**Interim Report**

Building Predictive model for diabetes treatment.

Submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM

**Members**:

* Gaurav Kantrod (R5V2OSNNXI)
* Komal Bamane (OV9R7Z3CCD)
* Nikhil Darade (NB4Z8HHGM6)
* Ronak Thakkar (GAR7IUSMGV)
* Shubham Singh (GOAAC5EZCF)

**Mentor**: Mr. Muppidi Srikar



Great Learning

**Abstract**

The goal of Predictive Analytics in healthcare domain is to help doctors make data-driven decisions within seconds and improve patients’ treatment. This is particularly useful in case of patients with complex medical histories, suffering from multiple conditions. Using Machine learning techniques and tools such as python would be useful to predict, for example, who is at risk of diabetes, and thereby suggest an appropriate treatment. Here, this work implements various features and compares the performance of multiple classifiers in predicting whether solo-insulin or conjunction of other drugs/ tests would be effective towards the treatment of the new patient using the data of 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks.

**Introduction**

Diabetes is a very common metabolic disease. Usually onset of type 2 diabetes happens in middle age and sometimes in old age. But nowadays incidences of this disease are reported in children as well. There are several factors for developing diabetes like genetic susceptibility, body weight, food habits and sedentary lifestyle. Undiagnosed diabetes may result in very high blood sugar level referred as hyperglycemia which can lead to complication like diabetic retinopathy, nephropathy, neuropathy, cardiac stroke and foot ulcer.

Diabetes can be treated and its consequences avoided or delayed with diet, physical activity, medication and regular screening and treatment for complications. So, early detection of diabetes is very important to improve quality of life of patients and enhancement of their life expectancy.

Data analytic is a process of examining and identifying the hidden patterns from large amount of data for drawing conclusions. In health care, this analytical process is carried out using machine learning algorithms for analysing the medical data to build machine learning models to carry out the medical diagnoses. Machine learning is a type of artificial intelligence (AI) that enables a system to learn by itself and develop the knowledge models to make decision, which drug to be given (solo insulin or combination of other drugs/treatment) to the new patients based on the analysis of data.

**Problem Statement and Objectives of Study**

**Problem Statement:**

The hospitals are evaluating efficiency of Insulin based treatment for patients. Recommend if solo insulin treatments work well towards the above stated objective. For a new patient, given his medical history and characteristics, should we recommend solo insulin or a conjunction of other drugs/ treatment?

**Objectives of the study:**

* The main objective of the study is to analyze the data of previously treated patient's and to get insights from their diabetes treatments to provide the best analytical answer to the raised question of should a new patient be treated directly with insulin shots or conjunction of medicine
* Due to the fact that the treatment that uses insulin shots causes a large “Out of pocket expenditure” which in turn affects the financial situations of patients notably.
* So the primary focus of this study is to provide the answer in 'yes or no' by providing facts, using statistics, algorithms which has taught us to prepare a machine learning model to compute a 'target variable' with the best possible success rate for corresponding problem which consist a large amount of data.
* Which would at least contribute in the debate of the question that every medical consultants and new diagnosed patients of diabetes face, isn’t it possible to use combination of medicine to maintain the sugar levels by thereby using conjunction of other drugs instead of solo insulin not affecting their financial situation considerably.

**Progress till date & Insights generated**

**Data Exploration and Preparation**

1. Combined dataset is obtained by merging the datasets from the five input files:
   1. Diabetic\_data
   2. Patient\_details
   3. Admission\_details
   4. Lab-session
   5. Diagnosis\_session
2. **Features and their description**:

List of features and their descriptions in the initial dataset (the dataset is also available at the website of Data Mining and Biomedical Informatics Lab at VCU (<http://www.cioslab.vcu.edu/>)).

|  |  |  |
| --- | --- | --- |
| Sr # | Feature | Description |
| 1 | Encounter ID | Unique identifier of an encounter |
| 2 | Patient number | Unique identifier of a patient |
| 3 | Race | Values: Caucasian, Asian, African American, Hispanic, and other |
| 4 | Gender | Values: male, female, and unknown/invalid |
| 5 | Age | Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) |
| 6 | Weight | Weight in pounds. |
| 7 | Admission type | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available |
| 8 | Discharge disposition | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available |
| 9 | Admission source | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital |
| 10 | Time in hospital | Integer number of days between admission and discharge |
| 11 | Payer code | Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay |
| 12 | Medical specialty | Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon |
| 13 | Number of lab procedures | Number of lab tests performed during the encounter |
| 14 | Number of procedures | Number of procedures (other than lab tests) performed during the encounter |
| 15 | Number of medications | Number of distinct generic names administered during the encounter |
| 16 | Number of outpatient visits | Number of outpatient visits of the patient in the year preceding the encounter |
| 17 | Number of emergency visits | Number of emergency visits of the patient in the year preceding the encounter |
| 18 | Number of inpatient visits | Number of inpatient visits of the patient in the year preceding the encounter |
| 19 | Diagnosis 1 | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values |
| 20 | Diagnosis 2 | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values |
| 21 | Diagnosis 3 | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values |
| 22 | Number of diagnoses | Number of diagnoses entered to the system |
| 23 | 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 |
| 24 | 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. |
| 25 | Change of medications | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” |
| 26 | Diabetes medications | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” |
| 27 | 24 features for medications | 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 |
| 28 | Readmitted | 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. |
| 29 | Admission\_id | 1 Emergency  2 Urgent  3 Elective  4 Newborn  5 Not Available  6 NULL  7 Trauma Center  8 Not Mapped |
| 30 | discharge\_disposition\_id | 1 Discharged to home  2 Discharged/transferred to another short term hospital  3 Discharged/transferred to SNF  4 Discharged/transferred to ICF  5 Discharged/transferred to another type of inpatient care institution  6 Discharged/transferred to home with home health service  7 Left AMA  8 Discharged/transferred to home under care of Home IV provider  9 Admitted as an inpatient to this hospital  10 Neonate discharged to another hospital for neonatal aftercare  11 Expired  12 Still patient or expected to return for outpatient services  13 Hospice / home  14 Hospice / medical facility  15 Discharged/transferred within this institution to Medicare approved swing bed  16 Discharged/transferred/referred another institution for outpatient services  17 Discharged/transferred/referred to this institution for outpatient services  18 NULL  19 Expired at home. Medicaid only, hospice.  20 Expired in a medical facility. Medicaid only, hospice.  21 Expired, place unknown. Medicaid only, hospice.  22 Discharged/transferred to another rehab fac including rehab units of a hospital .  23 Discharged/transferred to a long term care hospital.  24 Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.  25 Not Mapped  26 Unknown/Invalid  30 Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere  27 Discharged/transferred to a federal health care facility.  28 Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital  29 Discharged/transferred to a Critical Access Hospital (CAH). |

1. **Rejected features:**
   1. ***Weight*** feature has 96.85 % missing values in the dataset.
   2. ***examide, citoglipton*** are single-value features
   3. ***payer\_code*** has 39.55% missing values
   4. ***medical\_speciality*** has 49.08% missing values
   5. The feature ***'diabetesMed'*** implies whether diabetes medicines were prescribed or not. Since, our objective is to check whether 23 drugs or solo insulin worked better for the patient. All the records which where none of the 24 drugs/tests have been prescribed does not come under the scope of our analysis. Removing these records where features 'diabetesMed' is not prescribed. Now, we have records of features where ***'diabetesMed'*** is 'Yes' i.e. single value feature so dropping those ***'diabetesMed'*** as well.
2. **Missing values treatment and Imputation:**
   1. ***Gender*** feature with 3 records with Unknown/Invalid values were dropped.
   2. ***Race*** feature with missing values (0.022%) were merged with ‘Other’ existing category.
   3. Missing values in the ***diag\_1***, ***diag\_2*** and ***diag\_3*** are imputed with mode of their respective columns (values are numeric as well as alphanumeric.
3. **Feature Engineering**:
   1. ***diag\_1***, ***diag\_2*** and ***diag\_3*** states primary diagnosis, secondary diagnosis and other additional secondary diagnosis prescribed to the patient respectively. Also, these features consists of International Statistical Classification of Diseases and Related Health Problems (ICD9) code which is the international "standard diagnostic tool for epidemiology, health management and clinical purposes and a standard list of six-character alphanumeric codes to describe diagnoses. First numeric or Alphabets before decimal indicates category while the numeric after the decimal indicates category, anatomic site and severity as shown in the below figure Fig. 1

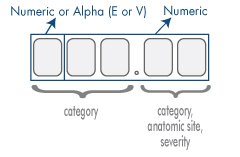


Fig. 1

These ICD9 codes are categorized into broader categories of the diseases they represent as seen in the Fig. 2

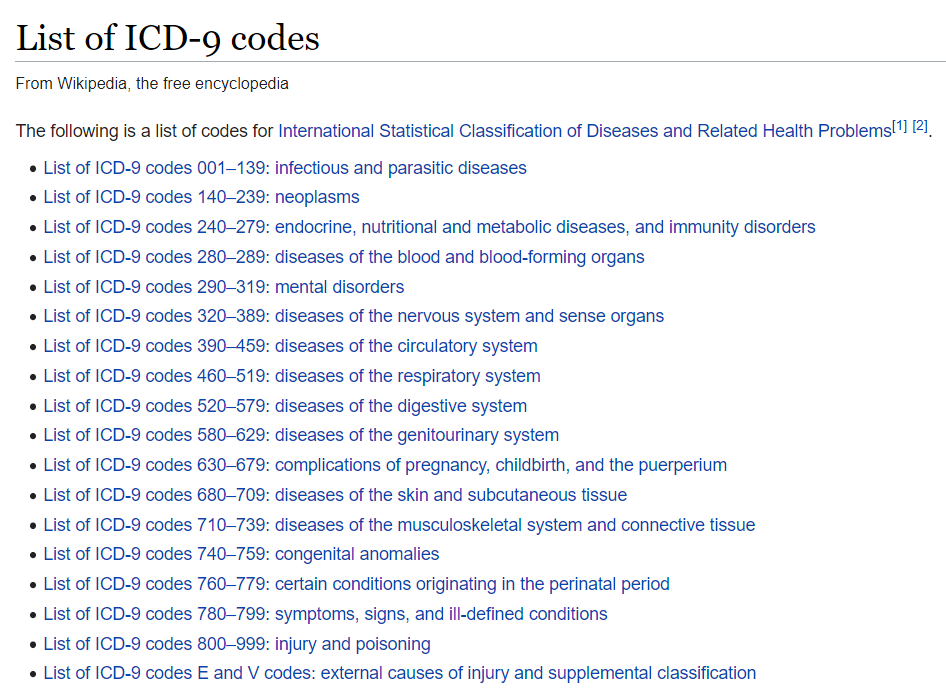


Fig. 2

So, these diag\_1, diag\_2 and diag\_3 features are converted into one of the above-mentioned categories depending upon on the range in which they fall making them categorical features.

* 1. ***Age*** feature which grouped into intervals of tens from 0-10 to 90-100 is converted into a discrete variable by imputing with the midpoint of its respective range.

1. **One-hot encoding**:
   1. ***readmitted*** encoded as: ‘No readmission’ as 0 and ‘readmitted’ as 1
   2. ***insulin*** and other 23 drugs/tests encoded as : ‘No’ as 0 and ‘Up’, ‘Steady’, ‘Down’ as 1
   3. ***change*** encoded as: ‘Ch’ as 1 and ‘No’ as 0
   4. performed one hot encoding on features ***diag\_1***, ***diag\_2*** and ***diag\_3*** which holds 19 categorical values of disease category.
   5. ***max\_glu\_serum\_dict*** encoded as: ‘None’ as 0, ‘Norm’ as 100, ‘>200’ as 200, ‘>300’ as 300
   6. ***A1Cresult*** encoded as: ‘None’ as 0, ‘Norm’ as 100, ‘>7’ as 7 and ‘>8’ as 8
2. ***Target Feature:***

Problem statement requires us to engineer a binary class target variable/ dependentvariable ***‘Treatment type’*** variable which will indicate that whether solo-insulin or a conjunction of other drugs were prescribed to the patient. Here, we need to consider the one hot encoded features like insulin and other 23 drugs to classify the patient records based on whether solo-insulin or other drugs were prescribed and gauge their effectiveness.

1. ***Data Visualization:***
   1. Most of the diabetic patients across 130 U.S hospitals fall in the age bracket of 55 to 85 as shown in the Fig. 3

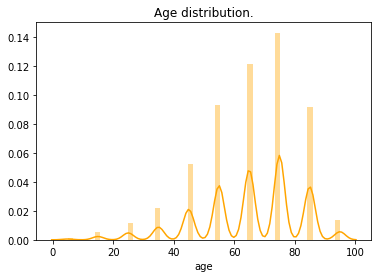
******

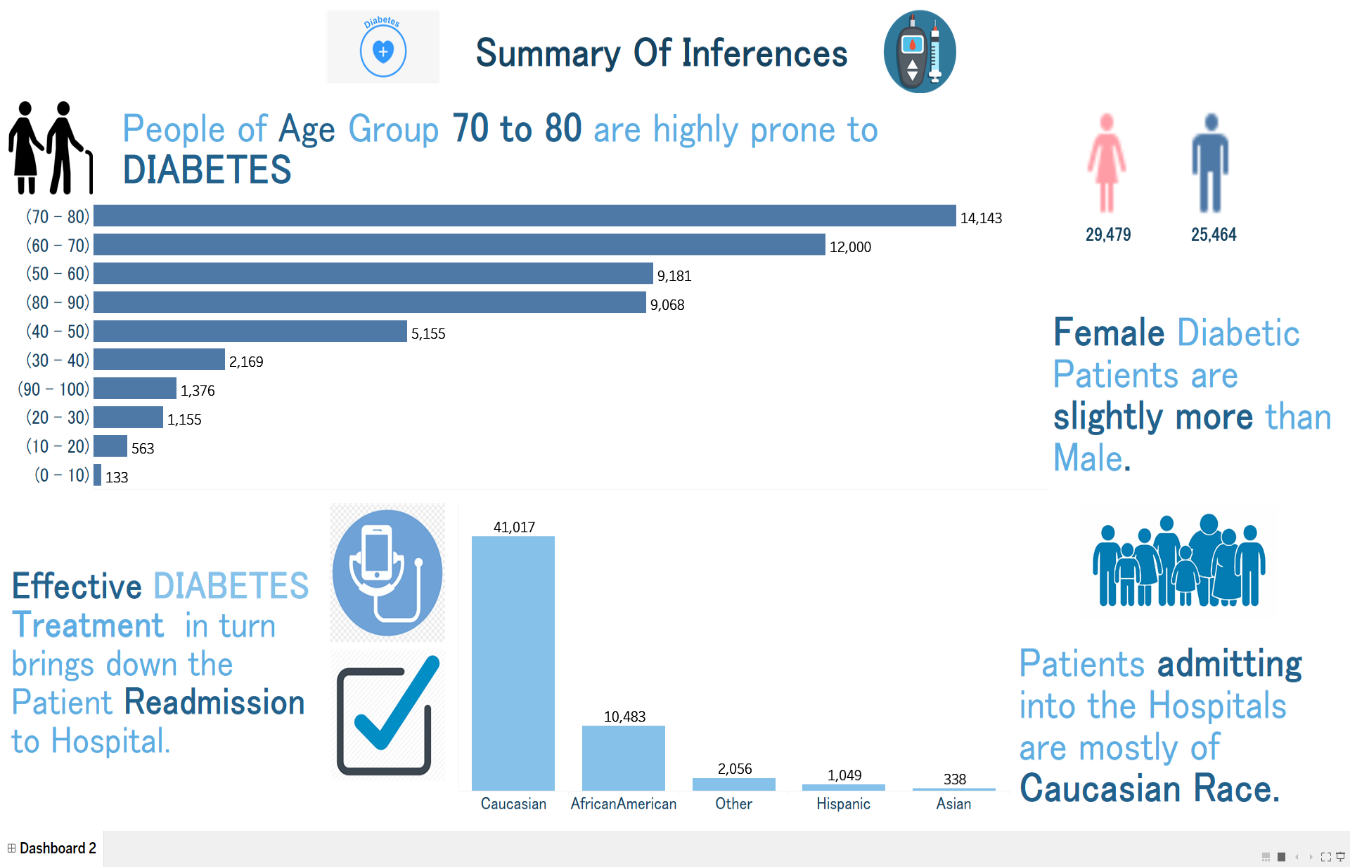
Fig.3

* 1. Fig. 4 and Fig.5 shows that dataset seems pretty balanced with respect to ‘readmitted’ and ‘gender’.

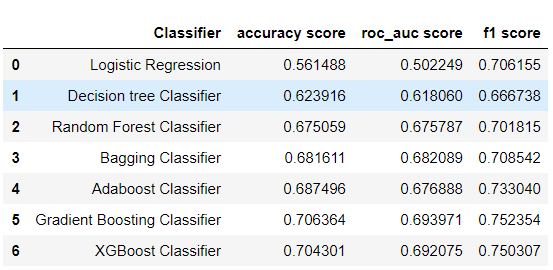
Fig.4 Fig.5

* 1. Below Fig.6 show data visualization of other features done in the tableau dashboard.



Dashboard Tableau Summary Fig. 6

1. **Data Modelling:**
   1. Data modelling was done on the baseline classification algorithms like Logistic Regression, Decision Tree, Random Forest, Adaboost, Gradient Boost and Xgboost. Table below shows their performance with respect to accuracy score, roc\_auc curve and f1-score:



* 1. Of the baseline classification models applied to the data, we observe that Gradient Boosting and Xgboost Classifiers performed better on above mentioned evaluation metrics.

**Next Steps:**

Further Steps may include performing –

* Outlier treatment, Standardization, transformation if necessary
* Selecting one of the above models – Gradient Boost and Xgboost.
* Tuning the hyperparameters of the selected model using GridSearchCV which searches exhaustively from the given subset of parameters

**Challenges:**

1. Understanding **ICD9 codes nomenclature**, Interpreting test results such as **Max Glucose serum test** and **Ac1 result** required Subject knowledge expert.

**References**:

1. <http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>
2. <https://www.hindawi.com/journals/bmri/2014/781670/tab1/>
3. <http://hppaccuchecker1.blogspot.com/2015/10/mapping-of-icd-9-to-icd-10.html>
4. <https://en.wikipedia.org/wiki/List_of_ICD-9_codes>
5. <https://www.webmd.com/diabetes/guide/glycated-hemoglobin-test-hba1c>

\*\*\*