ExplAIning Diabetes Predictions

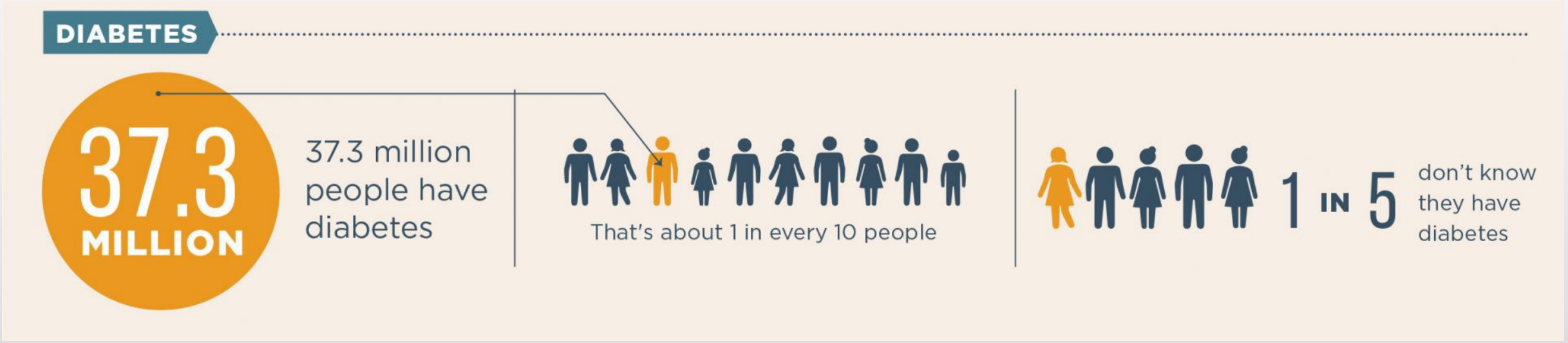
DBA5102

RACHEL SNG (A0231921N)
WONG CHENG AN (A0232039M)
FELIPE CHAPA CHAMORRO (A0179033E)
GINO MARTELLI TIU (A0231956Y)
SUSAN KORUTHU (A0231905L)
WIDYA GANI SALIM (A0231857Y)

Accurate and early diabetes prediction is critical to prevent patients from suffering more serious, long-term damage

BUSINESS BACKGROUND

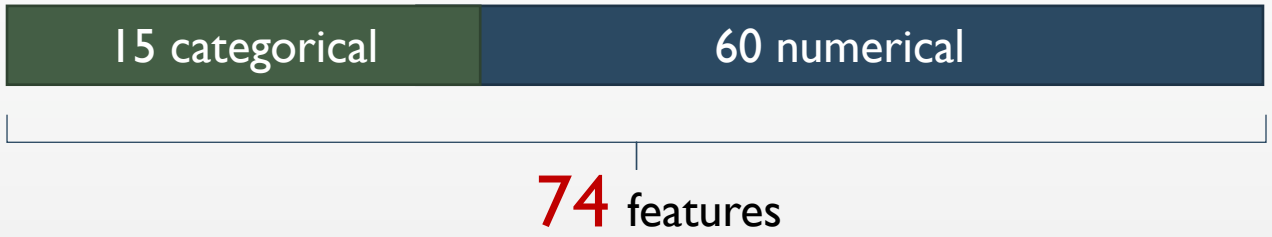
- Diabetes affects 1 in 4 people over the age of 65 in USA, and 1 in 10 overall
- Many patients **do not know** that they have the disease as early onset symptoms of prediabetes can be mild and unnoticeable
- If diabetes is diagnosed and treated early it can prevent more serious long-term organ damage and even be reversible

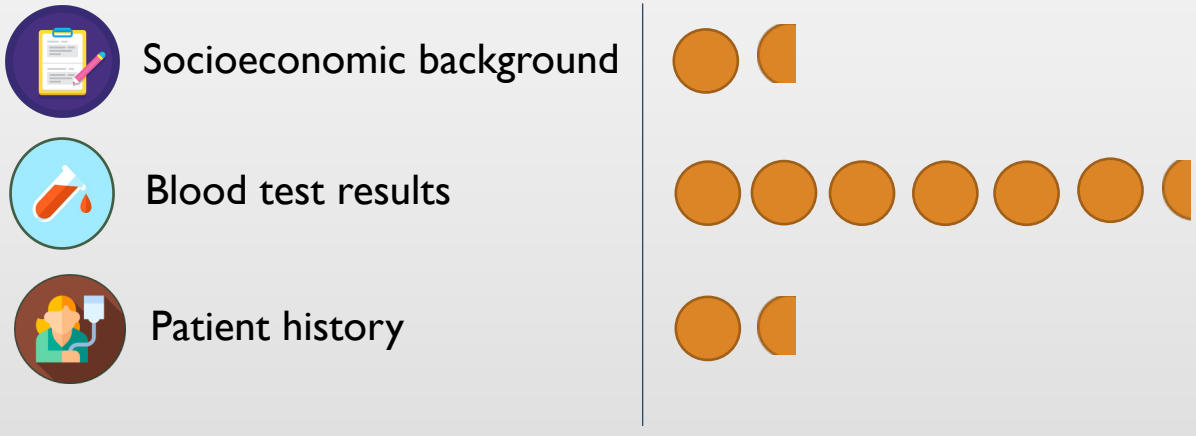


Predict propensity for diabetes given demographic and socioeconomic determinants

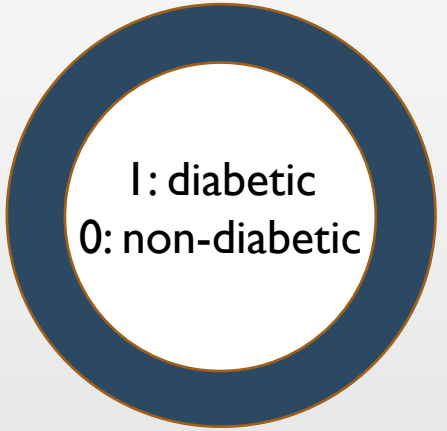
Dataset: National Health and Nutrition Examination Survey (NHANES)

8,610 unique individuals





PREDICTION GOAL



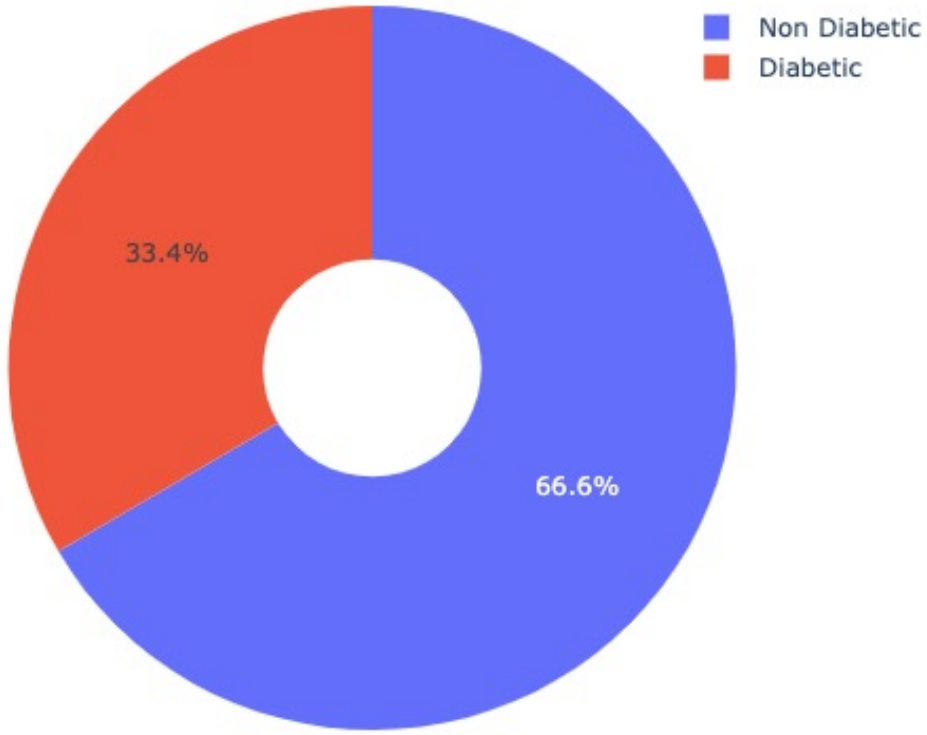
Target Variable
Is_diabetic (binary)

Understanding the Data

EXPLORATORY DATA ANALYSIS

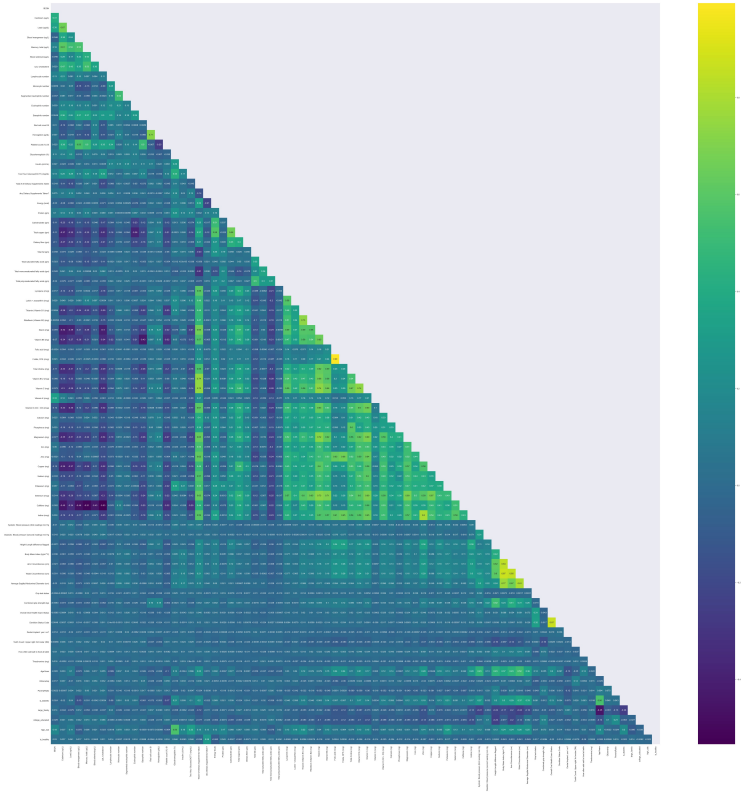
Exploratory Data Analysis: Initial Insights to Dataset

1 **Imbalanced Samples:** % of diabetic individuals within the sample is low



To improve our model performance, we adopted resampling techniques, SMOTE, to acquire a more balanced sample.

2 **Multicollinearity:** There were high correlations between some features



7 features (e.g: Folic Acid and Foleate) were dropped, this ensured that Permutation Feature Importance (PFI) score is accurate. We also implemented feature selection algorithm in our pipeline to ensure we include only important features.

Exploratory Data Analysis: Differences between Diabetic and Non-Diabetic Groups

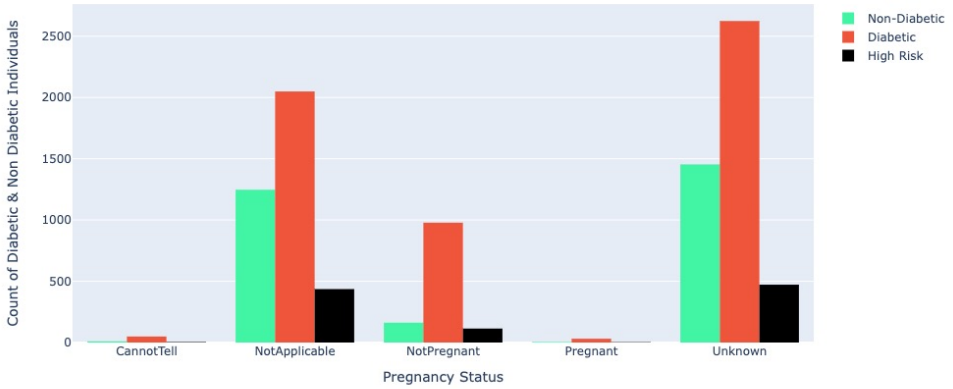
On average, blood nutrients that are indicative of diabetes (e.g: glycohemoglobin, insulin and platelets count) are around 5% higher in diabetic groups. It is also notable that average BMI of diabetic group is obese. More than half of overweight and obese population is diabetic, higher than that in other healthy population. Percentage of diabetic individuals in married population is lowest as compared to other groups. Meanwhile, pregnancy status does not seem to be indicative of diabetes.

Blood Nutrients Level for Diabetic and Non-Diabetic Groups

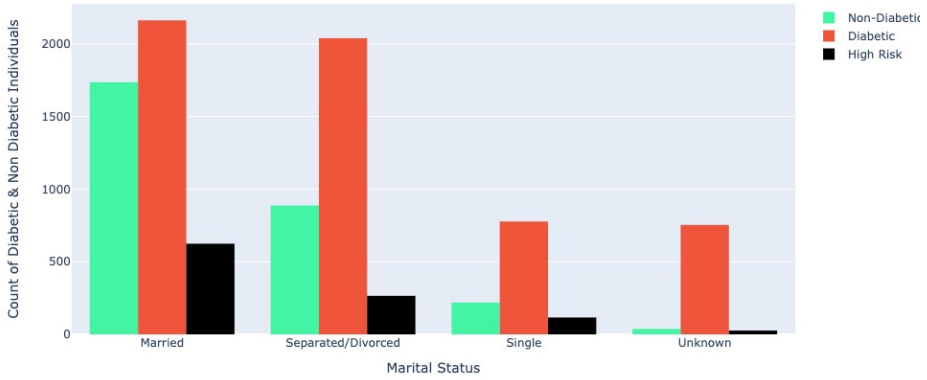
Non-Diabetic			
216.18 Average of LDL-Cholesterol (mg)	2.15 Average of Lymphocyte Number	0.22 Average of Eosinophilis Number	0.059 Average of Basophilis Number
274.96 Average of Platelet Count (%) SI	5.62 Average of Glycohemoglobin (%)	12.34 Average of Insulin (uU/mL)	2.33 Average of Total Sugars (gm)
2.03 Average of Dietary Fiber (gm)	1.91 Average of Total Fat (gm)	25.74 Average of Body Mass Index (kg/m**2)	64.06 Average of Combined Grip Strength (kg)

Diabetic			
216.32 Average of LDL-Cholesterol (mg)	2.11 Average of Lymphocyte Number	0.23 Average of Eosinophilis Number	0.055 Average of Basophilis Number
262.89 Average of Platelet Count (%) SI	5.84 Average of Glycohemoglobin (%)	12.90 Average of Insulin (uU/mL)	2.30 Average of Total Sugars (gm)
2.04 Average of Dietary Fiber (gm)	1.91 Average of Total Fat (gm)	30.84 Average of Body Mass Index (kg/m**2)	63.49 Average of Combined Grip Strength (kg)

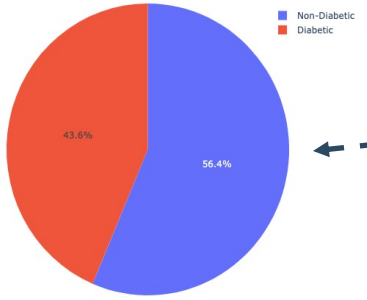
Count of High Risk, Diabetic and Non Diabetic Individuals based on Pregnancy Status



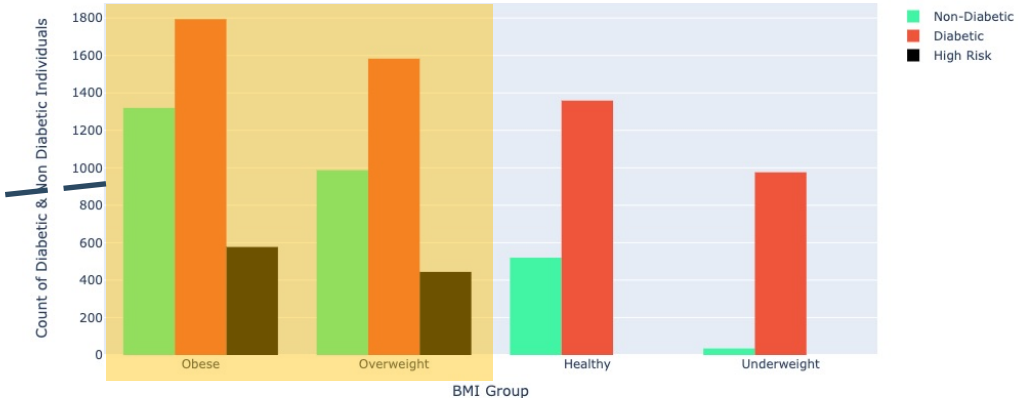
Count of High Risk, Diabetic and Non Diabetic Individuals based on Marriage Status



% of Diabetic Individuals among Overweight & Obese Population

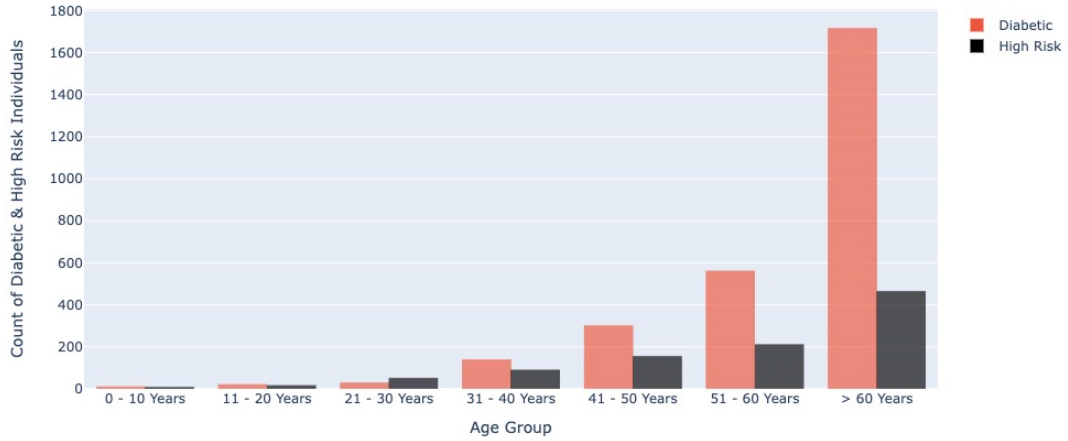


Count of High Risk, Diabetic and Non Diabetic Individuals based on BMI Group



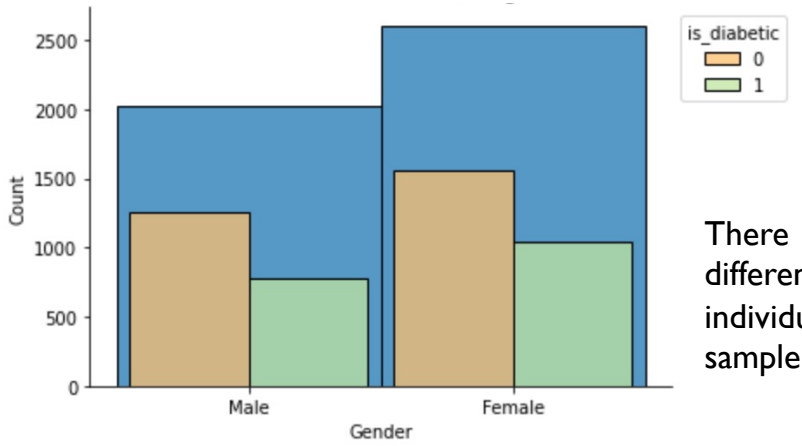
Exploratory Data Analysis: Further Investigation on Demographic Data

Count of Diabetic and High Risk Individuals per Age Group



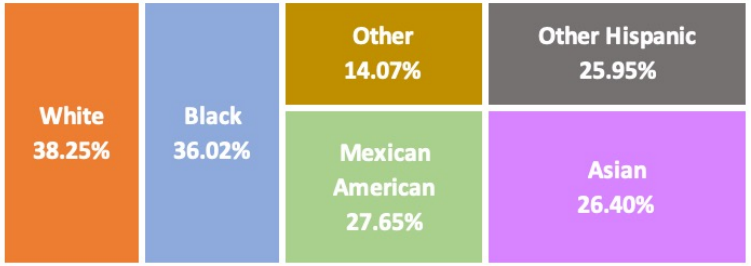
Older age groups have higher diabetic and high risk individuals count. This is unsurprising given that the risk of becoming diabetic increases with age.

Gender Distribution with Target Class



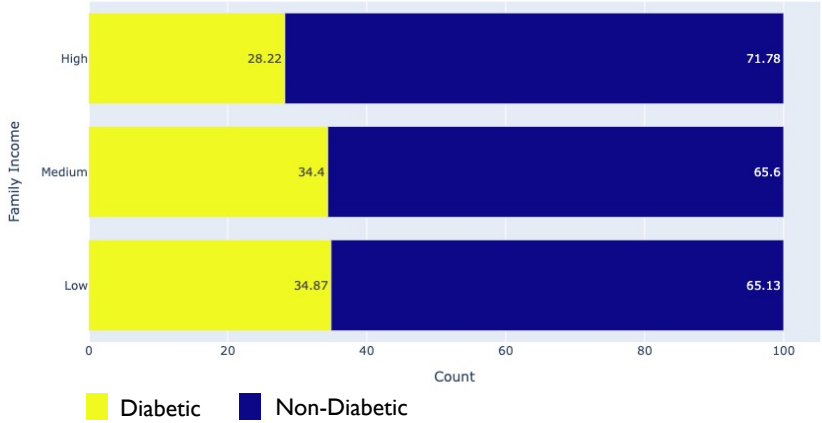
There seems to be minimal difference between diabetic individuals among genders in the sample.

% of Diabetic Individuals per Race



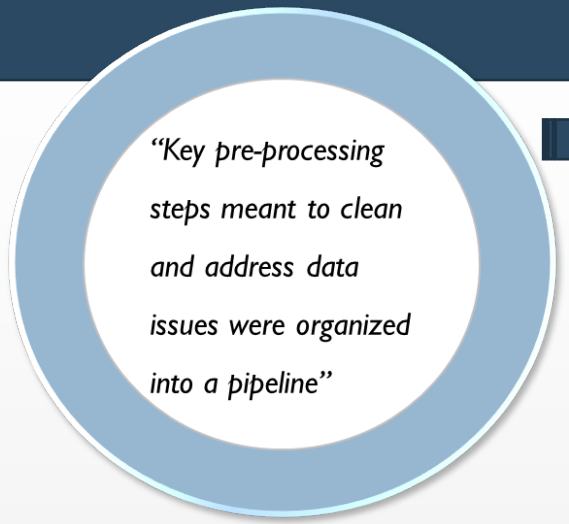
White and black population have more 10% more diabetic individuals as compared to the remaining population. This could be due to lifestyle, such as sugar consumption, diet and exercise habits.

% of Diabetes per Family Income



Family with higher income seems to have lower percentage of diabetes. This could be caused by their ability to afford a healthier lifestyle or diet.

4-Step Pipeline to Transform, Encode and Select Features

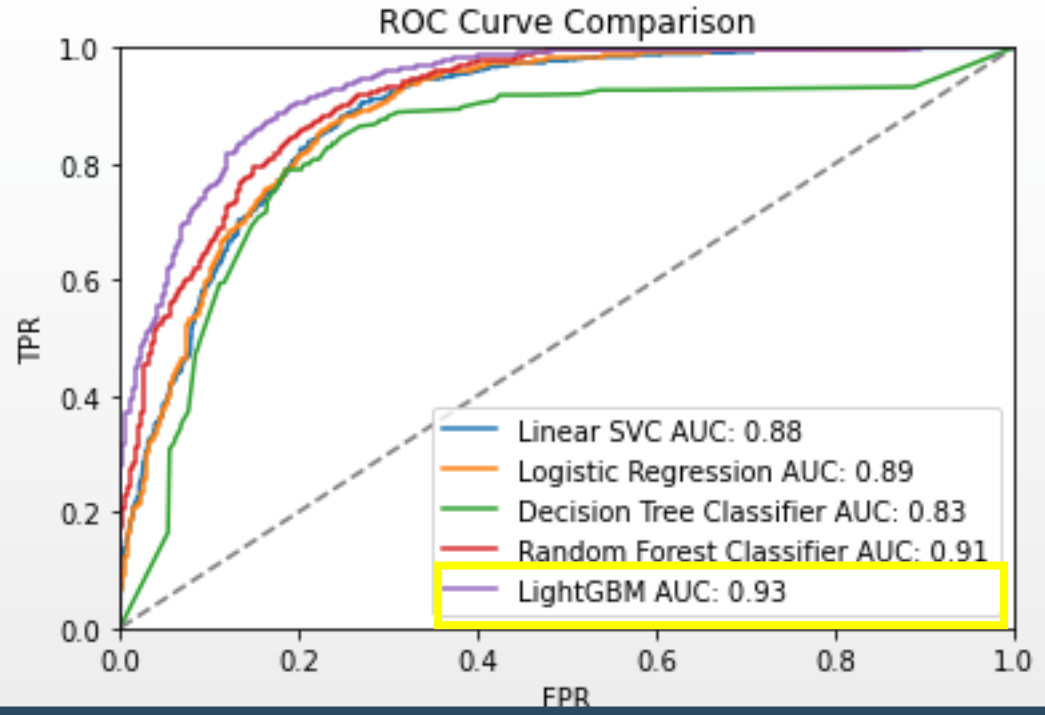


Component	Objective	Target Columns
1. Standard Scaler	Normalize numerical values given difference in scale	Numerical features
2. One Hot Encoder	Dummify non numerical columns	Categorical features
3. SMOTE	Address class imbalance by creating synthetic samples of the diabetic class	All feature columns
4. SelectFromModel	Reduce dimensionality and choose top X features that most likely predict propensity for diabetes	All feature columns

Modelling & Results

Results: LightGBM delivered the best overall AUC and is our model of choice

- 5 different classifiers were trained and tuned
- Overall best-performing model was LightGBM, which has **balanced performance** as seen through maximum AUC
- LightGBM also has the maximum accuracy and precision overall, with only a small tradeoff in terms of recall



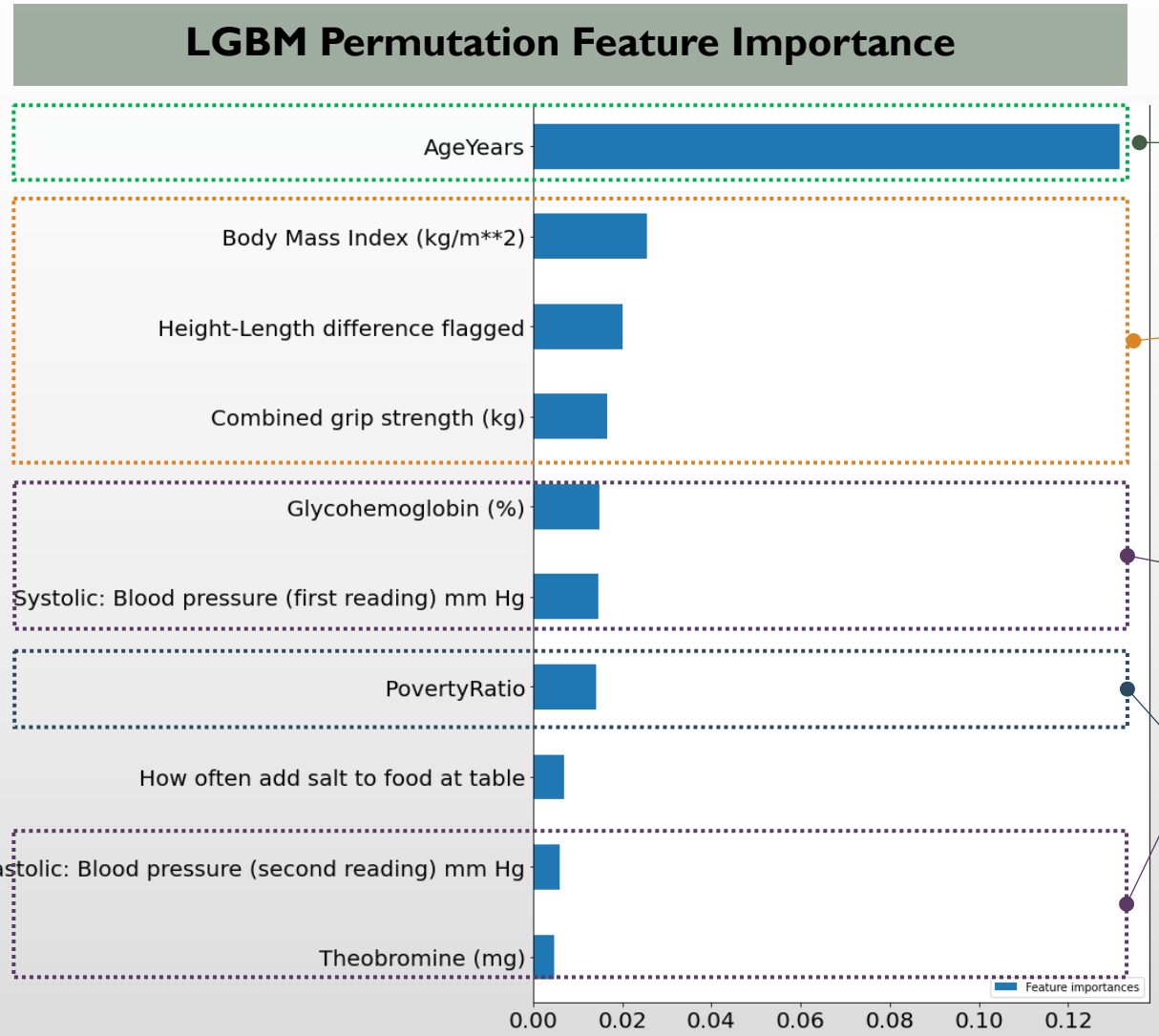
Test Results	AUC	Accuracy	Precision	Recall
Log Reg (baseline)	0.89	0.80	0.65	0.87
Linear SVC	0.88	0.80	0.63	0.88
Decision Tree	0.83	0.81	0.68	0.78
Random Forest	0.91	0.82	0.69	0.84
LightGBM	0.93	0.85	0.78	0.77

Understanding Diabetes Risk Factors

Explainability of prediction is critical for stakeholders to understand risk factors and sense check outcomes

- Top performing models (LightGBM) are relatively **opaque**
- Further techniques need to be applied to allow stakeholders (patients, doctors, insurers) to understand and trust predictions
- This can also shed light on what factors are associated with diabetes risk in general

Permutation Feature Importance highlights that Age is the most important feature



Age is the #1 most important feature

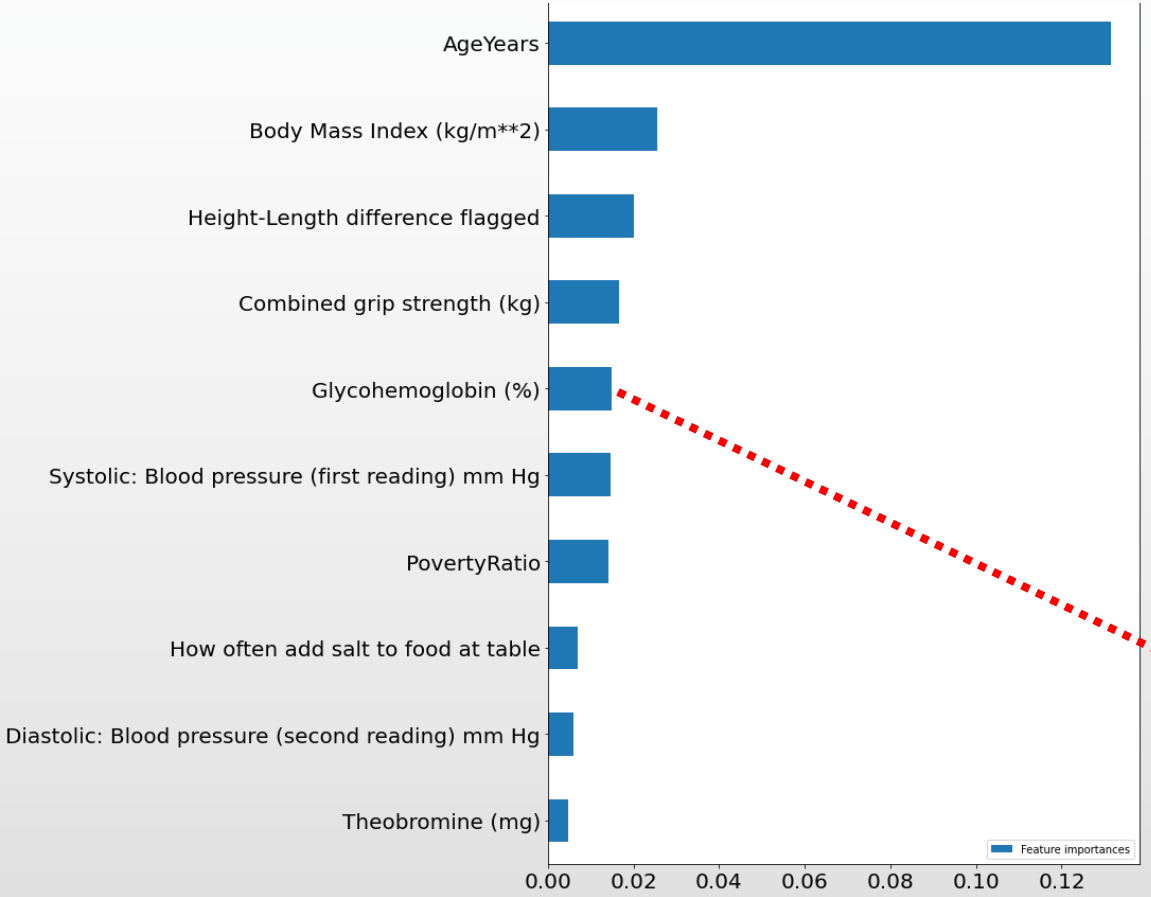
Weight and physical health indicators come out top

Out of the many patient blood test features, only **blood sugar level** (glycohemoglobin), blood pressure and theobromine levels are in the top 10

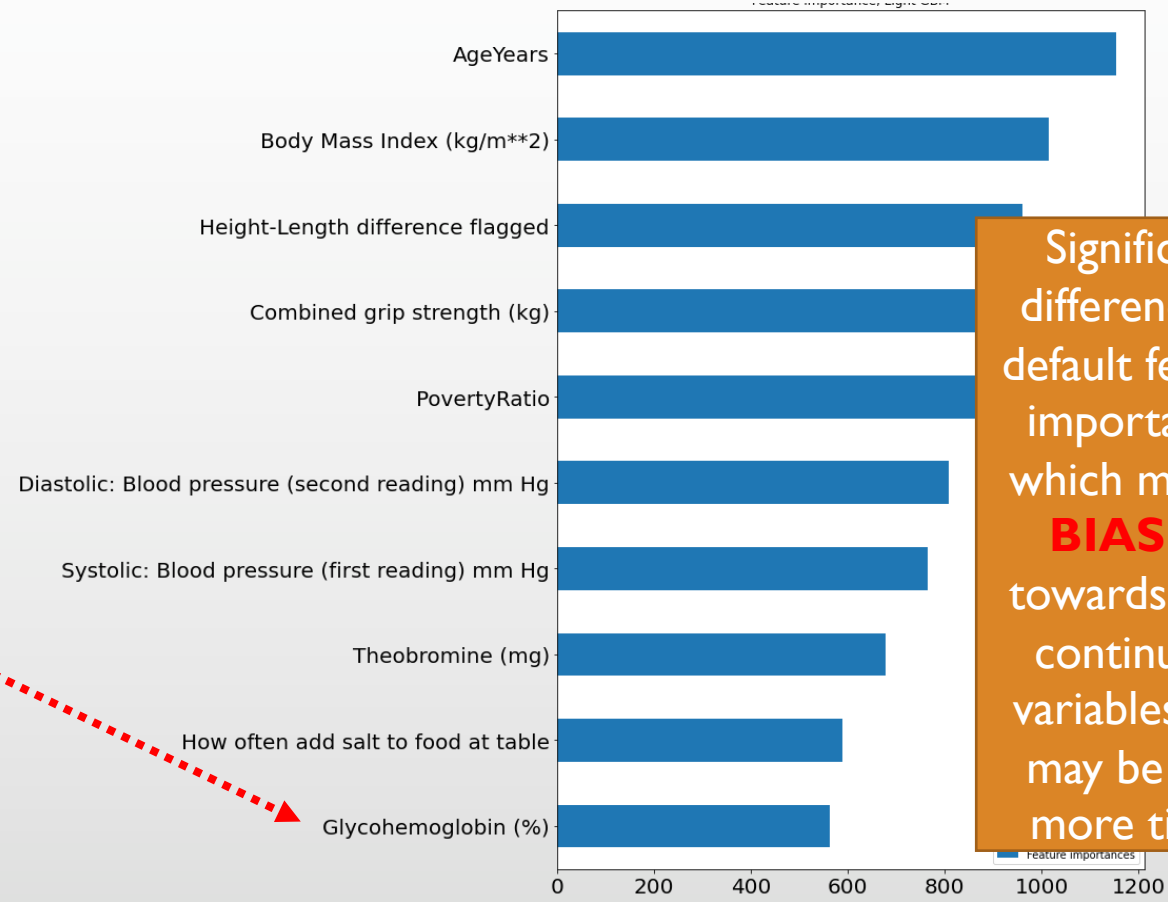
PovertyRatio is the only socio-economic factor in the top 10

Permutation Feature Importance gives more reliable results compared to default feature importance

LGBM Permutation Feature Importance



LGBM Default Feature Importance ('Splits')



Significant difference vs default feature importance, which may be **BIASED** towards large continuous variables that may be split more times

Global SHAP values add further explainability to the direction of feature relationships

Increased probability of diabetes if…

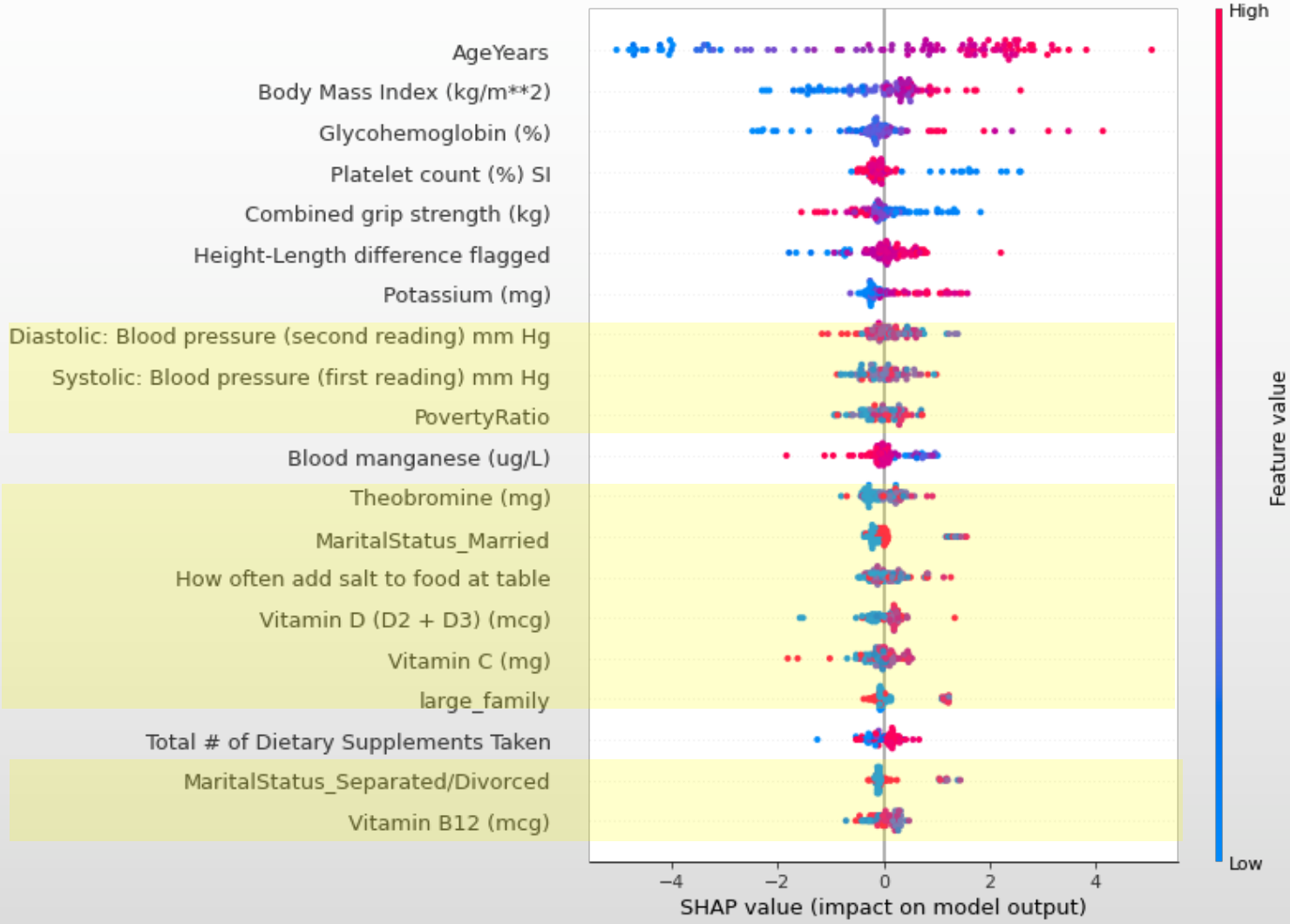
- Patient is older
- Patient is on the heavier side
- Patient has high blood sugar
- Patient has lower readings on platelets and blood manganese
- Patient has low grip strength
- Patient has higher readings on potassium
- Patient has a high height-to-leg length disparity (i.e. short stature)
- Patient takes more supplements

Global SHAP values are mixed for certain important features and local values are needed instead

However, many relationships are not as easy to visualize such as…

- Blood pressure
- Impact of poverty
- Eating habits
- Marital Status

Hence, local explainability (at an individual diagnosis level) is needed

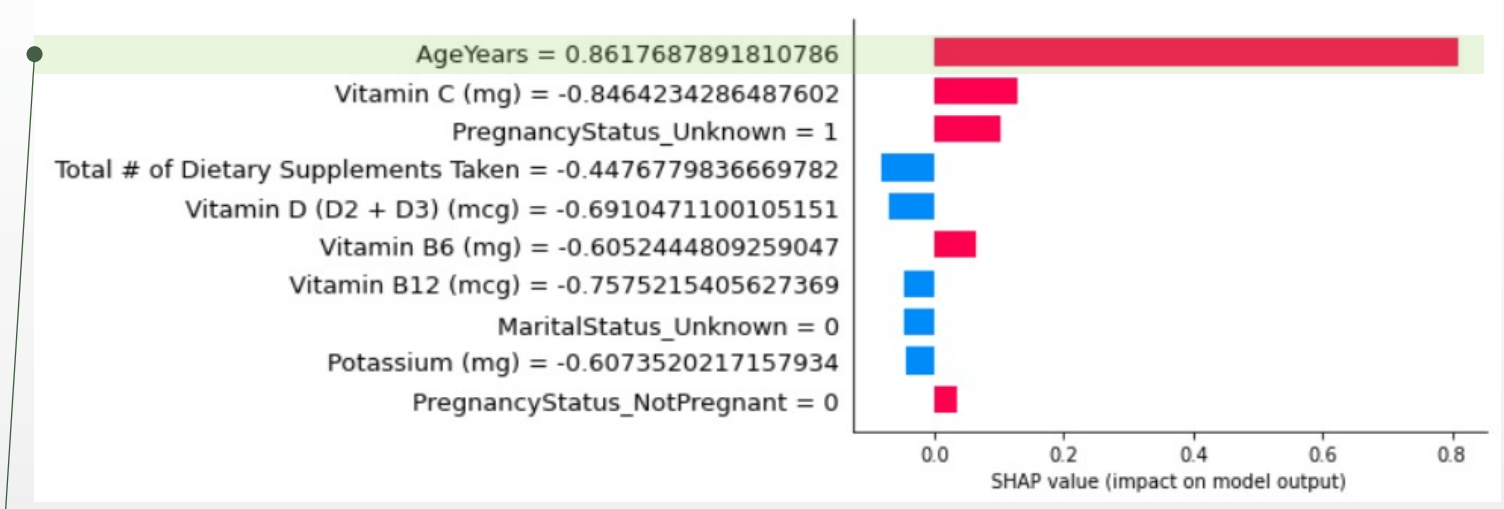


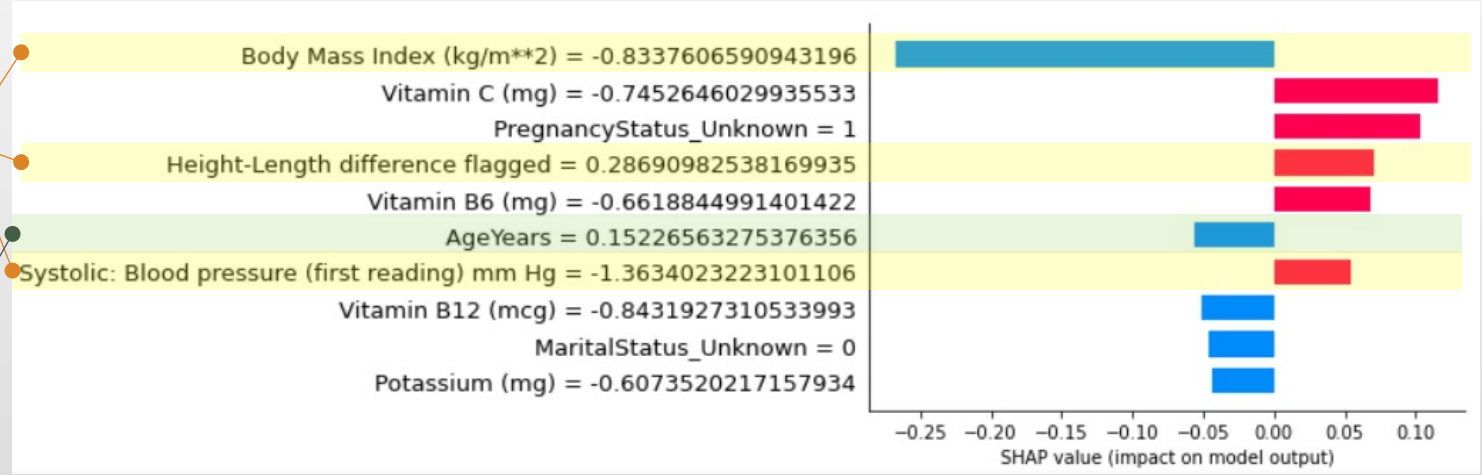
Local Explainability: Case Study with SHAP

- Top features **age years** and **marital status** have local impact
- Top global features such as **platelet count** and **waist circumference** are locally unimportant
- Two local samples may have very different reasons for prediction

Globally ambiguous/unimportant features may have local importance

The same feature have opposite impacts in different predictions





Local Explainability: Case Study SHAP vs LIME

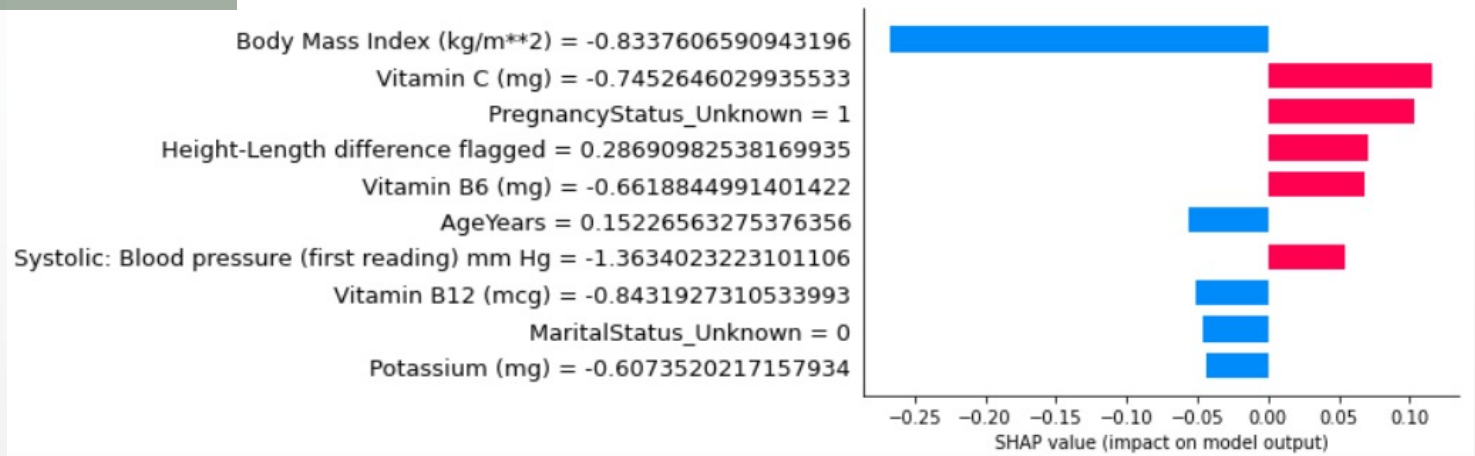
Similarities

- All features have the same direction of contribution
- Most features identified as important are the same. This person has:
  - *A good BMI*
  - *Abnormal levels of vitamin C*
  - Other abnormal values for not being pregnant
- **Body mass index** is mentioned as a main indication of diabetes by both models

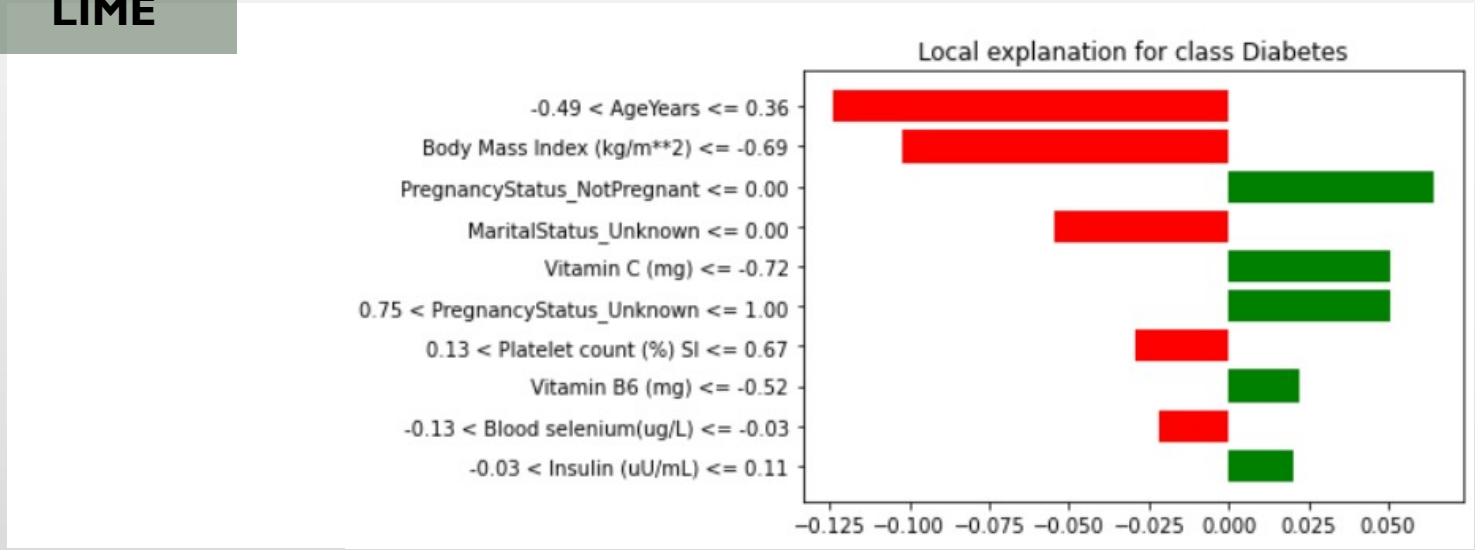
Differences

- Few low importance features vary (*Systolic Blood P, Potassium, Height-Length*)
- **Age years** given more importance by LIME

SHAP



LIME



Thank You!