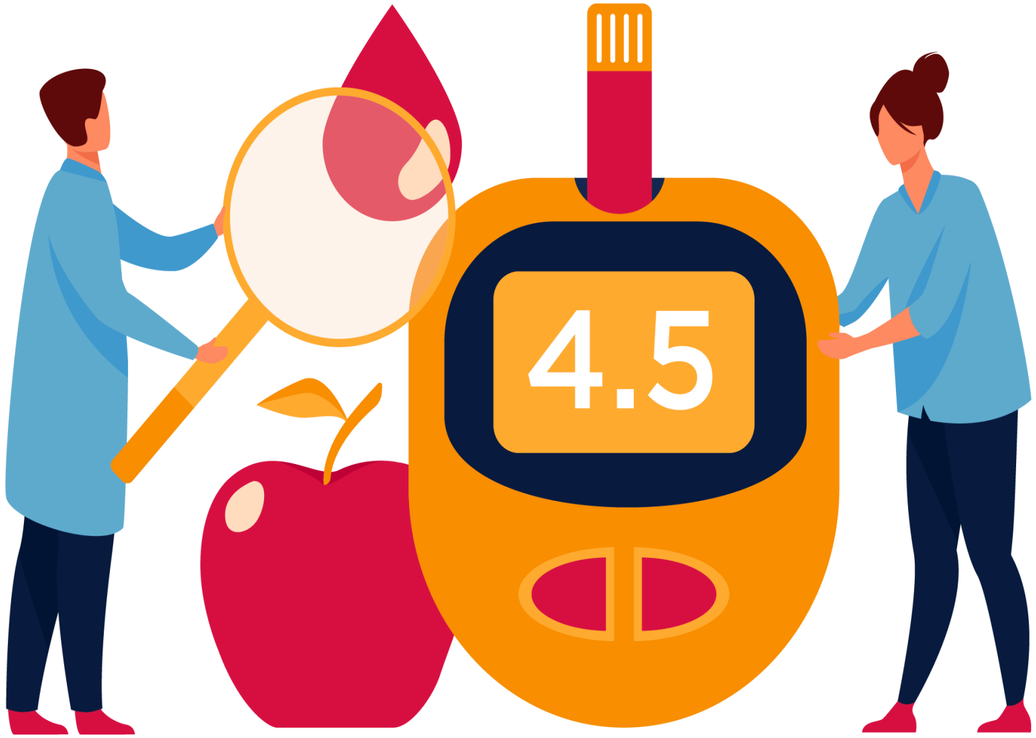


Diabetes Prediction

Vidya Venugopal Sharma
Tarunvel Venugopal Santhamani

Diabetes and its impact on healthcare

- Approximately 537 million adults (20-79 years) are living with diabetes. The total number of people living with diabetes is projected to rise to 643 million by 2030 and 783 million by 2045.

- Almost 1 in 2 (240 million) adults living with diabetes are undiagnosed. It has caused 6.7 million deaths

- Diabetes caused at least USD 966 billion dollars in health expenditure – 9% of total spending on adults

- The prevalence of diabetes worldwide is substantial and continues to rise. Its impact on public health with significant health complications, economic burdens, and health disparities.

- The goal of the project is Diabetes prediction to follow the comprehensive approach of prevention, early detection, and effective management to mitigate the impact of diabetes.

Project Goal

- To analyze the dataset for diabetes prediction that aids in the comprehensive approach of prevention, early detection, and effective management to mitigate the impact of diabetes

- To classify the patients as ‘having diabetes’ or ‘not having diabetes’ based on the dataset

Dataset

- The data source we are analyzing is from Kaggle: kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

- The data has around 1 lakh records and is a collection of medical as well as demographic data from patients, along with their diabetes status (positive or negative).

- The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.

Data Preparation and Transformation

Step 1: Data Cleaning

- **Handling missing data:** The missing data (“No info”) in the field ‘ Smoking History’ were removed
- **Random Sampling:** Created a “Random Sampling Variable” column in the data using the formula (=RANDBETWEEN(1,64185)
- **Selection of data:** 10,000 records of data were picked for analysis.

Step 2: Variable Selection

- **Numerical Variables:** Age, bmi, hbA1c_level, blood_glucose_level
- **Categorical Variables:** Gender, hypertension, heart_disease, smoking_history_trans, diabetes

Step 3: Exploratory Data Analysis

- **Correlation:** To understand the relationship between numerical variables
- **Data visualization:** we have used histogram, boxplot to see how the data is spread out and find outliers,

Model and Variable Selection

- **Model Selection:** Logistic regression is used for the following reasons:
  1. The dependent variable is binary classification i.e. the output is classified as 'a person having diabetes' or 'not having diabetes'
  2. The variables exhibit linearity

- **Variable selection**: Based on the P-Value in the table below, we have selected the independent variables are: Age, BMI, HbA1c_level, blood_glucose_level, gender_Male, hypertension_1 , heart_disease_1

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-26.95977524	-29.26650387	-24.6530466	1.95667E-12	1.176923991	524.7297681	3.9585E-116
age	0.052787262	0.043267077	0.062307446	1.054205351	0.004857326	118.1037493	1.64552E-27
bmi	0.088536631	0.067546697	0.109526564	1.092574273	0.010709347	68.34709057	1.37107E-16
HbA1c_level	2.173805479	1.898340969	2.449269989	8.791677002	0.1405457	239.2248176	5.80405E-54
blood_glucose_level	0.034511954	0.030560749	0.038463158	1.035114402	0.002015958	293.073291	1.06396E-65
gender_Male	0.524457395	0.23169166	0.81722313	1.689541846	0.149373018	12.32753139	0.000446326
hypertension_1	0.851984038	0.483797845	1.22017023	2.344293407	0.18785355	20.56951739	5.75045E-06
heart_disease_1	0.598546399	0.113576883	1.083515914	1.81947209	0.247437973	5.851442544	0.015564288
smoking_history_trans_Past	0.195351064	-0.117007517	0.507709644	1.215737713	0.159369551	1.502522076	0.220283707
smoking_history_trans_Present	0.378511734	-0.068502716	0.825526184	1.46010994	0.228072788	2.754303242	0.096993063

- **Variable Selection Technique:** Stepwise Selection
- **Cut Off Value** = 0.4

Model Output and Equation

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-26.82172593	-29.11626002	-24.52719184	2.24632E-12	1.170702171	524.9048599	3.6261E-116
age	0.052449024	0.043163432	0.061734615	1.053848839	0.004737634	122.5608719	1.73994E-28
bmi	0.088354249	0.067367263	0.109341235	1.092375026	0.010707843	68.08491946	1.56604E-16
HbA1c_level	2.175090293	1.899726361	2.450454224	8.802979929	0.140494384	239.6826808	4.61207E-54
blood_glucose_level	0.034404234	0.030462027	0.038346441	1.035002905	0.002011367	292.5776709	1.36431E-65
gender_Male	0.56389213	0.274740673	0.853043587	1.757499623	0.147528964	14.60957152	0.000132241
hypertension_1	0.846646041	0.479643696	1.213648385	2.331812916	0.187249535	20.44383035	6.14072E-06
heart_disease_1	0.600428506	0.116589475	1.084267537	1.822899756	0.246861185	5.915847739	0.015005281

Logit (Diabetes = 1) = -26.8217+ 0.05244age + 0.08835bmi + 2.1751HbA1c_level + 0.0344blood_glucose_level + 0.56389gender_Male + 0.8466hypertension_1 + 0.6heart_disease_1

*Odds (Diabetes = 1) = 2.24632E-12 * 1.054^age * 1.0925^bmi * 8.8029^HbA1c_level *1.035^blood_glucose_level * 1.7574^gender_Male * 2.332^hypertension_1 * 1.8229^heart_disease_1*

P (Diabetes = 1) = 1/(1+E^(-26.8217+ 0.05244age + 0.08835bmi + 2.1751HbA1c_level + 0.0344blood_glucose_level + 0.56389gender_Male + 0.8466hypertension_1 + 0.6heart_disease_11))

Research Questions

1. What are the impacts of gender, age, hypertension, heart disease, smoking history, BMI, HBAC1 level, and blood glucose level on diabetes? What combinations of parameters make a person more likely to get diabetes?

The Age, BMI, Normal Hemoglobin A1c Level, Blood Glucose level has higher impact on person having diabetes than other independent variables.

2. Which are the most significant predictors of diabetes risk?

BMI,Age, Normal Hemoglobin A1c Level, Blood Glucose Level, Hypertension, Heart Disease

3. How does BMI of a person affect the chances of having diabetes?

Based on the model and analysis, the BMI increases then the person getting diabetes also increases i.e there exists a direct relationship. As per the model, If BMI increases by one unit, the odds of a person getting diabetes would be multiplied by 1.0925 holding other variables constant.

5. How does Blood Glucose level & HbA1c level of a person affect the risk of having diabetes?

When “Blood Glucose” level and “Normal Hemoglobin A1c” levels increase, the risk of getting diabetes is higher

6. Does age and gender have an impact on getting diabetes?

Age has an impact on getting diabetes whereas gender does not impact the person from getting diabetes.

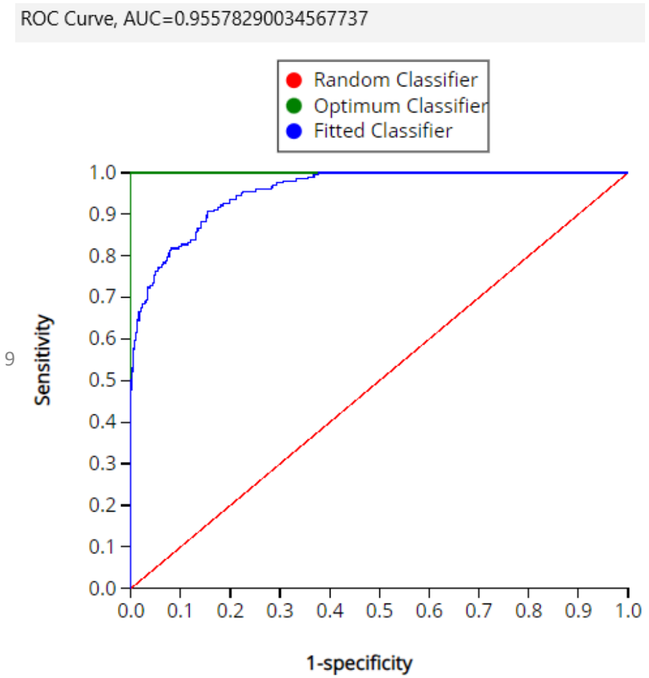
Model Performance and Evaluation

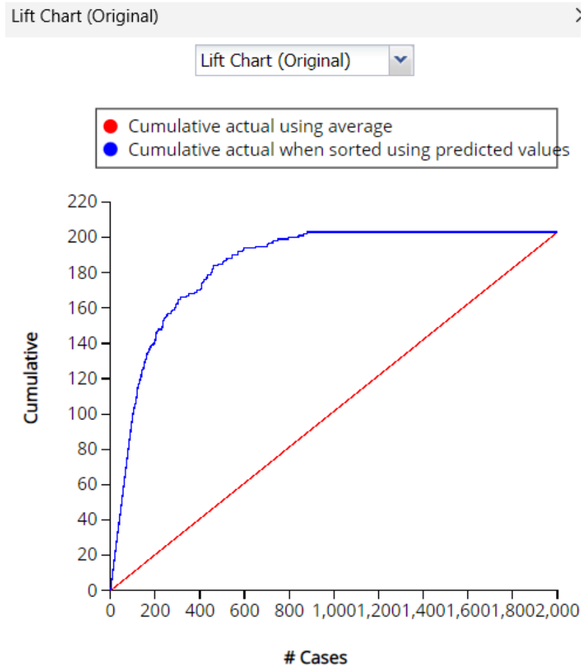
- **Based on the confusion matrix**

Confusion Matrix		
Actual\Predicted	0	1
0	1757	40
1	65	138

- Accuracy = 94.75%
- Specificity = 97.77%
- Sensitivity (Recall) = 67.98%
- Precision = 77.53%

- **Based on ROC Curve and Lift Chart**





Conclusion

Based on the analysis of the logistic regression model, when the age is increasing the risk of getting diabetes is also increasing.

Further, when Hemoglobin (HbA1c) and Blood Glucose level increases in the blood, the person needs to take care of his health because the increase in the HbA1c and blood glucose level would lead to diabetes. Therefore, maintaining the optimal level of HbA1c and blood glucose level is very important to mitigate diabetes.

When it comes to prior health conditions such as hypertension and heart disease, 3 in 10 people with hypertension would have a chance of diabetes. There is a one in five people having heart disease would have the chance of having diabetes.

From the model, we conclude that age, Hemoglobin (HbA1c) level, Blood Glucose level along with the prior health conditions such as hypertension and heart disease increases the risk of getting diabetes.