Diabetes Detection with Data Analytics

This project aims to predict diabetes using machine learning. We utilized a comprehensive Kaggle dataset. Our goal is early detection and intervention for better patient outcomes.

Data Source and Description

Dataset Overview

We sourced a comprehensive health survey dataset from Kaggle. It contains 100k rows with 16 features. This allows for robust analysis.

Key Features

- Demographics: Age, Gender, Location, Race
- Health Indicators: BMI, HbA1c, Blood Glucose
- Medical History: Smoking, Hypertension, Heart Disease

The 'diabetes' column is our binary target variable.

Data Cleaning and Preprocessing

1

Handle Missing Data

We imputed or removed missing values. About 5% of rows had missing location data. This improved data quality.

2

Feature Engineering

New features were created. Examples include BMI categories and age groups. This enriches the dataset.

3

Data Transformation

Scaling and normalization were applied. This prepares data for algorithms. It ensures consistent feature ranges.

4

Encode Variables

Categorical variables were encoded. One-hot encoding was used. This converts text to numerical formats.

Exploratory Data Analysis (EDA)

Feature Distributions

Histograms and box plots revealed data patterns. We understood value ranges and outliers.

Correlation Analysis

A heatmap showed relationships between features. This identified strong predictors.

Diabetes Prevalence

Analyzed diabetes by age, gender, and race. Identified demographic risk factors.

BMI and HbA1c Insights

Examined BMI and HbA1c levels against diabetes. Discovered critical thresholds.

Model Selection and Training

Algorithm Choice

Logistic Regression, Random Forest, and XGBoost were considered. XGBoost showed superior initial accuracy.

Data Splitting

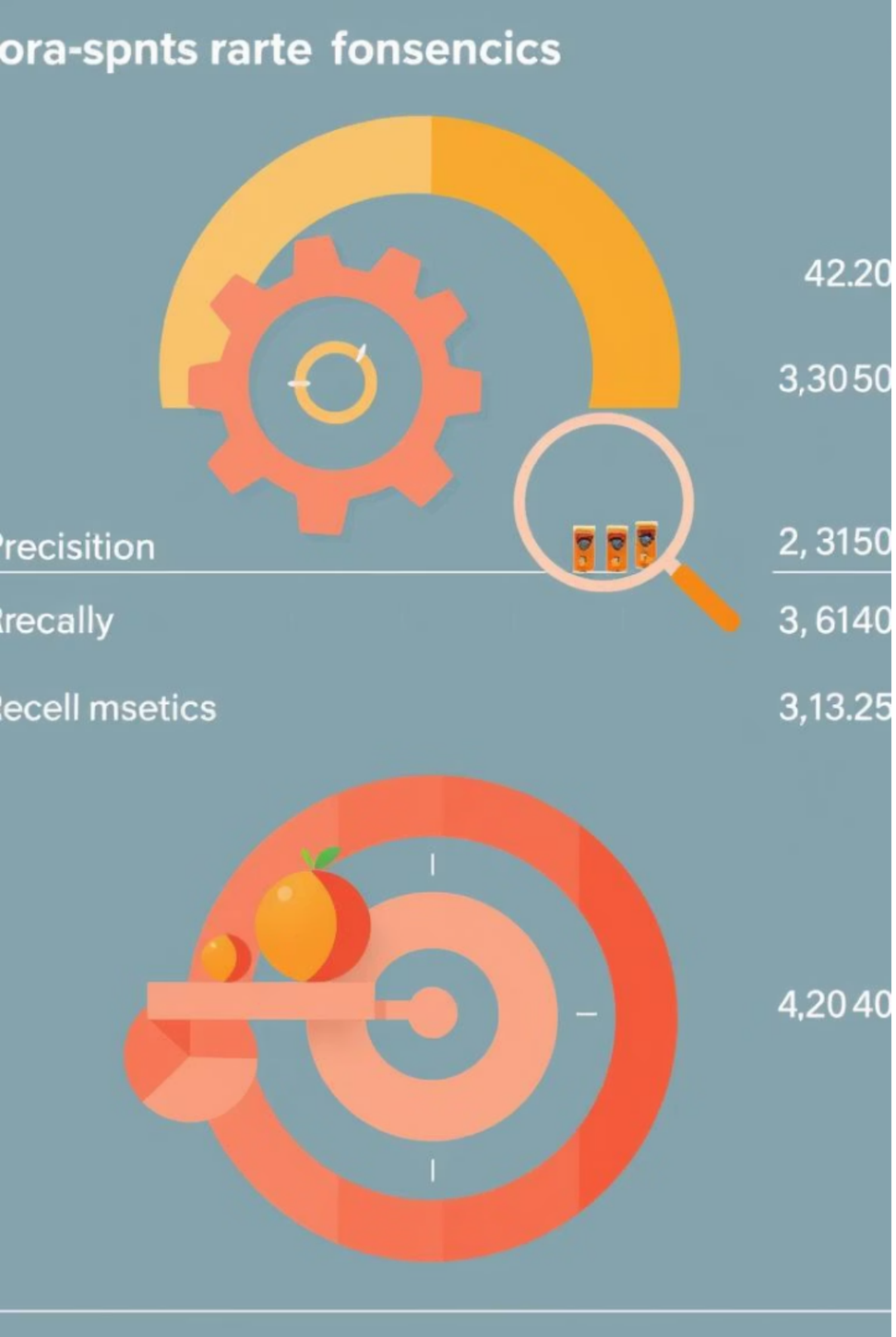
The dataset was split into 80/20 or 70/30 ratios for training and testing. This prevents overfitting.

Cross-Validation

K-fold cross-validation was used. This ensures robust model evaluation. It provides reliable performance estimates.

Hyperparameter Tuning

GridSearchCV and RandomizedSearchCV optimized model parameters. This fine-tuned performance.



Model Evaluation and Results

Evaluation Metrics

Accuracy, Precision, Recall, F1-score, and AUC-ROC were used. These provided a comprehensive view of model performance.

Confusion Matrix

Analyzed true positives, false positives, and other classifications. This helped understand prediction errors.

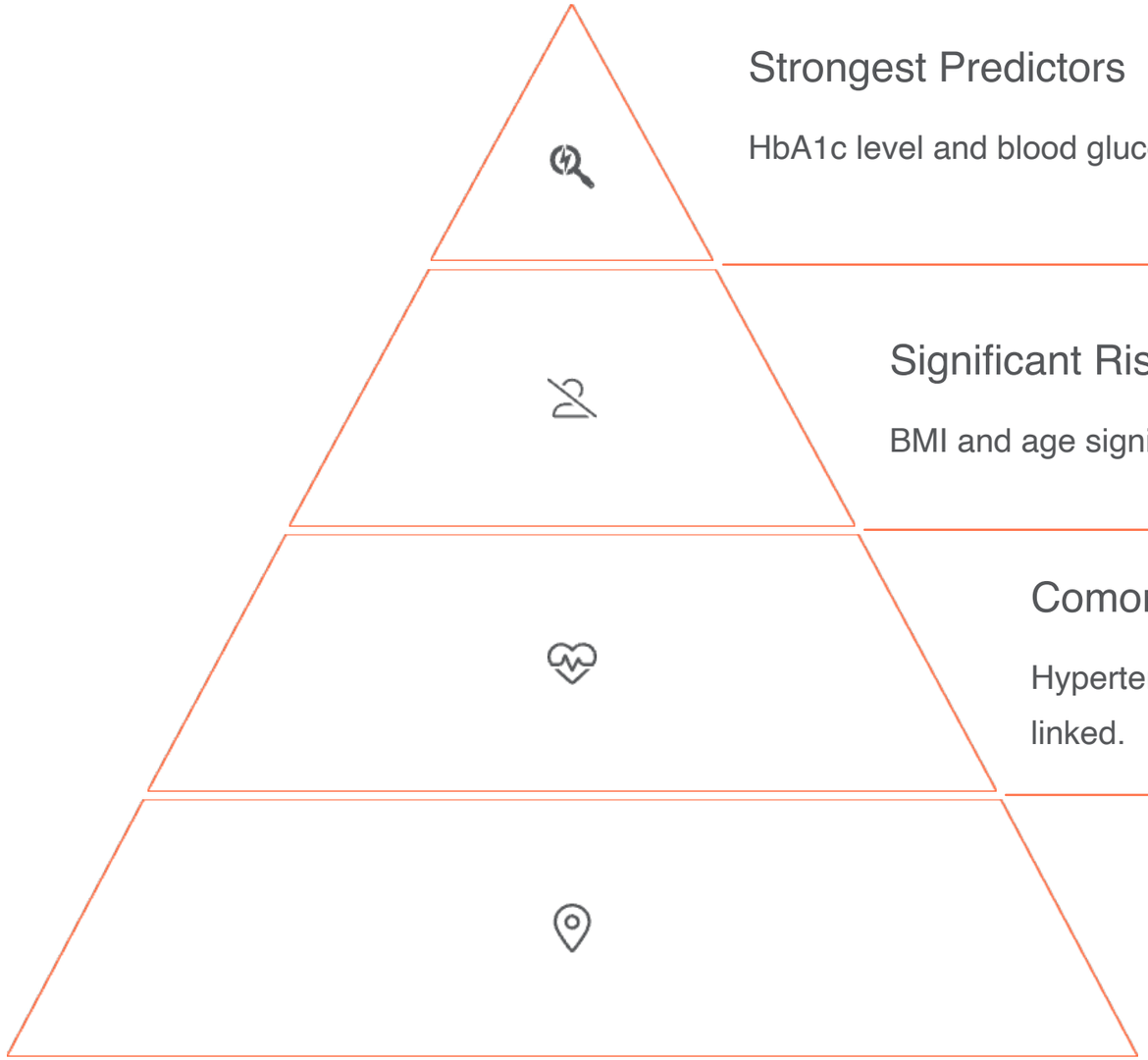
Feature Importance

Identified the most influential predictors. This highlights critical health factors.

ROC Curve

The ROC curve assessed classifier performance across thresholds. An optimized XGBoost model achieved 95% accuracy.

Key Findings and Insights



Strongest Predictors

HbA1c level and blood glucose level are critical indicators. They are primary drivers of diabetes risk.

Significant Risk Factors

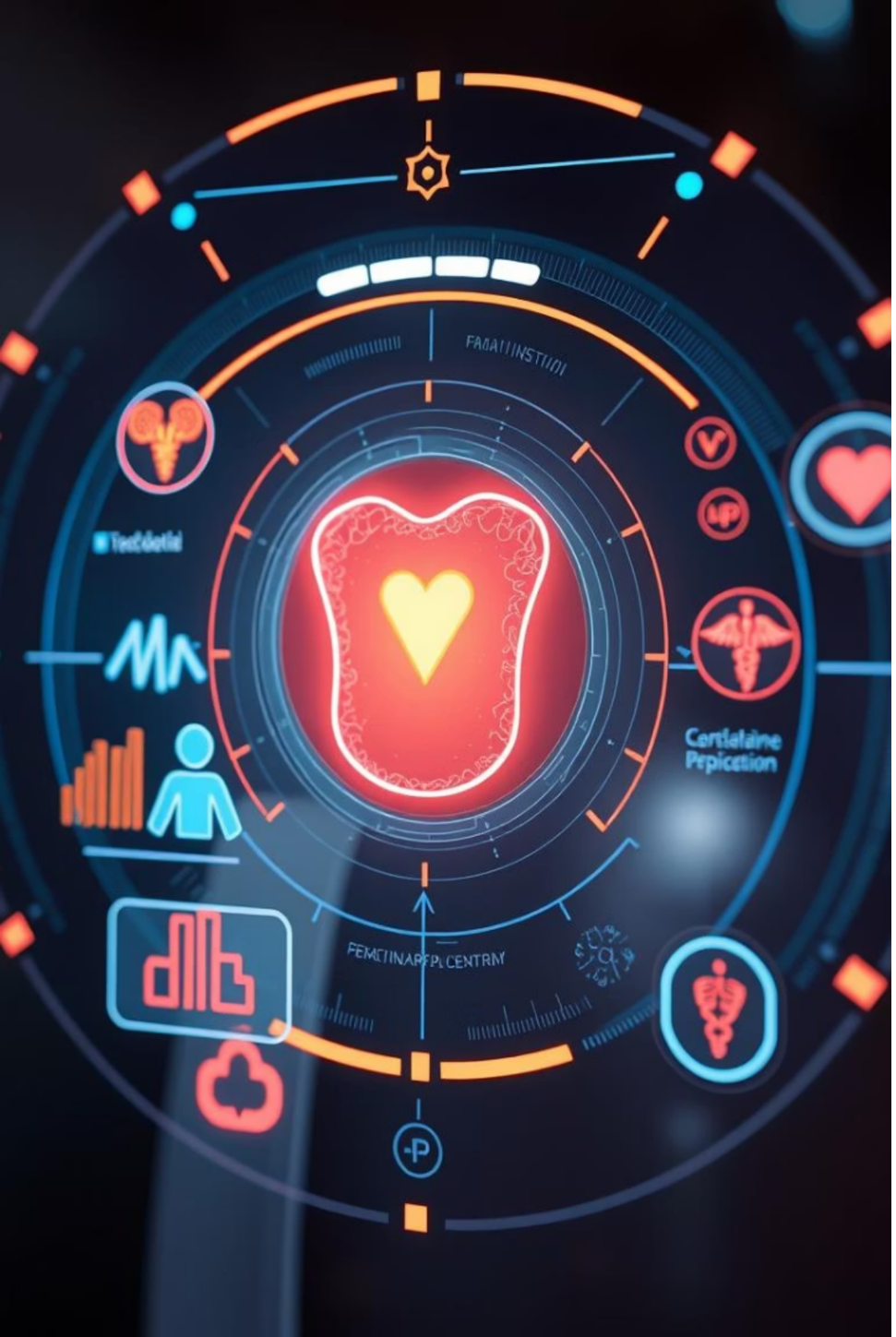
BMI and age significantly influence diabetes likelihood. Higher values increase risk.

Comorbidity Impact

Hypertension and heart disease elevate diabetes risk. These conditions are closely linked.

Demographic Patterns

African Americans showed higher predisposition. Specific geographic hotspots were identified.



Conclusion and Future Work

Robust Prediction Model

We developed a highly accurate diabetes prediction model. It offers potential for proactive healthcare.

Future Directions

Deployment as a web application is planned. Expanding the dataset will enhance capabilities.

Enhanced Data Integration

Incorporating cholesterol levels and other health data. This will improve model comprehensiveness.

Real-World Impact

Collaboration with healthcare providers for implementation. This will drive practical benefits.