Data Set: Diabetes 130-US hospitals for years 1999-2008 Data Set

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.

1. Source : https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008#
2. Size: 18 MB
3. 101766 rows × 50 columns
4. Important Features:
   a) **Encounter ID** Unique identifier of an encounter
   b) **Patient number** Unique identifier of a patient
   c) **Race** Values: Caucasian, Asian, African American, Hispanic, and other
   d) **Gender** Values: male, female, and unknown/invalid
   e) **Age** Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
   f) **Weight** in pounds (missing 97%)
   g) **Admission type** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
   h) **Discharge disposition** Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
   i) **Admission source** Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
   j) **Time in hospital** Integer number of days between admission and discharge
   k) **Payer code** (missing 52%) Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
   l) **Medical specialty** (missing 53%) Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
   m) **Number of lab procedures** Number of lab tests performed during the encounter
   n) **Number of procedures** Numeric Number of procedures (other than lab tests) performed during the encounter
   o) **Number of medications** Number of distinct generic names administered during the encounter
   p) **Number of outpatient visits** Number of outpatient visits of the patient in the year preceding the encounter
   q) **Number of emergency visits** Number of emergency visits of the patient in the year preceding the encounter
   r) **Number of inpatient visits** Number of inpatient visits of the patient in the year preceding the encounter
   s) **Diagnosis 1** The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
   t) **Diagnosis 2** Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
   u) **Diagnosis 3** Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
   v) **Number of diagnoses** Number of diagnoses entered to the system 0%
   w) **Glucose serum test result** Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
   x) **A1c test result** 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.
   y) **Change of medications** Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”

z) **Diabetes medications** Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”

aa) 24 features for medications For the generic names (Label): **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

bb) **Readmitted** (Label) 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

Measure tasks to be performed:

1. EDA: Analysis of all the features in the dataset.
2. Feature importance: To fetch the better features for the prediction.
3. Data preprocessing:
    a. Diagnosis 1, Diagnosis 2, Diagnosis 3 - these 3 variables will be encoded from the ICD-9 numeric values to nominal values (Groups of disease).
    b. Possibly remove columns of Weight, Payer Code and Medical Speciality (97%, 52%, 53% missing data, respectively)
4. Prediction: Our Goal here is to predict the effective treatments (out of 24 treatments available) and to predict whether the patient will be readmitted to the hospital and if so, then whether they will be readmitted within 30 days or after 30 days .
    a. Predicting Effective Treatments
    b. Hospital Readmission prediction (Please note: a further step can be taken, wherein we can predict readmission for multiple diseases based on Alcohol test results)

References:

1. https://www.hindawi.com/journals/bmri/2014/781670/