing the counts of inpatient stays, emergency room visits, and outpatient appointments.
2. med_change: With the dataset documenting 23 different medications administered during hospital stays, this metric indicates whether any medication underwent a dosage adjustment. By simplifying the original categorizations (No, Up, Down, Steady) to a binary scale, we categorized 'No' and 'Steady' as no change, and 'Up' and 'Down' as indicative of a change.
3. num_med: A new variable to quantify the total variety of medications a patient received during their hospital visit. Given the potential correlation between medication adjustments and the likelihood of readmission, this total count emerges as a significant factor for consideration.

Recoding Existing Variables:

1. Diagnosis Recoding: The dataset includes three diagnostic variables (diag_1, diag_2, diag_3), initially coded using ICD-9 classifications. We recoded these into a more manageable form, with the specifics of this process detailed in Appendix A.
2. Age Recoding: To explore the correlation between age and readmission rates, we converted the original ten age categories into numerical values by assigning the mean age of each category.
3. Readmission Recoding: Focusing on readmission within 30 days, we simplified this outcome into a binary variable: instances of readmission under 30 days were coded as 1, while instances beyond 30 days or without a need for readmission were coded as 0. This adjustment aims to directly address our study's focus on short-term readmissions.
4. Recode other variables:

- For variable “change”, we recoded change into 1 and no change into 0. For gender, we recoded male into 1 and female into 0.
- For diabetes_Med, we recoded yes into 1 and no into 0.

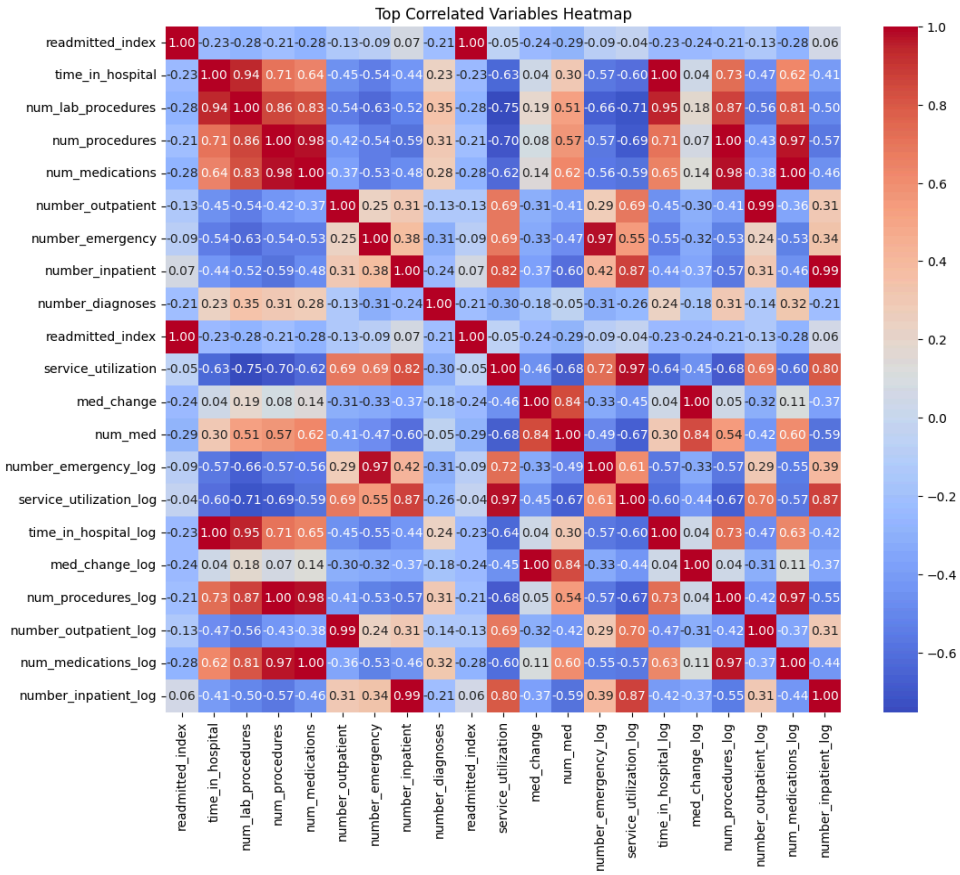
- For race, we recoded the categorical variables into dummy variables: Caucasian-1, African American-2, Hispanic-3, Asian-4, and others-0. For A1Cresult, we recoded >7 and >8 into 1, Norm into 0, and None into 99.
- For max_glu_serum, we used the similar method, namely, we recoded >200 and >300 into 1, Norm into 0, and None into 99.

Feature Engineering:

1) Data Type Conversion: For nominal features, we converted them into object type, in order for the later numerical variables processing.

2) Log Transformation, Standardization, and Correlation: For numeric features- most numericals are highly skewed and have high kurtosis. Therefore, we used log transformation to normalize the numerical variables to make sure numeric variables had a Gaussian-like or normal distribution. Since the numerical variables are not using the same scale so we rescale our data using the standardization method. After all data are standardized, we checked the correlation between the variables using a heat map to find the top correlated variables. There is not too much correlation between the variables.

3) Outliers: For detecting and processing the outliers, we used the coverage rule for normal distribution to deal with outliers. So we removed the outliers.



V. Results and Finding:

This report provides a comprehensive analysis comparing the performance of Logistic Regression and Random Forest models on a healthcare dataset focused on predicting patient readmissions. Additionally, it delves into the importance of various features in influencing the predictions made by the Random Forest model and explores the use of SMOTE for addressing data imbalance, the implications of PySpark's limitations, the effects of oversampling, and the potential of incorporating expert medical opinions for enhancing model efficiency. The findings suggest that both models perform similarly across several key metrics, with slight variations in weighted precision. The feature importance analysis further highlights the significant predictors for patient readmissions.

Model Performance Comparison

Accuracy:

Logistic Regression: 64.12%

Random Forest: 64.06%

Weighted Precision:

Logistic Regression: 83.62%

Random Forest: 83.73%

Weighted Recall:

Logistic Regression: 64.12%

Random Forest: 64.06%

F1 Score:

Logistic Regression: 70.86%

Random Forest: 70.82%

Interpretation and Next Steps

SMOTE (Synthetic Minority Over-sampling Technique): To address the imbalance in the dataset, SMOTE can be utilized for generating synthetic examples of the minority class, thereby improving the model's ability to identify underrepresented classes. However, it's important to monitor for overfitting as oversampling can sometimes lead to models that perform well on training data but poorly on unseen data.

Limitations of PySpark: While PySpark offers scalability and the ability to handle large datasets efficiently, it has limitations in terms of the variety and sophistication of machine learning algorithms available, especially for handling imbalanced data and advanced feature selection methods. These limitations can impact the performance and flexibility of model development and experimentation.

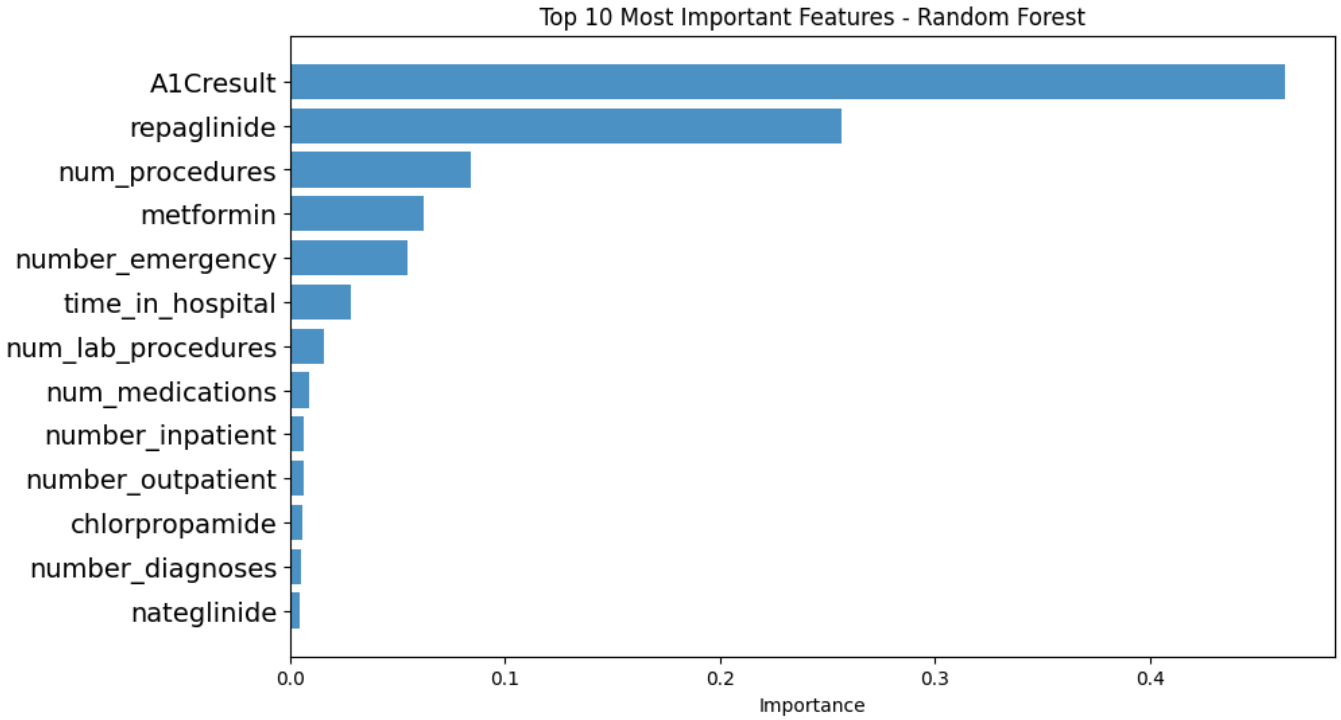
Effects of Oversampling on Models: Oversampling, including techniques like SMOTE, can significantly alter the distribution of classes in the training dataset. While it can improve recall by making the models more sensitive to the minority class, it may also increase the false positive rate, affecting precision. Careful evaluation of model performance metrics is essential to ensure that oversampling benefits outweigh its drawbacks.

Incorporating Expert Medical Insights: Beyond data-driven approaches, consulting with medical experts to identify relevant features can enhance model performance and interpretability. Experts can provide insights into clinically relevant variables not captured by data alone, potentially uncovering new avenues for feature engineering and model improvement.

Feature Importance Analysis (Random Forest)

- **A1C result**: This is the most important feature according to the model. The A1C test result reflects a patient's blood sugar levels over the past 3 months. In real-world settings, this test is crucial for managing diabetes, and its outcome can indicate how well the disease is being controlled. High scores here suggest that tighter control of diabetes could be essential in preventing readmissions.
- **Repaglinide**: This is a medication used to control high blood sugar in people with type 2 diabetes. Its ranking suggests that the use or dosage of repaglinide is a strong predictor of readmission, which might be related to its effectiveness in managing the patient's condition.
- **num_procedures**: The number of procedures a patient undergoes might indicate the severity of their condition. Frequent procedures could lead to a greater likelihood of readmission due to complications or the underlying severity of their health status.
- **Metformin**: Like repaglinide, metformin is another antidiabetic medication. Its importance indicates that the management of diabetes, possibly including the choice and administration of medication, is a key factor in readmissions.
- **number_emergency**: Frequent emergency visits could signal unstable health conditions or insufficient outpatient care, leading to repeated admissions.

- **time_in_hospital**: Longer stays in the hospital might be associated with more serious illnesses or complications that could result in readmission.
- **number_diagnoses**: A higher number of diagnoses could indicate a complex medical condition, which inherently carries a higher risk of readmission.
- **num_lab_procedures**: The number of lab tests can reflect the extent of monitoring or the complexity of a condition, potentially correlating with readmission risks.
- **num_medications**: The number of medications might suggest a complicated drug regimen, which could increase the risk of medication errors or interactions, potentially leading to readmission.
- **number_outpatient**: This may indicate the level of post-hospitalization care needed. Frequent outpatient visits could either reflect ongoing care management or issues that could lead to readmission if not managed properly.
- **chlorpropamide and nateglinide:** Both are less important than other features but still significant. These are also diabetes medications, emphasizing again the importance of diabetes treatment in the context of readmissions.
- **number_inpatient**: Interestingly, this is the least important of the top 10 but still noteworthy. Previous inpatient stays may be related to chronic issues that could predict future readmissions.



VI. Conclusion:

The analysis demonstrates the utility of Logistic Regression and Random Forest models in predicting patient readmissions, highlighting the close performance across key metrics. Incorporating techniques like SMOTE and leveraging expert medical insights for feature selection represent promising avenues for further improving model efficiency and accuracy. Future work should focus on refining the models, exploring advanced data balancing techniques, and integrating clinical expertise to enhance the predictive power and relevance of the models in healthcare settings.

VII. Appendix:

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%
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	For the generic names: 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%

Group name	icd9 codes	Number of encounters	% of encounter	Description
Circulatory	390–459, 785	21,411	30.6%	Diseases of the circulatory system
Respiratory	460–519, 786	9,490	13.6%	Diseases of the respiratory system
Digestive	520–579, 787	6,485	9.3%	Diseases of the digestive system
Diabetes	250.xx	5,747	8.2%	Diabetes mellitus
Injury	800–999	4,697	6.7%	Injury and poisoning
Musculoskeletal	710–739	4,076	5.8%	Diseases of the musculoskeletal system and connective tissue
Genitourinary	580–629, 788	3,435	4.9%	Diseases of the genitourinary system
Neoplasms	140–239	2,536	3.6%	Neoplasms
Other (17.3%)	780, 781, 784, 790–799	2,136	3.1%	Other symptoms, signs, and ill-defined conditions
	240–279, without 250	1,851	2.6%	Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
	680–709, 782	1,846	2.6%	Diseases of the skin and subcutaneous tissue
	001–139	1,683	2.4%	Infectious and parasitic diseases
	290–319	1,544	2.2%	Mental disorders
	E–V	918	1.3%	External causes of injury and supplemental classification
	280–289	652	0.9%	Diseases of the blood and blood-forming organs
	320–359	634	0.9%	Diseases of the nervous system
	630–679	586	0.8%	Complications of pregnancy, childbirth, and the puerperium
	360–389	216	0.3%	Diseases of the sense organs
	740–759	41	0.1%	Congenital anomalies