Machine Learning I

Diabetes Screening

Interactive Predictive Modelling with Streamlit

A Logistic Regression Application

Group 8: Omer Althwaini, Emiliano Puertas, Joel James Alarde, Maria do Carmo Abreu and Africa Bajils

1. Problem Statement
2. Objectives
3. Our Approach & Data Preparation
4. Model Training & Evaluation
5. User Interface
6. Findings & Conclusions

PROBLEM STATEMENT

The Problem

A Hospital approached us with an issue in their Diabetes Diagnosis Process:

“Our Diabetes Screening Procedures are inconsistent, regularly delayed or missing and time and resource consuming.”

OBJECTIVES

Our Goals

01  Short-Term

Improve the overall Diabetes Diagnosis process in a fast and easy way

02  Mid-Term

Continuous Learning & Optimisation

03  Long-Term

Wide-adoption from Hospitals as the first step in Diagnosis

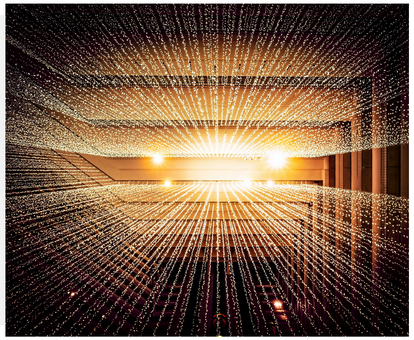
1. Problem Statement
2. Objectives
3. Our Approach & Data Preparation
4. Model Training & Evaluation
5. User Interface
6. Findings & Conclusions

OUR APPROACH

The Process

1. Data Preparation & Preprocessing

Handling Outliers, Dealing with Missing Values, Data Type Handling and Correlation Analysis.
Data Set



2. Model Development

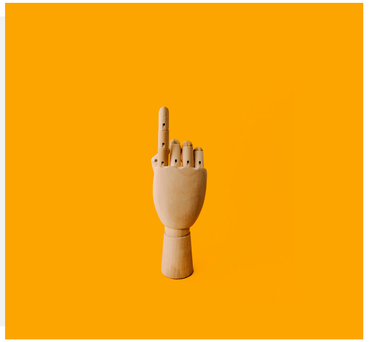
Dataset Splitting, Model Selection, in this case, Logistic Regression as our classification model as we learned in class. And Feature Importance Evaluation.

3. Model Evaluation

Performance metrics: compared accuracy on training, testing, and overall data. Used classification reports and accuracy scores. Final Model Optimisation: retrained the model, improved generalisation and avoided overfitting.

4. User Interface

Created a Streamlit web app where users can input patient data. Using the classification model, the interface returns a prediction on wether the patient has diabetes or not.



OUR APPROACH

Diabetes patient data

Data Set

Index
A unique identifier for each row

Pregnancies
Number of times the patient has been pregnant.

Blood Pressure
Patients blood pressure measurements.

Age
Patient's age in years.

Glucose
Blood glucose level (mg/dL) measured during an oral glucose tolerance test.

Skin Thickness
Thickness of skin fold at the triceps (mm), used as an indirect measure of body fat.

Insulin
Insulin level, which can indicate how well the body regulates blood sugar.

BMI
Body Mass Index, an indicator of body fat.

Diabetes Pedigree Function
A score that estimates genetic predisposition to diabetes based on family history.

Outcome
The target variable (0 = No diabetes, 1 = Diabetes).

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   index                     768 non-null    int64
 1   Pregnancies               768 non-null    int64
 2   Glucose                   768 non-null    int64
 3   BloodPressure             768 non-null    int64
 4   SkinThickness             768 non-null    int64
 5   Insulin                   768 non-null    int64
 6   BMI                       768 non-null    float64
 7   DiabetesPedigreeFunction  768 non-null    float64
 8   Age                       768 non-null    int64
 9   Outcome                   768 non-null    int64
dtypes: float64(2), int64(8)
```

OUR APPROACH

Data Preparation & Preprocessing

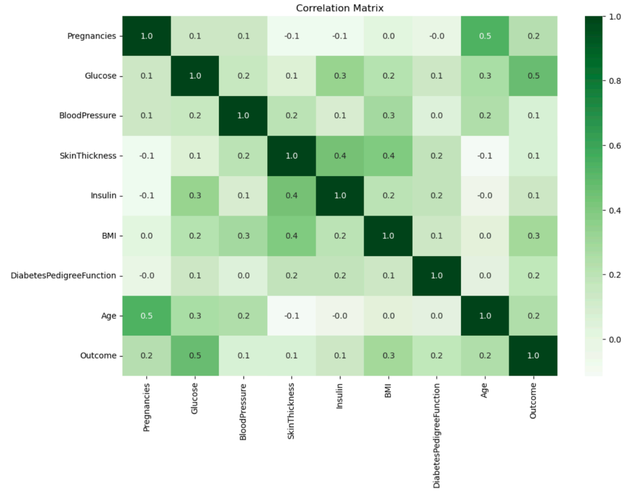
Feature	Lower Limit	Upper Limit
Age (years)	0 (Newborn)	122 (Oldest recorded human)
BMI	12–14 (Starvation)	100+ (Extreme obesity)
Blood Pressure (mmHg)	50/30 (Shock)	300/200 (Extreme hypertension)
Diabetes Pedigree Function	0 (No family history)	2.5+ (Very high risk)
Glucose (mg/dL)	40 (Severe hypoglycemia)	600+ (Diabetic crisis)
Insulin (µU/mL)	<2 (Type 1 diabetes)	300+ (Severe insulin resistance)
Pregnancies	0	15–20+ (Rare)
Skin Thickness (mm)	~5 (Very lean)	~100 (Extreme obesity)

Dealing with Missing Values

- No null values were found in the dataset.

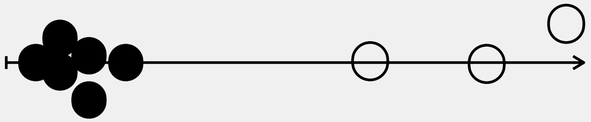
Correlation Analysis

- Identified glucose and BMI as the most correlated features with diabetes. Index didn't show any kind of correlation with other features, so we removed it.
- The other features have a positive correlation with the outcome.



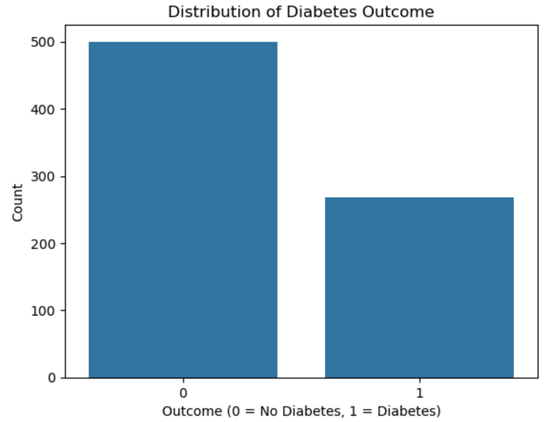
Handling Outliers

- Analysed the different feature value limits.
- Decided not to remove the outliers.



Data Type Handling

- Ensured all columns had appropriate data types. In the Diabetes dataset there were only columns of the type integer or float.
- Also looked into the Outcome distribution. Data with no diabetes is almost double of people with diabetes. Unbalanced dataset.



MODEL TRAINING & EVALUATION

Model Development

Dataset Splitting

Used different Data splits with varying random seeds. Selected the seed that resulted in the best model performance.
Applied hyper-parameter tuning (e.g., adjusting random_state). Random State - to peak the best performing model seed

Normalisation & Standardisation

Applied techniques like Min-Max Scaling and Standard Scaling to selected features and compared best results. Decided to adopt Min-max Scaling, also called Normalisation.

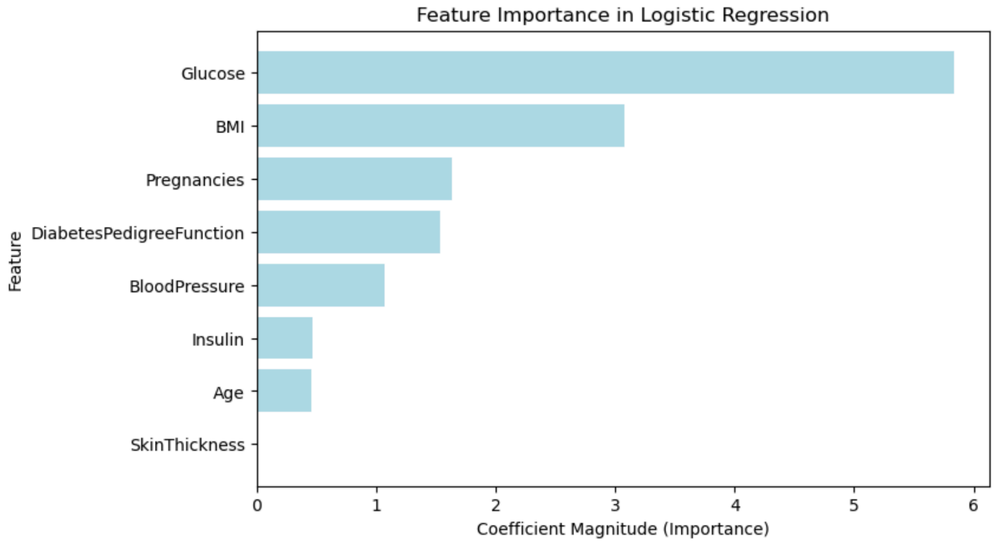
Machine Learning Model

Logistic Regression as the primary classification model. Tested different model splitting approaches, but opted to keep an 80% Training and 20% Validation model.

Regularisation

Initially included all features and analysed feature importance.

Used **GridSearchCV** to refine feature selection and to avoid overfitting.



Model Evaluation

Model Performance

Compared accuracy on **training, testing, and overall data**.

The model is doing very well at identifying people that are not diabetic.

Better performance prediction No-Diabetes vs Diabetes

Final Model Optimisation

The model able to generalize numbers

No overfitting

Consistent outcomes

- **Model Performance in Testing data**

```
Accuracy Rate = 0.8246753246753247
Best Hyperparameters: {'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}
              precision    recall  f1-score   support

Non Diabetic       0.81      0.95      0.88       101
    Diabetic       0.86      0.58      0.70        53

    accuracy                           0.82       154
   macro avg       0.84      0.77      0.79       154
weighted avg       0.83      0.82      0.81       154
```

- **Model Performance in Training data**

```
Accuracy Rate = 0.7817589576547231
              precision    recall  f1-score   support

Non Diabetic       0.78      0.92      0.85       403
    Diabetic       0.78      0.51      0.62       211

    accuracy                           0.78       614
   macro avg       0.78      0.72      0.73       614
weighted avg       0.78      0.78      0.77       614
```

1. Problem Statement
2. Objectives
3. Our Approach & Data Preparation
4. Model Training & Evaluation
5. User Interface
6. Findings & Conclusions

USER INTERFACE

Diabetes Diagnosis Streamlit App

Scan Me!

Link: https://mligroupassignment-bnfqradu7voomwttebw344.streamlit.app/

1. Problem Statement
2. Objectives
3. Our Approach & Data Preparation
4. Model Training & Evaluation
5. User Interface
6. Findings & Conclusions

OBJECTIVES

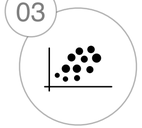
Our Goals & Conclusions

01  Short-Term

Improve the overall Diabetes Diagnosis process in a fast and easy way

02  Mid-Term

Continuous Learning & Optimisation

03  Long-Term

Wide-adoption from Hospitals as the first step in Diagnosis

Our Outcome

Deploy the ML Model. Automate higher risk patients prioritisation. Set up performance monitoring.

Time and cost efficiency improvements.

Future Recommendations

Connect the model with the hospital's EHR system for real-time screening.

Expand available datasets and continuous model development.

Implement the model in all hospital departments and expand to other institutions. Use model insights to shape public health guidelines for diabetes screening.

Thank you!