**Executive Summary for Project on Diabetes**

1. What is the problem you want to solve?
   * Identify the markers for Diabetes
   * Identify the key People who are non-diabetics but have some chances of Diabetes
   * Create a logistic model that would predict diabetes for people as an outcome for known BMI,Triceps Thickness
   * Use this model to leverage early detection of Diabetes and or use the conventional alternate therapy to sample group with higher diabetic predictability.
   * Alternative Treatments to this control group will help in better monitoring and prevent worsening of disease naturally.
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

Diabetes is a disease that occurs when the insulin production in the body is inadequate or the body is unable to use the produced insulin in a proper manner, as a result, this leads to high blood glucose. The body cells break down the food into glucose and this glucose needs to be transported

to all the cells of the body. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal range as per different testing mechanism

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WHO diabetes diagnostic criteria[[44]](https://en.wikipedia.org/wiki/Diabetes_mellitus#cite_note-who2006-44)[[45]](https://en.wikipedia.org/wiki/Diabetes_mellitus#cite_note-45)  [edit](https://en.wikipedia.org/w/index.php?title=Template:OGTT&action=edit) | | | | |
| **Condition** | **2 hour glucose** | **Fasting glucose** | **HbA1c** | |
| Unit | mmol/l(mg/dl) | mmol/l(mg/dl) | mmol/mol | DCCT % |
| Normal | <7.8 (<140) | <6.1 (<110) | <42 | <6.0 |
| [Impaired fasting glycaemia](https://en.wikipedia.org/wiki/Impaired_fasting_glycaemia) | <7.8 (<140) | ≥6.1(≥110) & <7.0(<126) | 42-46 | 6.0–6.4 |
| [Impaired glucose tolerance](https://en.wikipedia.org/wiki/Impaired_glucose_tolerance) | ≥7.8 (≥140) | <7.0 (<126) | 42-46 | 6.0–6.4 |
| **Diabetes mellitus** | ≥11.1 (≥200) | ≥7.0 (≥126) | ≥48 | ≥6.5 |

Lot of Data on Healthcare is not analyzed. While manual intervention works well for individual particular cases- health care companies need to identify people who have higher chances of diabetes so that regular monitoring of data can be done. We tend to solve this problem by analyzing large amount of sample data

* + Imperative need to identify mark up for diabetes at early stages for better clinical Outcome
  + Help in Better Cognition and advance detection of disease
  + Create a Model which could help in Early Detection of Diabetes

While our decision model cannot replace individual health decision- the probabilistic model could be leveraged by:

1. Laboratory Companies in Health Care Sector for Identifying people who need regular screening or alternative intervention to reduce BMI
2. Insurance Companies
3. Hospitals & Health Care Organizations
4. What data are you going to use for this? How will you acquire this data?

The Data has been acquired from UCI database below

* + <https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>

1. In brief, outline your approach to solving this problem (knowing that this might change later)

The model needs to run for both Training Data and Test Data

**High Level Methodology**

Known Individual

Predictors

Iterative Probabilistic Logistic Regression/

Bayesian Network

Other Predictors Not Known Predictor

Legend

* 1. Body Mass Index- It is one the risk factors for Type 2 Diabetes
  2. Diastolic Pressure(DB)- Another risk factor for Diabetes
  3. TT- Triceps Thickness- Another risk Factor of Diabetes
  4. Hereditary – As some Chronic Conditions are embedded in cells –there is tendency of humans to suffer from chronic diseases due to diabetes

Unknown Factors- Pregnancy/No of Times Pregnant is also one of the factors for Diabetes.

The probabilistic model can be run differently for patients without hereditary inheritance.

How do we improve the performance of the model?

* 1. The performance of the model will be improved by creating subsets or sub domains by creating different factors of BMI- High BMI, Medium BMI & Low BMI. This could be done further with Diastolic Blood Pressure and Triceps Thickness..
  2. Once we have subsets – then we create logistic regression (Diabetes & No Diabetes) of different subjects within that subset.

Mathematical Approach for Known Predictors

…………………EQUATION 1

Leverage the above formula y(Diabetes) for different subsets by creating different subsets & factors like High BMI, Low BMI, Triceps Thickness. Using Natural Log

……………………..Equation 2

Now we use the above formula for three different sub sets generated from General Dataset.

Data set1= Factor High BMI

…(when ………………………………………..Equation 3

Now from the test data set separate High factor BMIs and run and test the prediction model based on subset created from equation 3.

|  |  |  |
| --- | --- | --- |
|  | Prediction –diabetic | prediction non Diabeteic |
| Diabetic | X1 | X2 |
| Non diabetic | X3 | X4 |

Specificity= X1/X1+X2

SENSTIVITY=X4/X3+X4

1. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

Paper/PPT

* 1. Code
  2. Methodology
  3. Graphs
  4. Mathematical Construct
  5. White Paper/Presentation.