Diabetes

Karthik Reddy Mathuru

A **hospital Readmission[1]** is an episode when a patient who had been discharged from a hospital is admitted again within a specified time interval. Readmission rates have increasingly been used as an outcome measure in health services research and as a quality benchmark for health systems. Insurance companies and other payers sometimes view unplanned hospital readmission's as wasteful spending .To penalize these readmission rates, measures are to be taken by hospitals. Machine learning models are used to overcome and analyse this problem. "Predicting whether a patient is readmitted to hospital or not" address the problem. A data set that represents 10 years (1999-2008) of clinical care at 130 US hospitals, which has 50 features and 101766 records of patients and hospital outcomes.

Introduction

Tracking patients who are readmitting to a hospital after a hospital stay is one category of data which is used to evaluate the quality of hospital care. Patients with diabetes[2] have high rates of readmission compared with patients without diabetes, according to a pilot study published in Clinical diabetes and Endocrinology. In the first study, the **readmission rate** was 26% in patients with diabetes vs 22% in patients without diabetes "US has the highest prevalence of diabetes among all developed countries across the world"- IDF[3]. Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Glucose comes from the foods you eat. Insulin is a hormone that helps the glucose get into your cells to give them energy.

We have data Set[4] that represents 10 years (1999-2008) of clinical care at 130 US hospitals, which has 50 features and 101766 records of patients and hospital outcomes. Information was extracted from the database for encounters that satisfied the criteria's such as inpatient encounter (a hospital admission), diabetic encounter(one during which any kind of diabetes was entered to the system as a diagnosis), length of stay was at least 1 day and at most 14 days, laboratory tests performed during the encounter, medications were administered during the encounter.

The data also contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

Attribute Information:

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. (International Classification of diseases)	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%
23 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%

Preprocessing:

- When the data is loaded into IDE, the attributes has to be changed to relevant data types such as numeric or nominal, based on attribute information.
- Data standardization is done.

- Class imbalance is handled using SMOTE(Synthetic minority oversampling technique). It uses k- nearest neighbors algorithm and calculate the similarity between records, high similarity records of minority class are over-sampled.
- Omitted three attributes (weight, payer code, medical specialty) which have missing values more than 50 % ,remaining attributes have less than 3% of missing values. Instead of imputing missing values(which may not be perfect), records (with missing values) are deleted(as data is huge).
- Unimportant features such as patient number, encounter ID, examide are dropped(features which have unique records / no variance).
- Feature engineering is done, number of lab procedures (number of lab tests performed during the encounter) and number of procedures (number of procedures performed other than lab tests) are combined which gave total number of procedures.

Approach:

There are three approaches for this data.

Approach – 1 :

Handling the class imbalance using SMOTE, which over samples the minority class.

- **Logistic Regression:**

  Logistic regression is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome.

Classification Report:

	precision	recall	f1-score	support
0	0.18	0.40	0.25	2213
1	0.44	0.27	0.34	6930
2	0.63	0.63	0.63	10468
avg / total	0.51	0.48	0.48	19611

- **Decision Tree:**

  Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

```
Classification Report:

             precision    recall  f1-score   support

          0       0.22      0.29      0.25      2213
          1       0.41      0.23      0.30      6930
          2       0.61      0.74      0.67     10468

avg / total       0.49      0.51      0.49     19611

Accuracy is 0.510784763653
```

- **Random Forest:**

Random forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the target.

```
Classification Report:

             precision    recall  f1-score   support

          0       0.67      0.00      0.00      2213
          1       0.49      0.32      0.39      6930
          2       0.60      0.86      0.70     10468

avg / total       0.57      0.57      0.51     19611

Accuracy is 0.572841772475
```

- **XG Boost:**

XGBoost is an optimized distributed gradient boosting library designed to be highly ***efficient***, ***flexible*** and ***portable***. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting.

```
Classification Report:

             precision    recall  f1-score   support

          0       0.40      0.06      0.10      2213
          1       0.50      0.35      0.41      6930
          2       0.61      0.83      0.70     10468

avg / total       0.55      0.58      0.53     19611

Accuracy is 0.576309214217
```

- **Majority voting:**

  Majority voting is an ensemble approach which combines different models (mode of the classes for classification and mean prediction for regression of the target).

```
Classification Report:

             precision    recall  f1-score   support

          0       0.31      0.08      0.12      2213
          1       0.50      0.32      0.39      6930
          2       0.60      0.84      0.70     10468

avg / total       0.53      0.57      0.53     19611

Accuracy is 0.569884248636
```

- **Stacking:**

  Stacking (also called meta ensembling) is a model ensembling technique used to combine information from multiple predictive models to generate a new model.

```
             precision    recall  f1-score   support

          0       0.06      0.40      0.10       328
          1       0.35      0.50      0.41      4957
          2       0.83      0.61      0.70     14326

avg / total       0.70      0.57      0.62     19611

Accuracy: 0.574881444088
```

Final test scores:

Model	Accuracy	Precision	F1 – score	Recall
Logistic Regression	47.5	51	48	48
Decision Tree	52	50	50	52
Random Forest	58	56	52	58
XG Boost	58	54	54	58
Majority Voting	52	53	50	52
Stacking	58	70	62	58

Approach – 2:

Random forest gives us variable importance. Using these important variables, models are built in this approach.

- **Logistic Regression:**

	precision	recall	f1-score	support
0	0.18	0.40	0.25	2213
1	0.43	0.27	0.34	6930
2	0.63	0.63	0.63	10468
avg / total	0.51	0.48	0.48	19611

Accuracy: 0.475804395492

- **XG Boost:**

	precision	recall	f1-score	support
0	0.04	0.42	0.08	226
1	0.38	0.50	0.43	5342
2	0.82	0.61	0.70	14043
avg / total	0.69	0.58	0.62	19611

Accuracy: 0.577584008975

Model	Accuracy	Precision	F1 – score	Recall
Logistic Regression	48	51	48	48
XG - Boost	59	71	63	59

Approach – 3:

In this approach, target variable is made to binary class from multi-class

- **Target as Yes or No:**
  - **Logistic Regression:**

	precision	recall	f1-score	support
1	0.64	0.42	0.50	7266
2	0.61	0.80	0.69	8423
avg / total	0.63	0.62	0.61	15689

Accuracy: 0.620817133023

  - **Random Forest:**

	precision	recall	f1-score	support
1	0.64	0.49	0.56	7266
2	0.63	0.76	0.69	8423
avg / total	0.64	0.64	0.63	15689

Accuracy is 0.635668302632

Model	Accuracy	Precision	F1 – score	Recall
Logistic Regression	60	62	59	61
Random Forest	62	60	61	62

- **Target as >30 or <30:**
  - **Logistic Regression:**

	precision	recall	f1-score	support
0	0.02	0.60	0.03	58
1	1.00	0.76	0.86	9085
avg / total	0.99	0.76	0.86	9143

accuracy is 0.759269386416

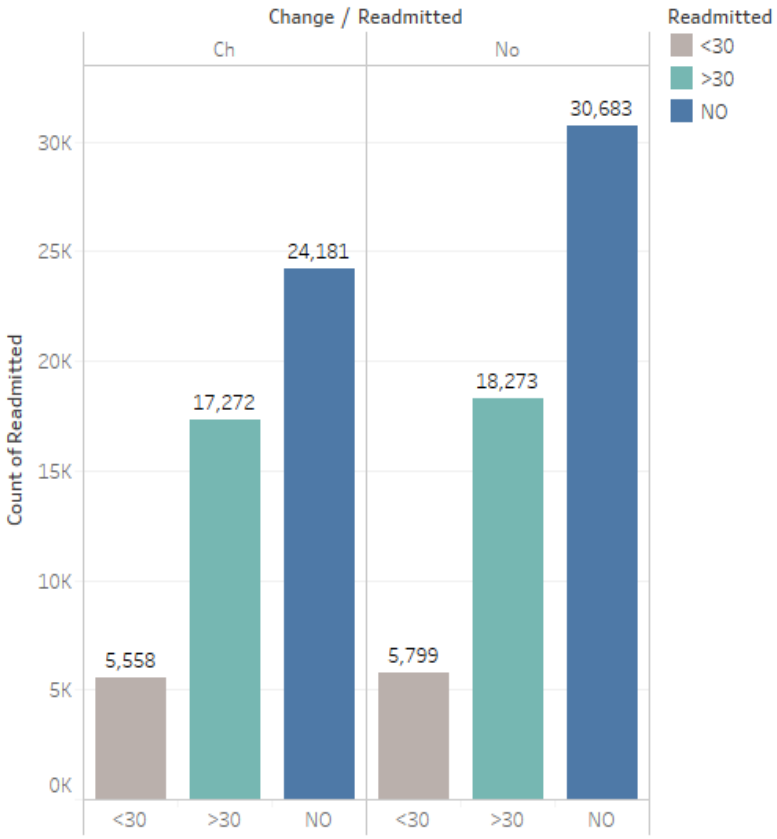
- **Random Forest:**

	precision	recall	f1-score	support
0	0.01	0.80	0.01	15
1	1.00	0.76	0.86	9128
avg / total	1.00	0.76	0.86	9143

accuracy is 0.758941266543

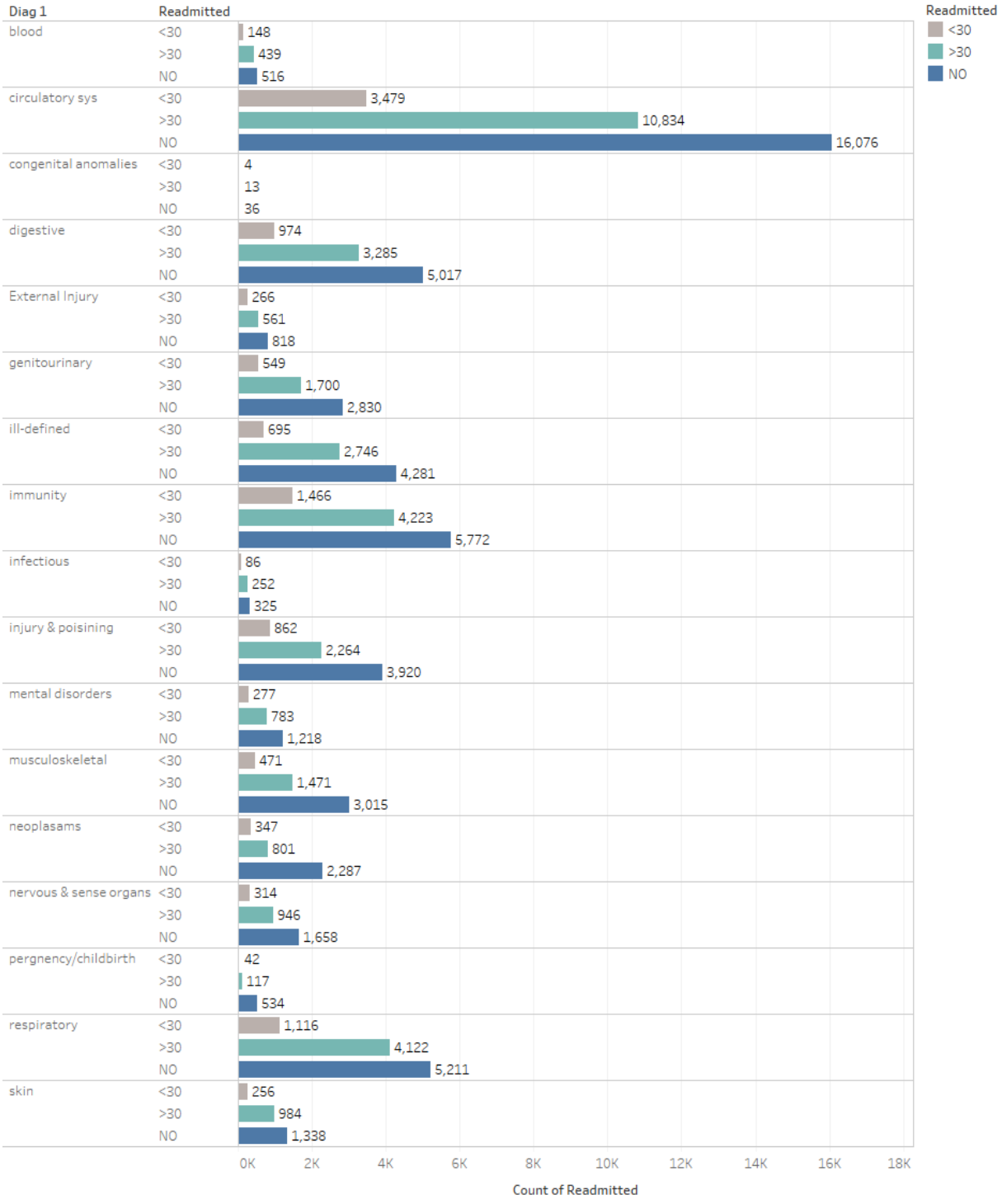
Model	Accuracy	Precision	F1 – score	Recall
Logistic Regression	76	93	86	76
Random Forest	77	95	87	77

Insights:



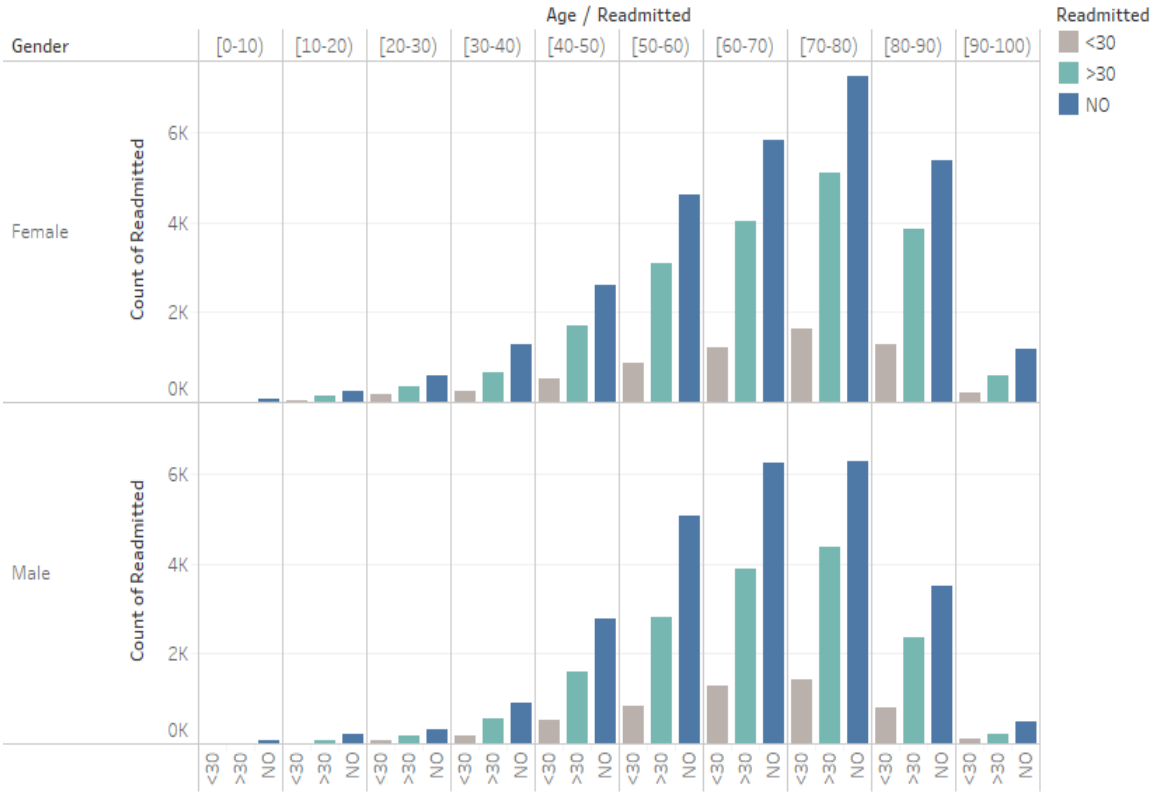
Count of Readmitted for each Readmitted broken down by Change. Color shows details about Readmitted. The marks are labeled by count of Readmitted.

From the above graph, we can infer that irrespective of the change in medications patients are readmitting.



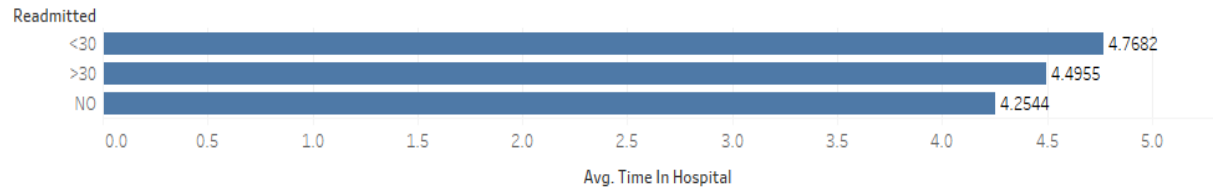
Count of Readmitted for each Readmitted broken down by Diag 1. Color shows details about Readmitted. The marks are labeled by count of Readmitted. The view is filtered on Diag 1, which excludes NA.

This graph tells us about the relationship between diagnosis 1 and readmission rates, which shows that majority of the patients have diagnosis as circulatory system.



Count of Readmitted for each Readmitted broken down by Age vs. Gender. Color shows details about Readmitted. The view is filtered on Gender, which keeps Female and Male.

This graph tells that there is a constant increase in readmission with respect to their age.



Average of Time In Hospital for each Readmitted. The marks are labeled by average of Time In Hospital.

By the we get to know that average time spent by a patient in the hospital is four to five days (4-5 days).

Conclusion:

Recall is the error metric, we cannot classify a patient wrongly as he cannot be readmitted (treatment should be done on time, this may also cost a patient's life).

Recall = True Positive / Total Actual Positive

Based on the approaches made during model building, it can be concluded that these models can be used by the client to predict hospital readmission

Approach	Model	Recall
1	Random Forest	58
2	Xg Boost	59
3	Random Forest	Target: Yes/No - 61 Target: >30 / <30 - 77

References

[1]https://en.wikipedia.org/wiki/Hospital_readmission.

[2] http://www.endocrinologyadvisor.com/type-2-diabetes/diabetes-hospital-readmission-rates/article/675957/.

[3] http://www.dailymail.co.uk/health/article-3341773/America-HIGHEST-rate-diabetes-developing-world-UK-Australia-Lithuania-nations-lowest-rates-condition.html.

[4] https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.