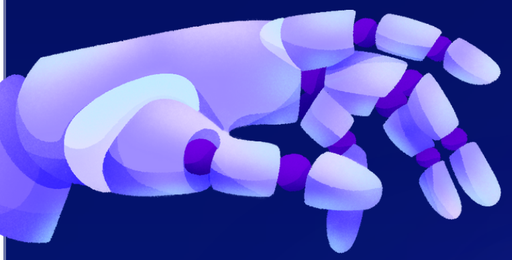
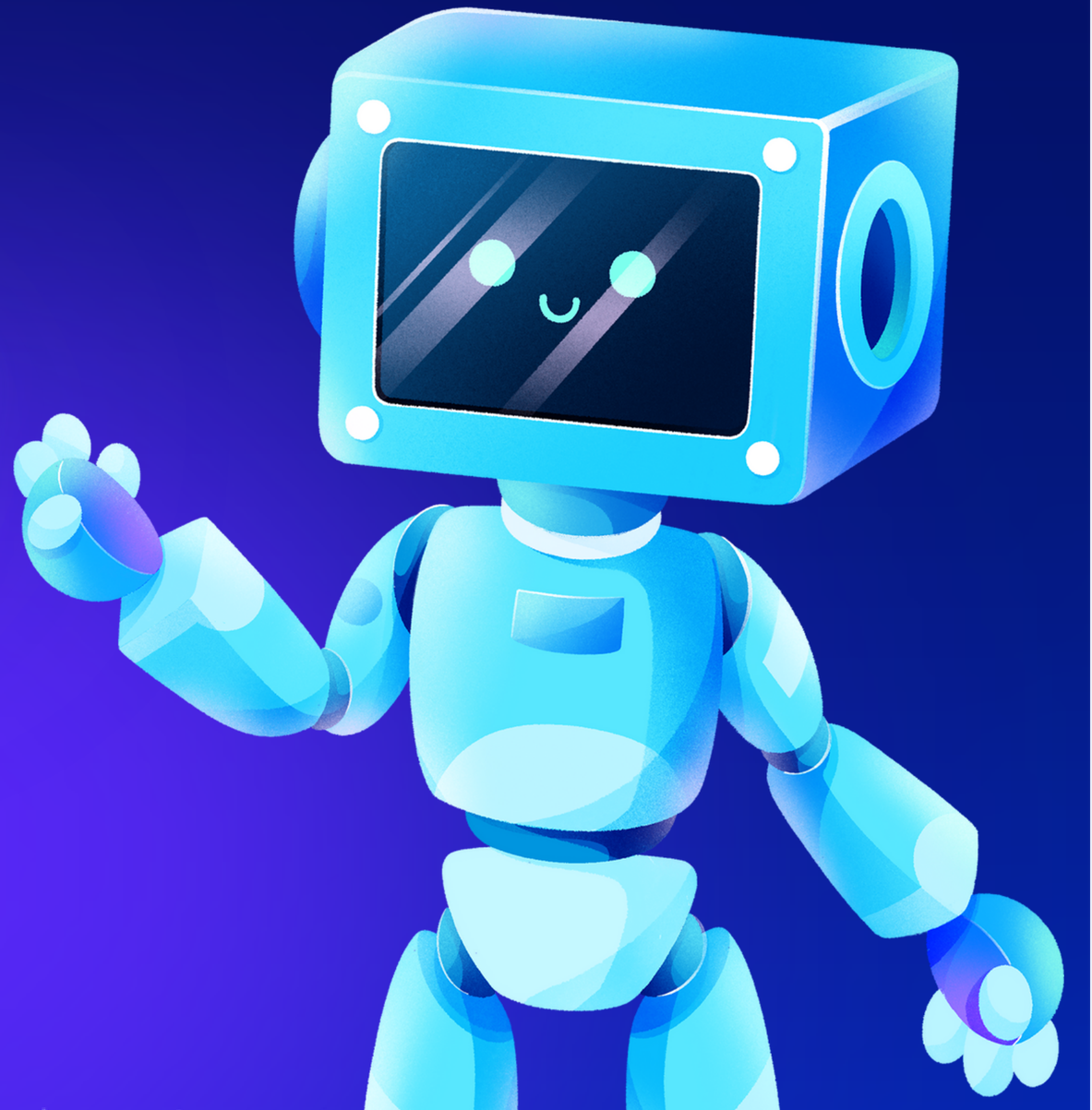
DIABETES PREDICTION

PROJECT

By Hazem Karoui





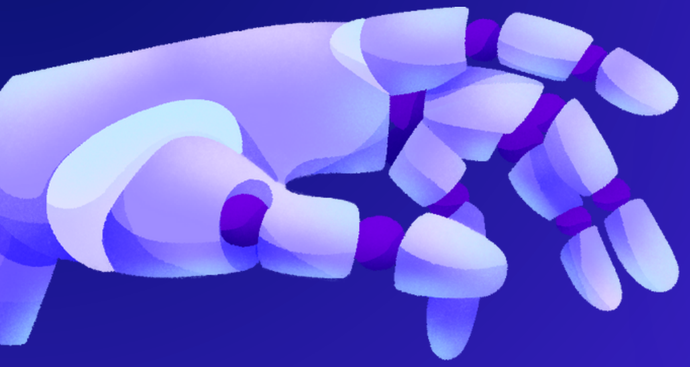
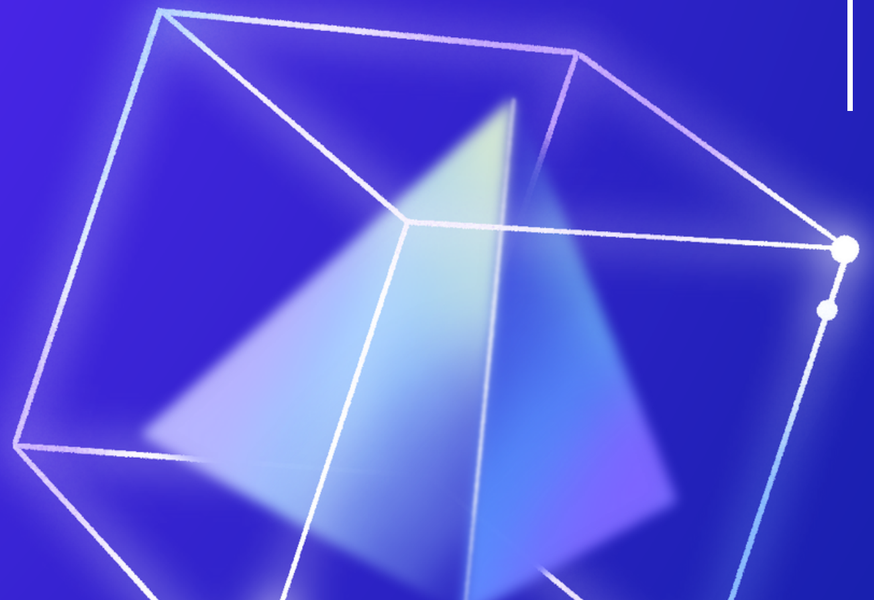


TABLE OF CONTENTS

- Introduction 01
- Dataset Overview 02
- Data Cleaning & Visualization 03
- Model Training & Evaluation 04
- Model Comparison 05
- Web App 06
- Conclusion 07



INTRODUCTION

Predicting diabetes is a critical task with far-reaching implications for healthcare. Early detection is key to effective management, prompting our exploration into data science solutions. In this presentation, we'll discuss our efforts to develop a practical diabetes prediction model. The goal is to leverage data-driven insights for timely intervention and empower individuals in managing their health proactively.

DATASET OVERVIEW

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
2	138	62	35	0	33.6	0.127	47	1
0	84	82	31	125	38.2	0.233	23	0
0	145	0	0	0	44.2	0.630	31	1
0	135	68	42	250	42.3	0.365	24	1
1	139	62	41	480	40.7	0.536	21	0

Statistics:

- Count: 2000 entries
- Mean (Age): 33.09
- Mean (Glucose): 121.18
- Outcome (Diabetic): 34.2%

DATA CLEANING & VISUALIZATION

HANDLING MISSING VALUES

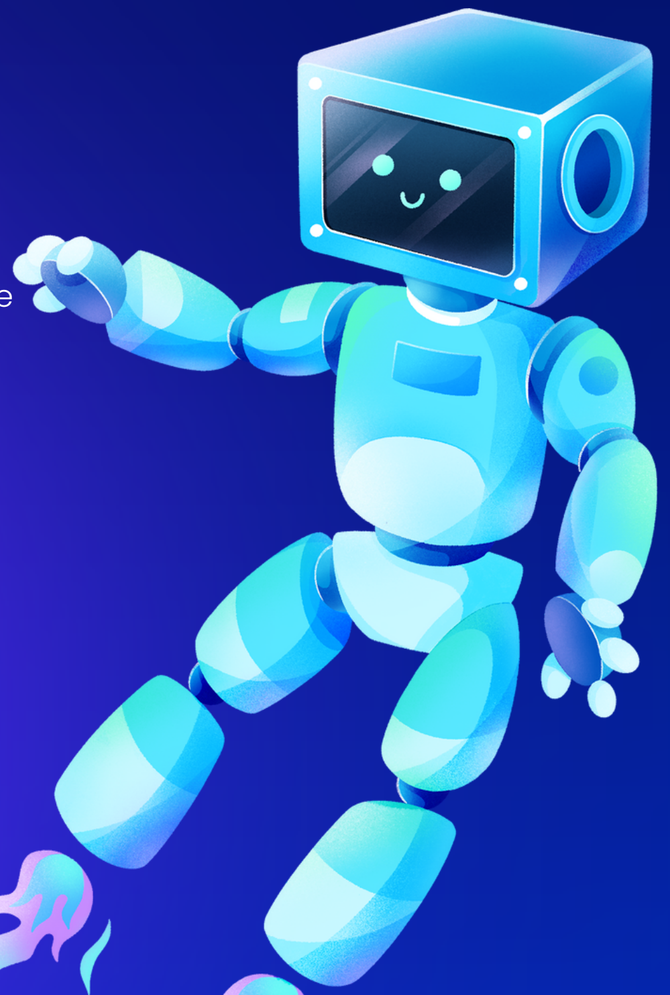
- Identified and addressed missing values using imputation techniques.
- Ensured completeness and accuracy in the dataset.

OUTLIER REMOVAL

- Utilized the Interquartile Range (IQR) method to identify and remove outliers.
- Enhanced the model's resilience to noise and irregularities.

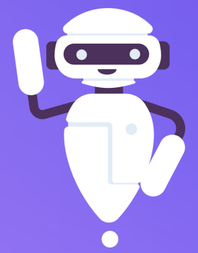
ZERO VALUES

- Investigated and addressed instances of zero values, ensuring data consistency.
- Applied domain knowledge to distinguish genuine zero values from data entry errors.



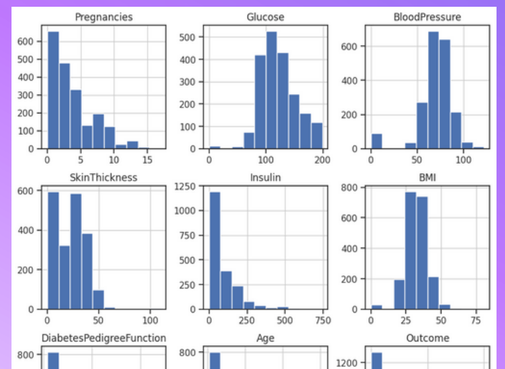
DATA VISUALIZATION

Using Seaborn and Matplotlib



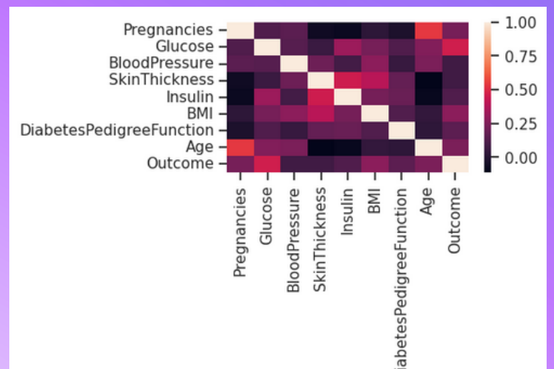
HISTOGRAMS

Examined distributions of key features (e.g., Glucose, BMI) to understand data patterns.



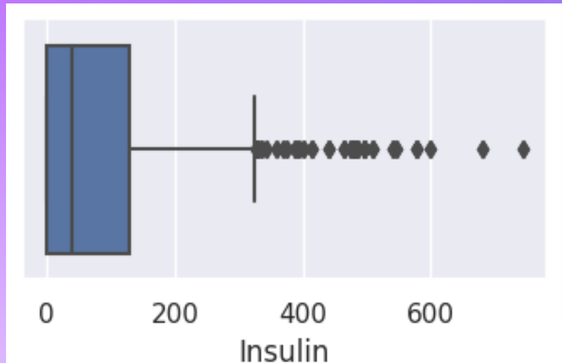
HEATMAPS

Explored correlations among variables, identifying potential multicollinearity.



BOX PLOTS

Visualized the distribution of features, identifying outliers in variables like Insulin and Blood Pressure.



MODEL TRAINING & EVALUATION

LOGISTIC REGRESSION

Overview

Linear model for binary classification.

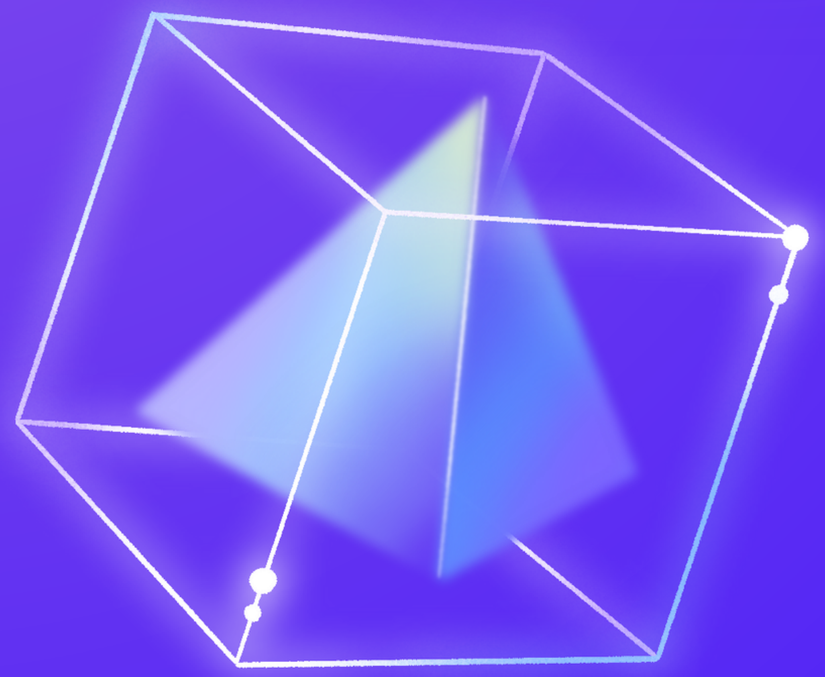
Training Process

Fitted the model on the training set.

Results

- Accuracy: 0.788
- ROC: 0.715

SUPPORT VECTOR MACHINE



Overview

Non-linear model for classification.

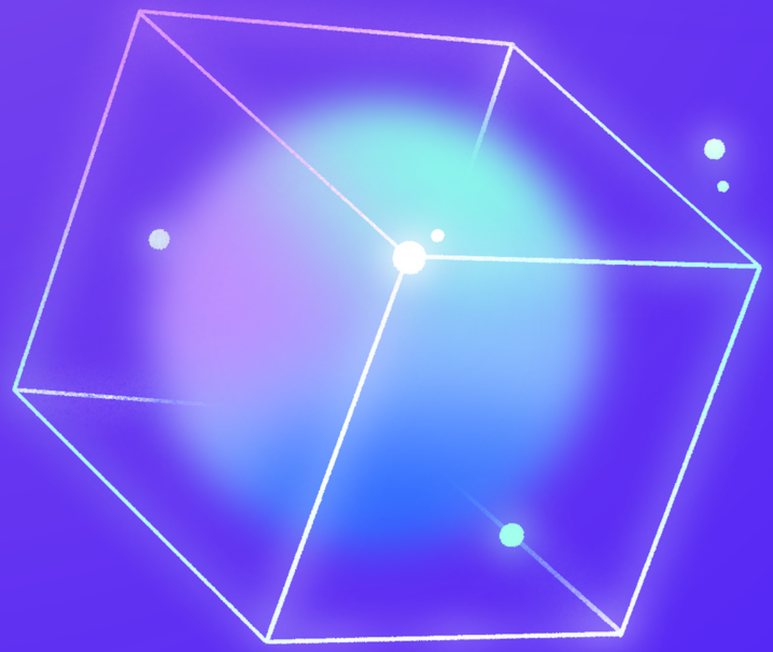
Training Process

Utilized a linear kernel.

Results

- Accuracy: 0.773
- ROC: 0.696

K-NEAREST NEIGHBORS



Overview

Instance-based learning for classification.

Training Process

Trained with three nearest neighbors.

Results

- Accuracy: 0.906
- ROC: 0.888

RANDOM FOREST

Overview

Ensemble learning method using decision trees.

Training Process

Constructed a forest of trees during training.

Results

- Accuracy: 0.978
- ROC: 0.979

NAIVE BAYES

Overview

Probabilistic model using Bayes' theorem.

Training Process

Assumed independence between features.

Results

- Accuracy: 0.764
- ROC: 0.715

GRADIENT BOOSTING CLASSIFIER

Overview

Boosted ensemble method combining weak learners.

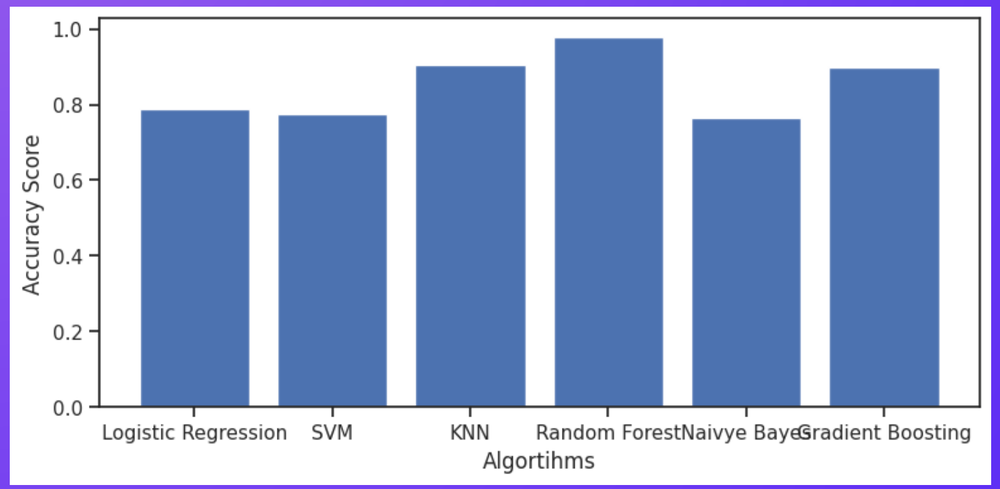
Training Process

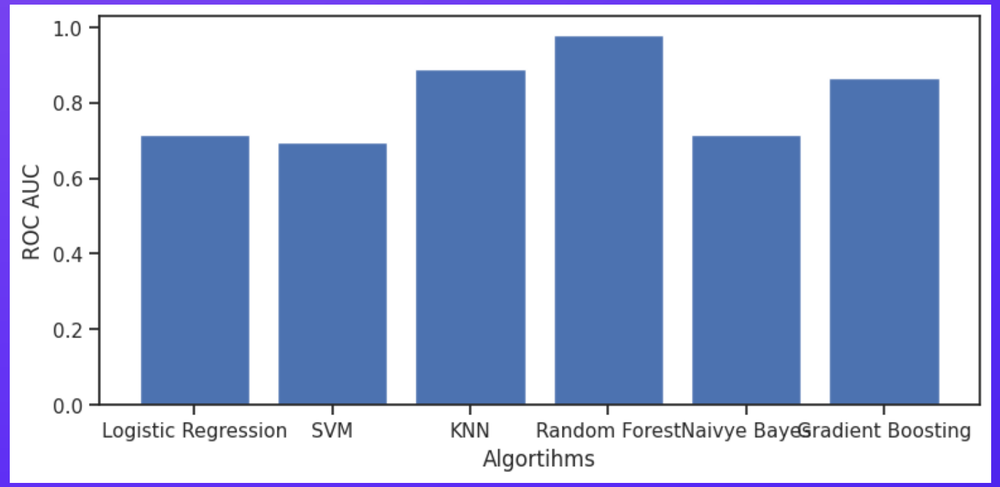
Iterative training of weak learners.

Results

- Accuracy: 0.897
- ROC: 0.866

MODEL COMPARISON





COMPARISON

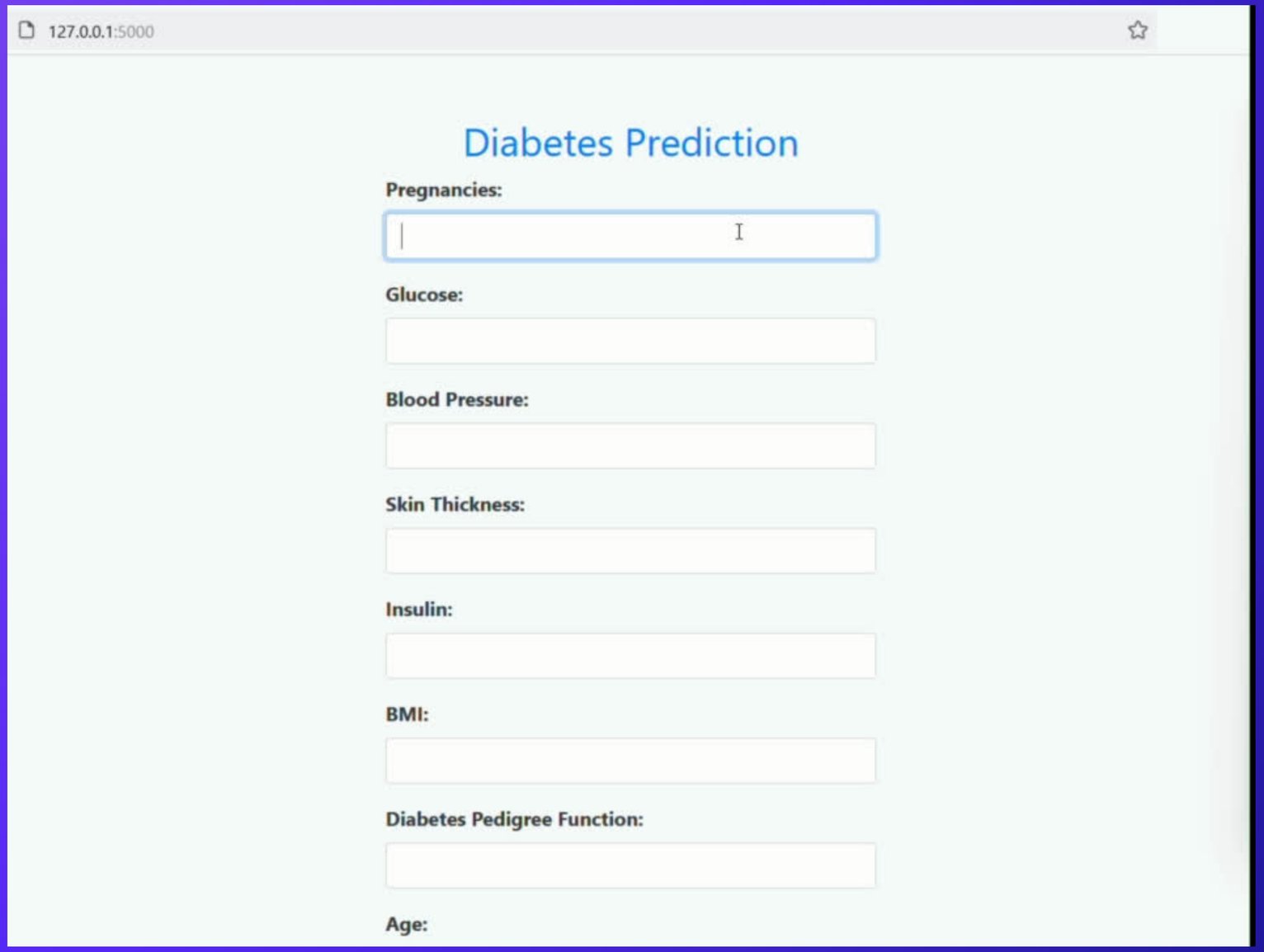
- We evaluated performances across all algorithms using Accuracy Score, ROC AUC Curve, and Confusion Matrix.
- Random Forest achieved the highest accuracy (98%) and robust ROC AUC curve (97%) which is why we chose it as the final model.

WEB APP

WEB APP IMPLEMENTATION

- Developed a Flask web app for diabetes prediction.
- Implemented input fields for relevant parameters (e.g., Glucose level, BMI, Age).
- Integrated the pre-trained Random Forest model for real-time predictions.

Demo


127.0.0.1:5000
Diabetes Prediction
Pregnancies:
Glucose:
Blood Pressure:
Skin Thickness:
Insulin:
BMI:
Diabetes Pedigree Function:
Age:

THANK YOU!