Diabetes Prediction Web App - Project Report

1. Dataset Description and Selection Rationale

Dataset: Pima Indians Diabetes Dataset
Features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age
Target: Outcome (0 = Non-Diabetic, 1 = Diabetic)
Rationale: Widely used benchmark dataset for binary classification in medical diagnosis, small enough for quick model training yet representative of real-world diabetes screening.

2. Data Preprocessing Steps Taken

- Handled missing values: Zero values in physiological measurements considered missing.
- Feature Scaling: Applied StandardScaler to numerical features.
- Train/Test Split: 80/20 ratio.
- Encoding: Not required as all features are numerical.

3. Model Selection and Evaluation Process

Final Model: Logistic Regression
Rationale: Interpretable, efficient, performs well on small datasets.
Evaluation Metrics: Accuracy, Confusion Matrix, Precision, Recall, F1-score. Accuracy: ~0.78 on the test set.

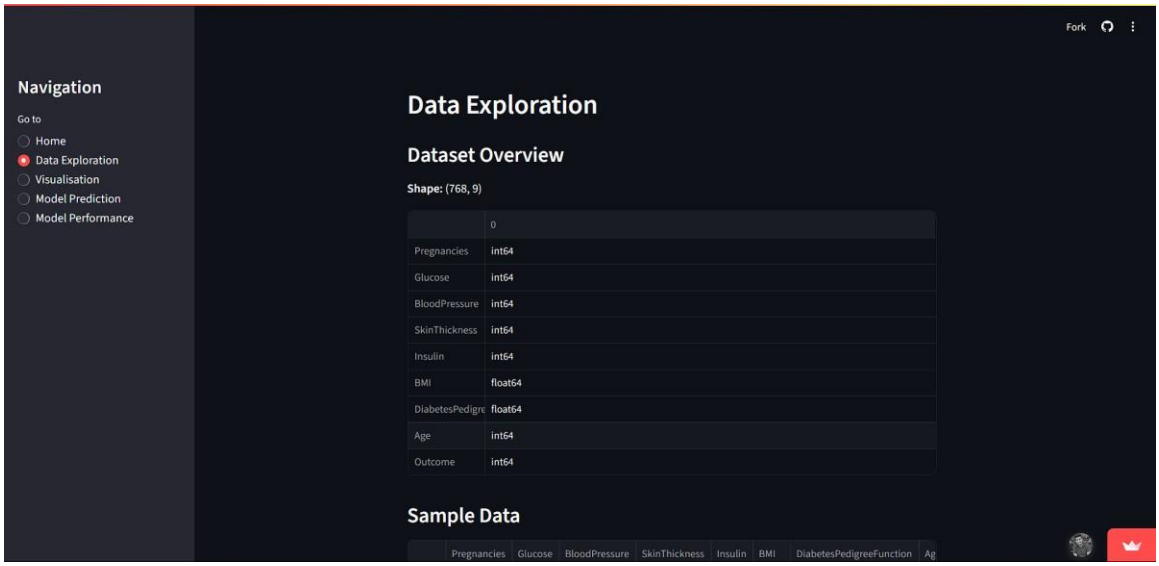
4. Streamlit App Design Decisions

- Sidebar navigation with sections: Home, Data Exploration, Visualisation, Model Prediction, Model Performance.
- Visualisations: Outcome histogram, correlation heatmap, BMI vs Glucose scatter plot.
- Prediction Interface: Numeric inputs for all features, prediction label, and probability.
- Performance Page: Displays accuracy, confusion matrix, classification report.

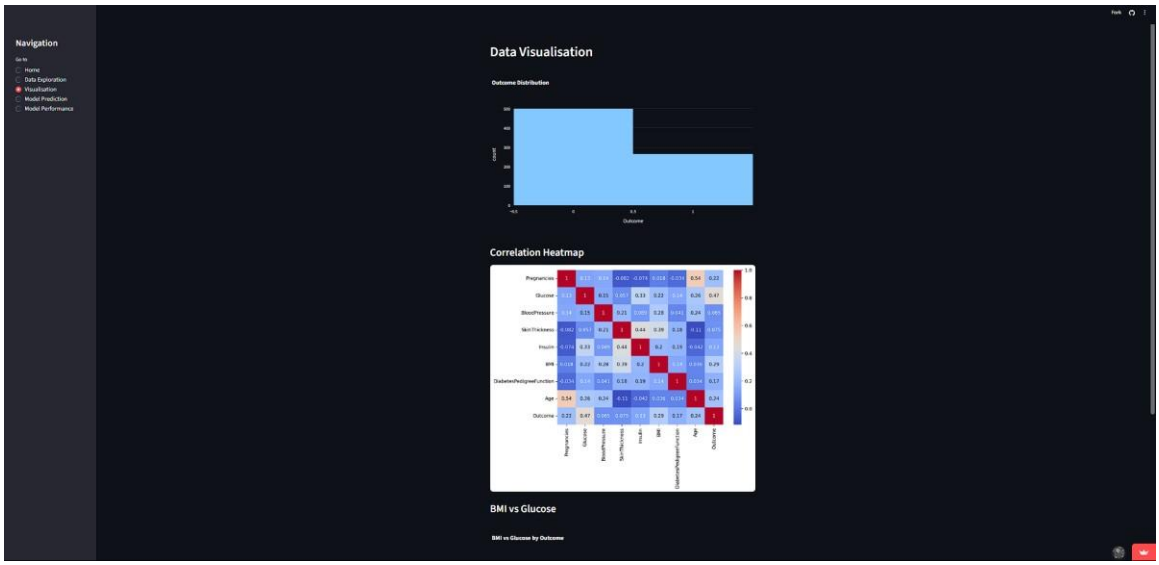
5. Deployment Process and Challenges Faced

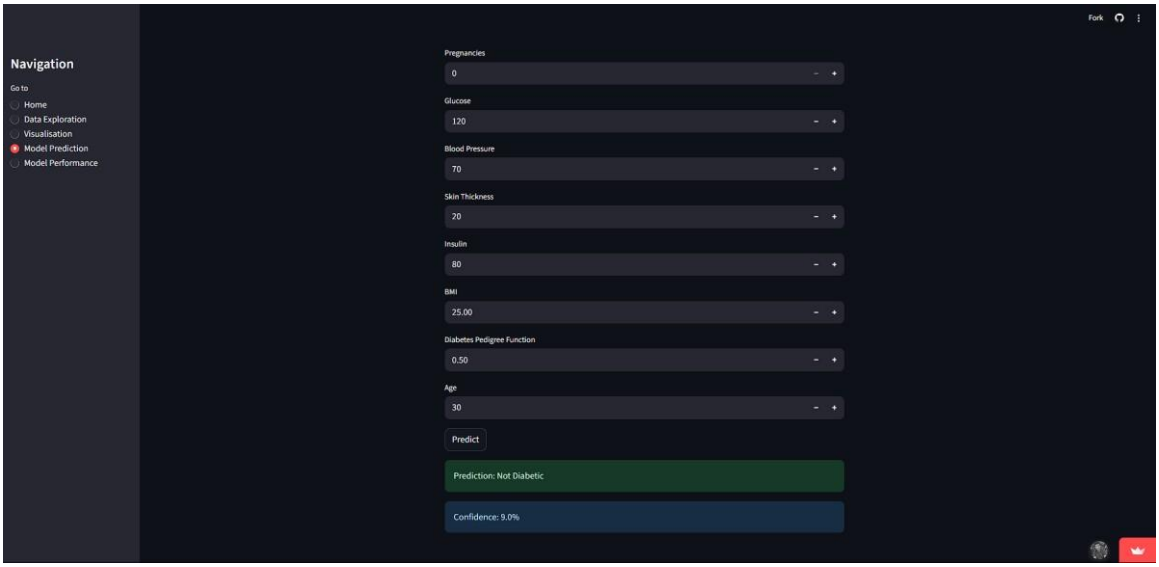
- Hosting: Streamlit Cloud.
- Challenges:
* Managing pickle model loading.
* Ensuring consistent scaling in prediction and training.
* Handling dependency issues in deployment.

6. Screenshots of the Application




Navigation
Go to
Home
Data Exploration
Visualisation
Model Prediction
Model Performance
Fork
Data Visualisation
Outcome Distribution
count
Outcome
Correlation Heatmap


Navigation
Data Visualisation
Outcome Distribution
Correlation Heatmap
BMI vs Glucose
BMI vs Glucose by Outcome



7. Reflection on Learning Outcomes

- Learned the end-to-end ML pipeline: data loading, preprocessing, model training, evaluation, and deployment.
- Gained experience with Streamlit for interactive apps.
- Understood importance of feature scaling and model interpretability.
- Improved UI design for medical decision support tools.

8. Links

Streamlet App: https://diabetesprediction2.streamlit.app/

GitHub Repository: https://github.com/Dulanga917/House