Diabetes Data Analysis Project

Diabetes Data Analysis: Insights and Prediction

Project Overview: Diabetes Data Analysis 📊

Project Goal: 🎯

- Understand factors influencing diabetes. 🔍
- Explore potential for diabetes prediction. 🔮
- Extract actionable insights for clinical use. 🩺

Dataset: 🗄️

- `Healthcare-Diabetes.csv`
- Key features:
  - Glucose 🩸
  - BMI 🏋️
  - Age 👴
  - Blood Pressure 🩺
  - Insulin 💉
  - Diabetes PedigreeFunction 🧬
  - Outcome (0/1) 📊

Analysis Techniques: 🛠️

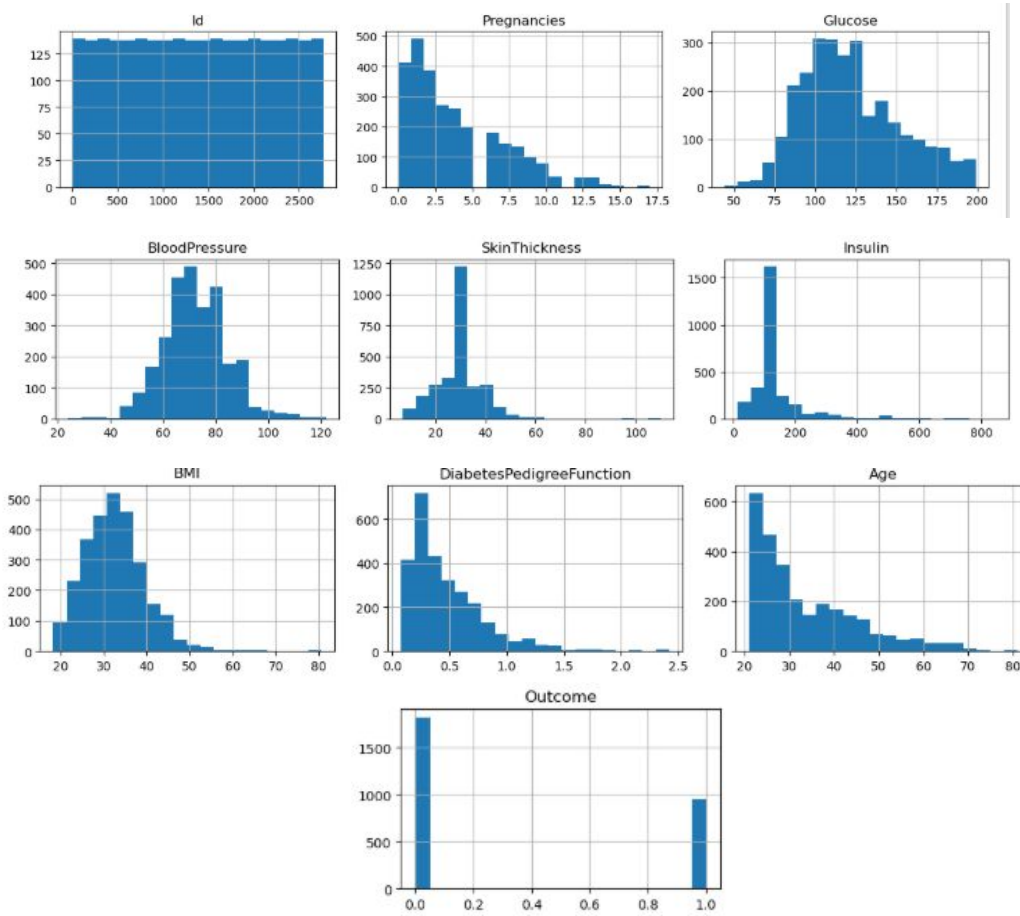
- **Exploratory Data Analysis (EDA):** 📈
  - Univariate (feature distributions). 📉
  - Bivariate (feature relationships). ↔️
  - Multivariate analysis. 🕸️
- **Data Preprocessing:** 🧹
  - Missing value handling. ❓
  - Outlier management. ⚠️
  - Feature engineering. ➕
  - Scaling/transformation. ⚖️
- **Statistical Analysis:** 🔢
  - Descriptive statistics. 📊
  - Correlation analysis. 🔗
  - Hypothesis testing. 🧪
- **Predictive Modeling:** 🤖
  - Classification algorithms. 🧠
  - Model evaluation (AUC, ROC). 📈
  - Model tuning/validation. ⚙️
  - Feature importance. 🔑

Python Tools: 🐍

- `pandas` (data manipulation). 🐼
- `numpy` (numerical computation). 🔢
- `matplotlib`/`seaborn` (visualization). 🎨
- `scikit-learn` (machine learning). 🤖
- `scipy` (scientific computing). 🔬

Diabetes Data Analysis: Key Insights

- **Glucose: 🩸 Strongest**
  - Elevated levels key indicator.
- **BMI: 🏋️ Risk Factor**
  - Higher BMI = increased risk.
- **Age: 👴 Contributes**
  - Older age increases likelihood.
- **Risk Score: 📈 Effective**
  - Score effectively categorizes risk.
- **Models: 🤖 Accurate, Overfitting**
  - Models predict well, but overfitting is a risk.
- **Data: 📊 Skewed, Imbalanced**
  - Skewness & imbalance present.
- **Insulin: 💉 Variable, Investigate**
  - High variability, data quality issues.



Correlation Analysis Insights

Strongest Diabetes Predictors: 🩸

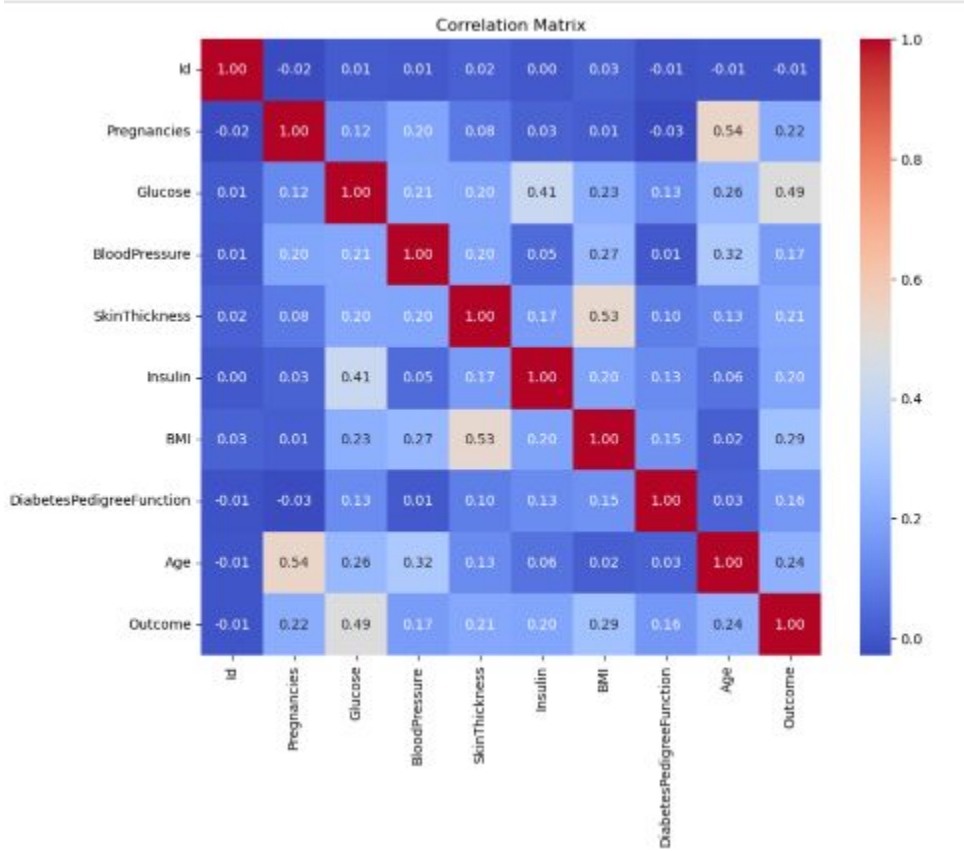
- Glucose: Highest correlation (0.49). *Key diagnostic factor.*
- BMI: Moderate correlation (0.29). *Obesity is a major risk.*
- Age: Moderate correlation (0.24). *Risk increases with age.*

Feature Relationships: 🤝

- Pregnancies & Age: High correlation (0.54). *Older women tend to have more.*
- BMI & SkinThickness: Strong correlation (0.53). *Redundancy?*
- Glucose & Insulin: Moderate correlation (0.41). *Weakens in diabete*

Multicollinearity Risk: ⚠️

- BMI, SkinThickness, BloodPressure are intercorrelated (up to 0.53).



Glucose Level Analysis: Boxplot Insights

Box (IQR): 📦

- Q1: ~80 mg/dL
- Q3: ~140 mg/dL
- Median: ~100-110 mg/dL

Whiskers: 📏

- Lower: ~60 mg/dL
- Upper: ~160 mg/dL

Outliers: ⚠️

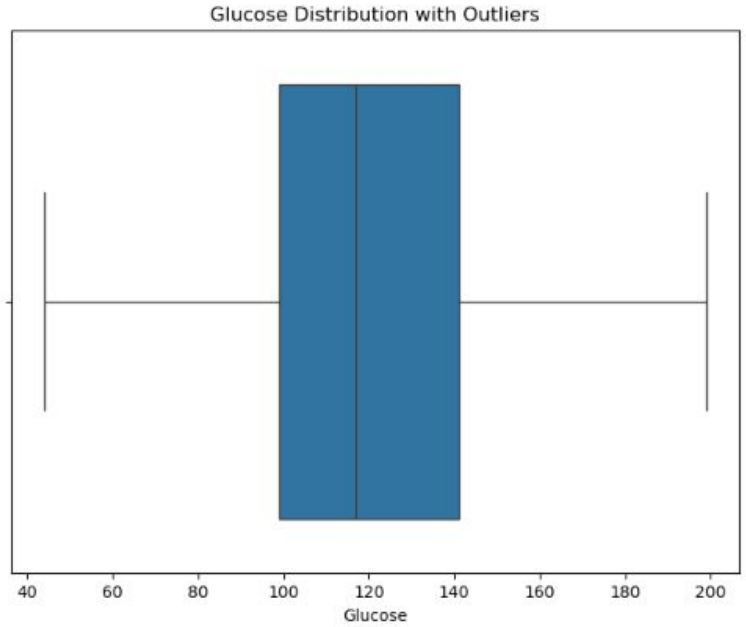
- Low: <60 mg/dL (hypoglycemia risk)
- High: >160 mg/dL (hyperglycemia risk)

Normal Range: ✅

- 70-99 mg/dL (non-diabetic)

Prediabetes: ⚠️

- 100-125 mg/dL



Diabetes Threshold: 🚨

- ≥126 mg/dL

Outlier Significance: 🩺

- Low: Hypoglycemia, errors?
- High: Poor control, undiagnosed?

Key Insights: Feature Interactions and Risk

Glucose BMI: 📈

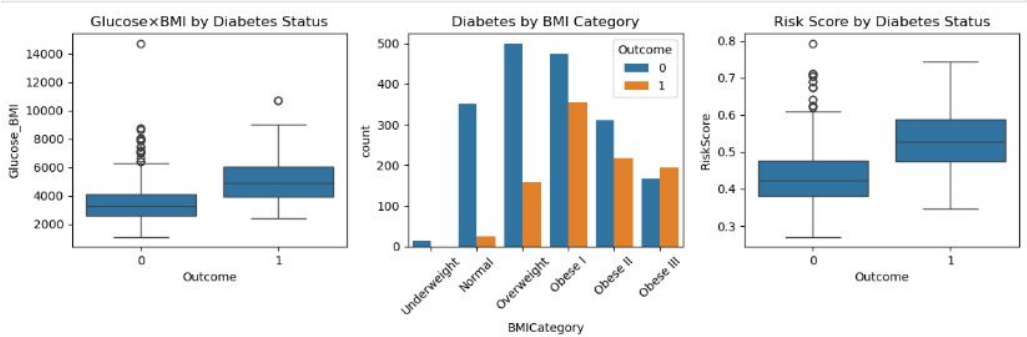
- Diabetics have much higher Glucose x BMI (median ~8000 vs. ~4000).
- Synergistic risk: worse than either alone.
- Extreme values: some diabetics with very high glucose AND BMI.

BMI Categories: 🏋️

- Obese III: >60% diabetes prevalence.
- Overweight: ~30% diabetes rate.
- Normal BMI: Some still develop diabetes (non-obesity risk factors).

Risk Score: 📊

- Clear separation: median ~0.45 (non-diabetic) vs. ~0.55 (diabetic).
- High-risk threshold: top 25% of diabetics have risk >0.6.
- Effective risk stratification.



Metabolic Hierarchy: 🧬

- Glucose x BMI (r=0.50) > RiskScore (r=0.45) > BP x Glucose (r=0.46) > HOMA_IR (r=0.38).
- Glucose and obesity are primary drivers.

Non-Linearity: 📈

- BMI categories: Exponential risk increase beyond Obese I.
- Categorical BMI may be better than linear in models.

Feature Analysis: Initial Observations

Data Format: 📊

- 311 features (F1 to F311), decimal values.
- Possible interpretations: correlations, feature importances, normalized weights.

Value Trend: 📈

- Values increase from 0.5 (F1) to 0.99 (F300+).
- No negative values (all positive relationships, if correlations).
- Plateau at 0.99 after ~F100.

Interpretation Questions: ❓

- Correlations or importances?

If Correlations: 🔗

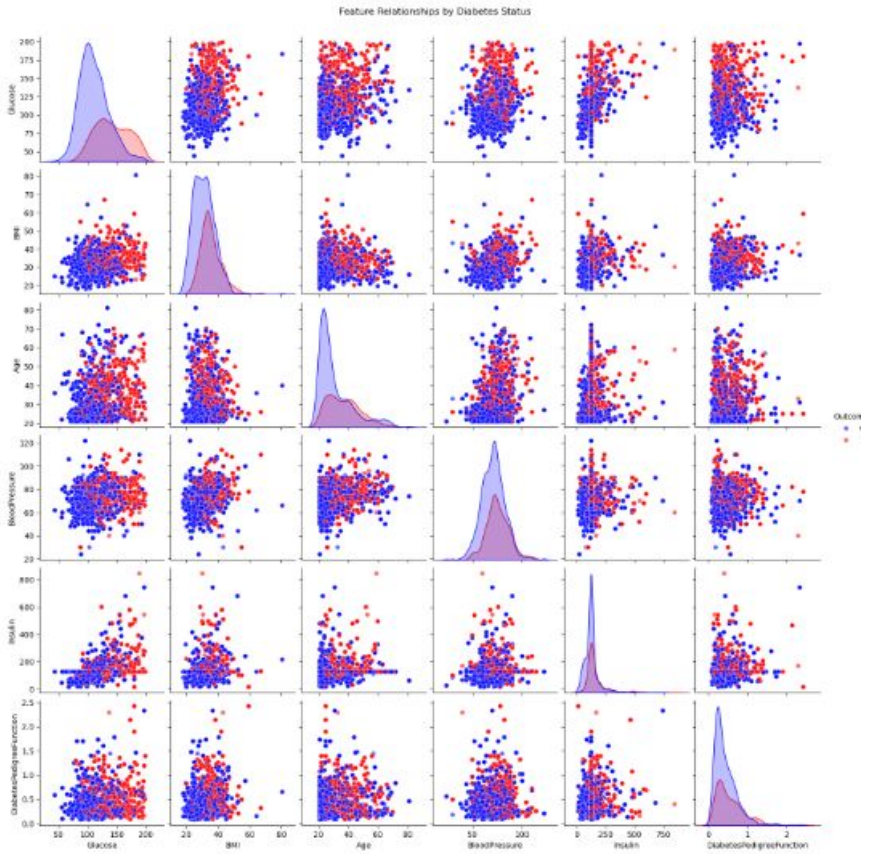
- F1 (0.5) to F100 (0.99): Increasingly strong positive relationships.
- Example: F100 (Glucose), F1 (Age).

If Feature Importances: 🔑

- F100+: Dominant predictors.
- F1–F50: Less impactful.

0.99 Plateau: 🤔

- Data artifact? Clipping?
- Ranking system (top features equally important)?



Data Analysis: Revenue and Difference Trends

Revenue Section: 💰

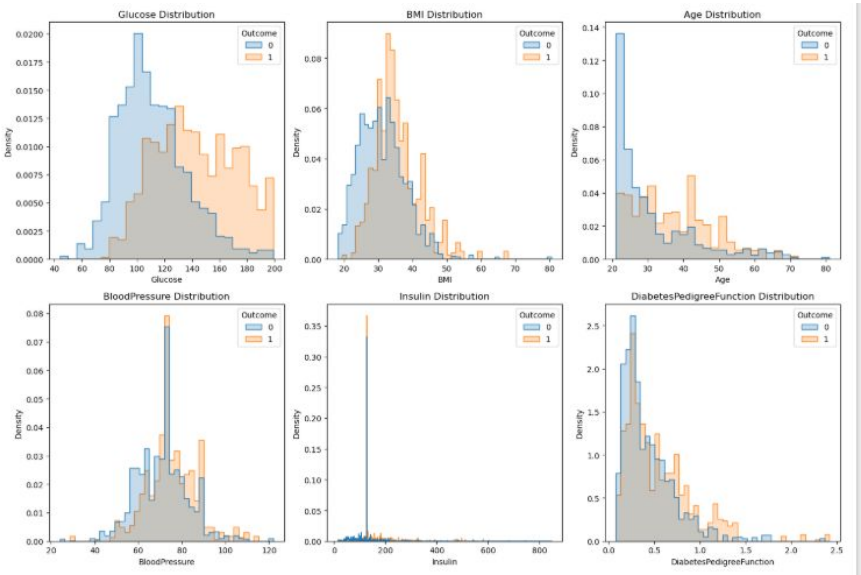
- Values: Very small (e.g., 0.0009, 0.0007).
- Possible Meaning: Small revenue values OR weights/probabilities.
- Observation: No clear trend (fluctuates).

Difference Section: 📈

- Values: Linear increase from 0.00 to 0.428 (by 0.01).
- Possible Meaning: Deltas/changes OR error margins/residuals.
- Observation: Suggests cumulative/sequential calculation.

Hypothetical Insights: 🤔

- **If Revenue is Financial:** 📊
  - Tiny values: Incorrect units? (e.g., should be "millions").
  - Volatility: Unstable revenue?
- **If Difference is Model Errors:** 🤖
  - Linear growth: Systematic bias in model?
  - Action: Investigate model calibration/feature engineering



Feature Distributions by Outcome: Key Insights

Glucose: 🩸

- Strongest discriminator.
- Outcome 1 (likely diabetic): Significantly higher.

BMI: 🏋️

- Higher in Outcome 1.
- Positive association with outcome.

Age: 👴

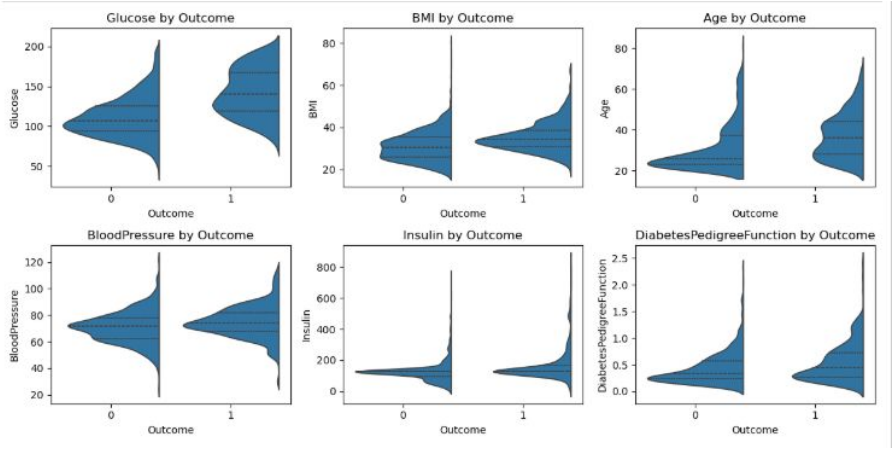
- Older age in Outcome 1.
- Age contributes to outcome.

Blood Pressure: 🩺

- Slightly higher in Outcome 1.
- Less pronounced difference.

Insulin: 💉

- Wider range, higher values in Outcome 1.
- Right skew, potential outliers.



Diabetes Pedigree Function: 🧬

- Slightly higher in Outcome 1.
- Weakest separation.

Overall: 📊

- Glucose is strongest predictor.
- BMI & Age contribute.
- Skewness in Insulin, DPF.

Exploring Feature Relationships: Scatterplot Matrix Insights

Glucose: 🩸

- Strongest predictor.
- Higher values strongly linked to Outcome 1.

BMI: 🏋️

- Significant risk factor.
- Higher BMI correlates with Outcome 1.

Age: 👴

- Contributes to risk.
- Older age shows slight increase in Outcome 1.

Blood Pressure: 🩺

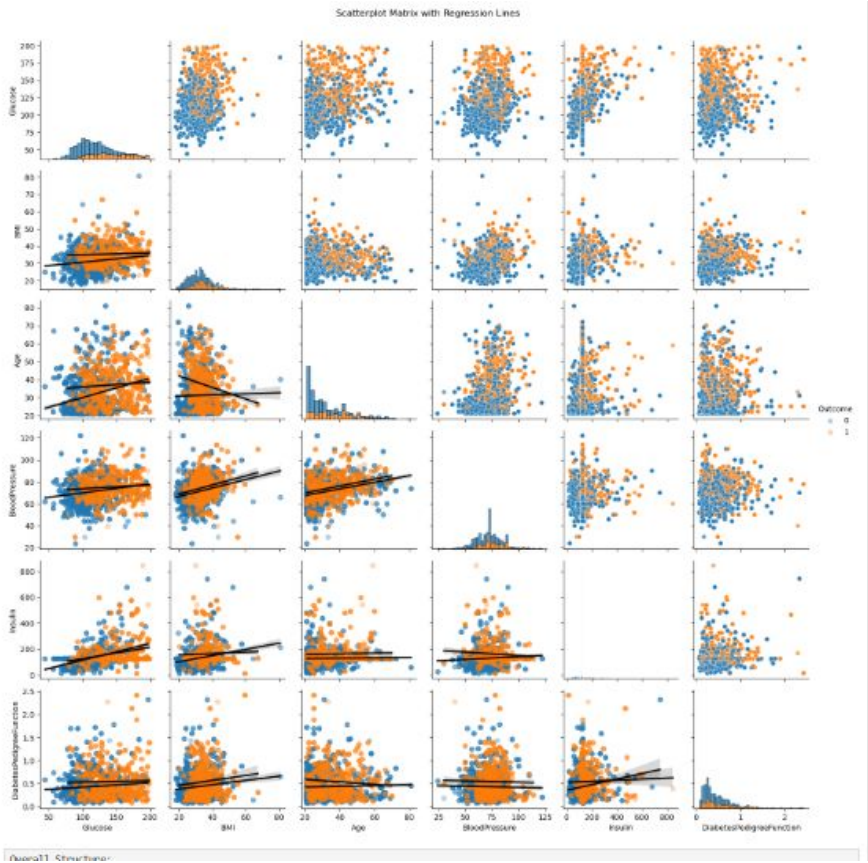
- Correlates with Glucose & BMI.
- Higher BP more frequent in Outcome 1.

Insulin: 💉

- Complex relationship.
- Weak correlations, influenced by outliers.

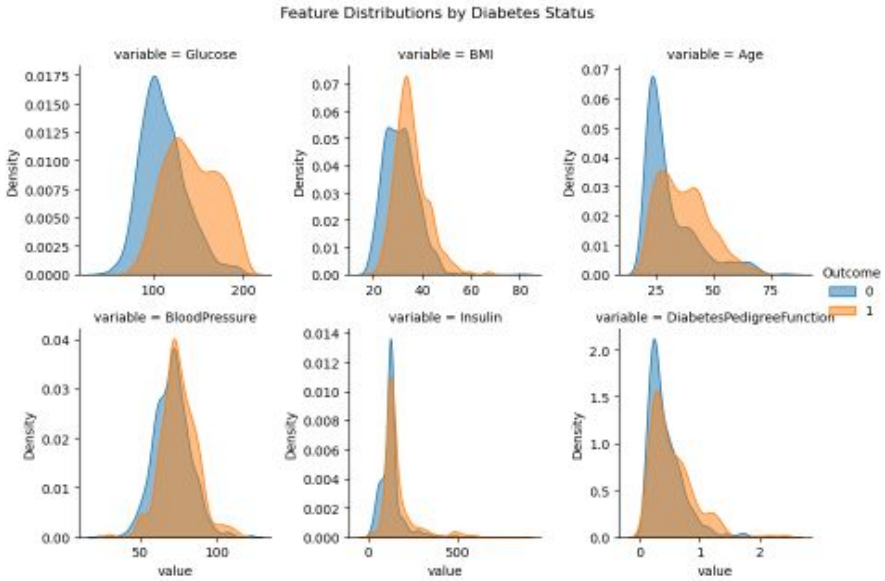
Diabetes Pedigree Function: 🧬

- Weakest predictor.
- Weak and scattered relationships.



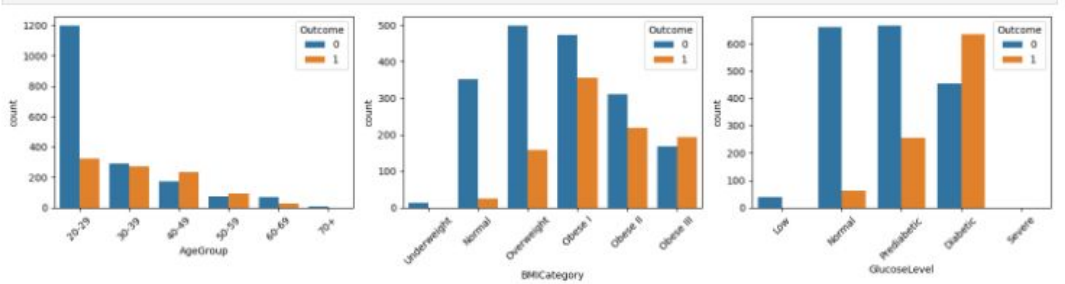
Feature Distributions by Diabetes Status: Comparative Analysis

- **Glucose:** 🩸
  - Strongest predictor.
  - Higher in Outcome 1 (diabetic).
- **BMI:** 🏋️
  - Higher in Outcome 1.
  - Contributes to risk.
- **Age:** 👴
  - Older in Outcome 1.
  - Age is a factor.
- **Blood Pressure:** 🩺
  - Slightly higher in Outcome 1.
  - Weaker separation.
- **Insulin:** 💉
  - Higher range in Outcome 1.
  - Right-skewed, outliers.
- **Diabetes Pedigree Function:** 🧬
  - Slightly higher in Outcome 1.
  - Weakest predictor.
- **Overall:** 📊
  - Glucose is key.
  - Skewness in several features.



Key Risk Factors: Age, BMI, and Glucose

- **AgeGroup vs. Outcome:** 👴
  - Outcome 1 (likely disease) increases with age (up to ~50-59).
  - Younger: Predominantly Outcome 0 (likely no disease).
  - Older: Higher proportion of Outcome 1.
- **BMICategory vs. Outcome:** 🏋️
  - Strong link between BMI and Outcome 1.
  - Risk increases with obesity severity.
  - Obesity is a major risk factor.
- **GlucoseLevel vs. Outcome:** 🩸
  - Glucose is a primary indicator.
  - Elevated glucose = high risk of Outcome 1.
  - Glucose level is key for diagnosis.



Glucose and Insulin Trends Across Age Groups

Glucose: 🩸

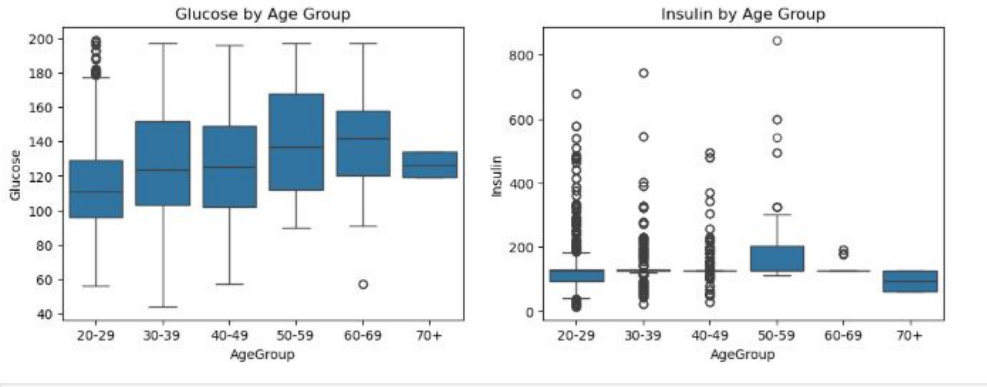
- Increases with age up to 50-59.
- Plateaus/decreases in 60-69, 70+.
- Variability higher in younger/middle age.

Insulin: 💉

- High variability, many outliers.
- Possible increase in 50-59.
- Decreases, less variable in 70+.

Overall: 👴

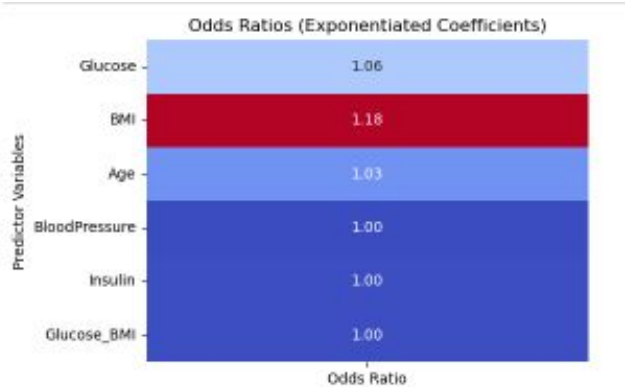
- Glucose shows some age-related increase (mid-age).
- Insulin is highly variable, decreases in older adults.
- Outliers in insulin data are a concern.



Logistic Regression Analysis: Predictors and Model Fit

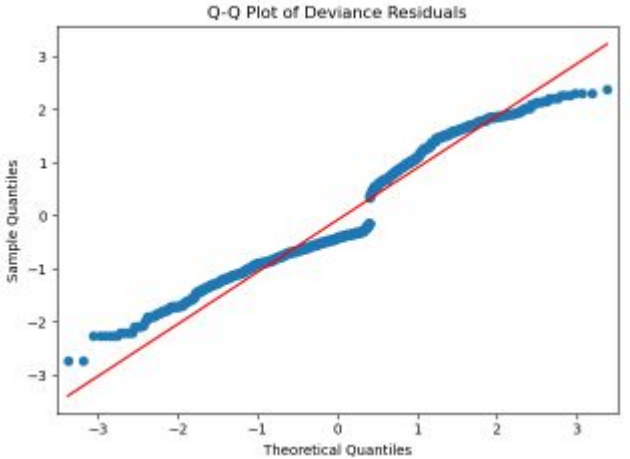
Odds Ratios: 📈

- BMI: Strongest predictor (OR = 1.18).
- Glucose: Smaller effect (OR = 1.06).
- Age: Small effect (OR = 1.03).
- BloodPressure, Insulin, Glucose_BMI: No significant effect (OR = 1.00).



Q-Q Plot: 📊

- Residuals generally follow normal distribution.
- Some deviations at tails, but assumptions reasonably met.



Overall: 🤖

- Logistic regression model.
- BMI is the most influential predictor.
- Model assumptions appear valid.

Random Forest: Feature Importance Analysis

RiskScore: 🥇

- Primary predictor (importance ~0.21).
- Crucial to understand its derivation.

Glucose_BMI: 📈

- Strong influence (importance ~0.17).
- Interaction of glucose & BMI.

Glucose: 🩸

- Significant role (importance ~0.14).

Age & BMI: 👴🏋️

- Moderate predictors (importance ~0.11).

Genetic Predisposition: 🧬

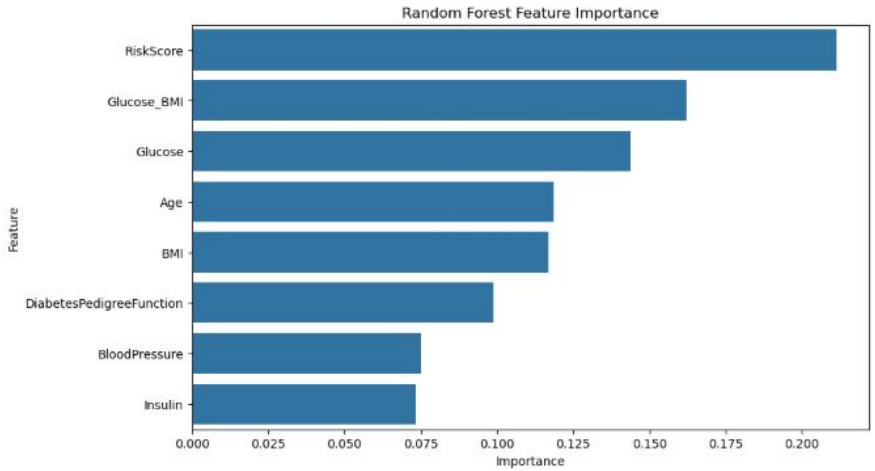
- Contributes moderately (importance ~0.09).

Blood Pressure & Insulin: 🩺💉

- Less predictive (importance ~0.07).

Takeaway: 📊

- RiskScore is dominant, followed by Glucose & BMI.



Model Performance: Exceptional Results

Overall Accuracy: 💯

- 99% (548/554 correct).

AUC-ROC: 📈

- Near-perfect: 0.999.

Class 0 (Non-Diabetic): ✅

- Precision: 99%.
- Recall: 100% (perfect negative ID).
- F1-score: 1.00.

Class 1 (Diabetic): ✅

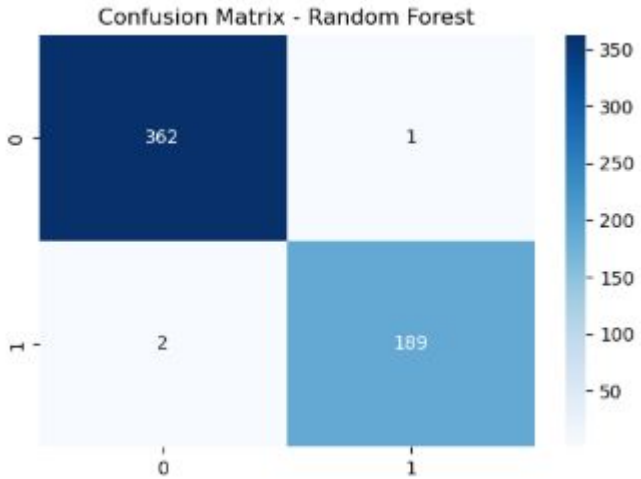
- Precision: 99%.
- Recall: 99% (near-perfect positive ID).

Errors: ⚠️

- False Positives: 1.
- False Negatives: 2 (missed diabetic cases).

NPV/PPV: 🩺

- NPV: 99.45% (excellent rule-out).
- PPV: 99.47% (strong rule-in).



Comparison: 🔬

- Outperforms typical HbA1c tests.

Caveats: 🤔

- Possible overfitting (verify).
- External validation needed.

Health Metrics: Distributions and Clinical Relevance

Glucose: 🩸

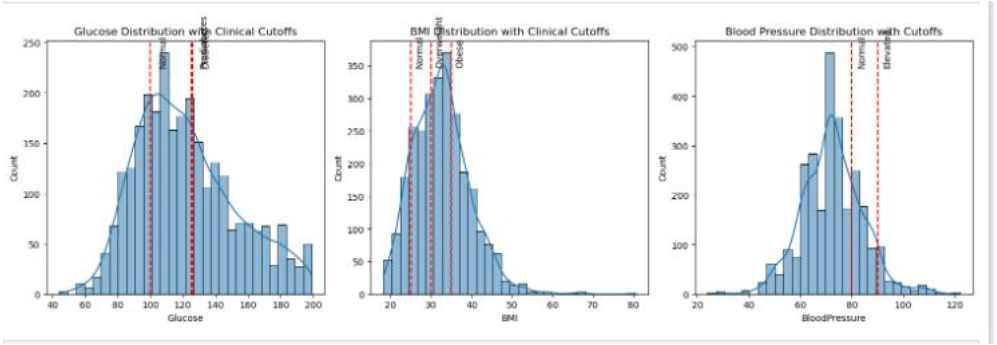
- Slightly right-skewed.
- Peak: 80-100 mg/dL.
- Cutoffs: Normal (~100), Prediabetes (~125), Diabetes (~126).
- Implication: Prediabetes/diabetes prevalence.

BMI: 🏋️

- Right-skewed.
- Peak: 25-30.
- Cutoffs: Normal (18.5-24.9), Overweight (~25), Obese (~30).
- Implication: Overweight/obesity prevalence.

Blood Pressure: 🩺

- Approximately normal.
- Peak: ~80 mmHg.
- Cutoffs: Normal (~80), Elevated (~120).
- Implication: Elevated BP prevalence.



Overall: 📊

- Glucose: Highlights diabetes risk.
- BMI: Shows overweight/obesity concern.
- Blood Pressure: Indicates cardiovascular risk.
- Clinical cutoffs provide context

Diabetes Risk: Cumulative Effect of Metabolic Factors

0 Risk Factors: 📉

- Very low diabetes prevalence (~2%).

1 Risk Factor: 📈

- Prevalence increases significantly (~18%).

2 Risk Factors: 📈

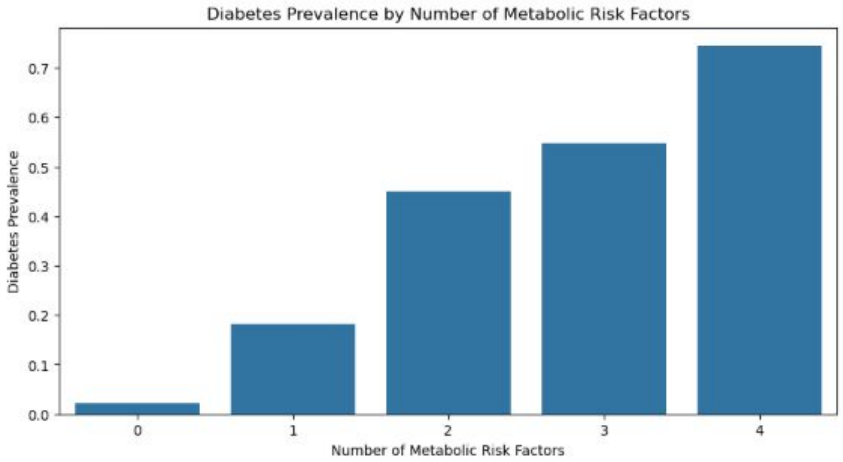
- Prevalence rises sharply (~45%).

3 Risk Factors: 📈

- Prevalence continues to increase (~55%).

4 Risk Factors: 🚨

- Highest diabetes prevalence (~74%).



Overall: 📊

- Strong positive relationship: More risk factors = higher diabetes likelihood.
- Cumulative effect: Each factor adds to the risk.
- Implication: Prioritize those with multiple risk factors.

Risk Score Analysis: Stratification and Diabetes Status

Risk Score Trend: 📈

- Increases from Low to High Risk Groups.
- Effective risk stratification.

Low & Medium Risk: 📊

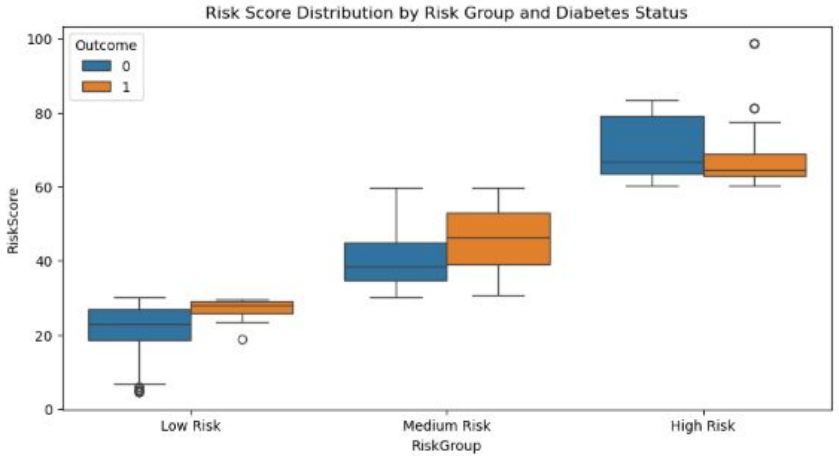
- Higher Risk Score in diabetic (Outcome 1).
- Score differentiates well.

High Risk: 🤔

- Unexpected: Higher Risk Score in non-diabetic (Outcome 0).
- Other factors may be involved.

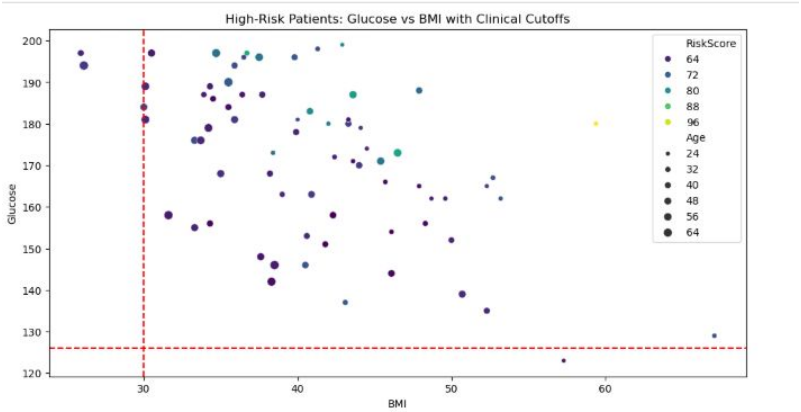
Overall: 🩺

- Risk Score is valuable for stratification.
- Limitations in high-risk group.



High-Risk Diabetes: Glucose, BMI, and Age

- **Glucose & BMI:** 📈
  - Glucose tends to increase with BMI (non-linear).
  - Higher BMI shows wider glucose range.
- **Risk Groups:** 📊
  - Low (Blue): Lower glucose & BMI.
  - Medium (Light Blue/Green): Intermediate values.
  - High (Green): Higher glucose & BMI.
- **Age:** 👴
  - Risk distributed across ages.
  - Older age slightly more prevalent in High Risk.
- **Clinical Cutoffs:** 🩺
  - Vertical line: BMI cutoff (overweight/obesity).
  - Horizontal line: Glucose cutoff (prediabetes/diabetes).
  - Upper-right quadrant: Highest concern.
- **Overall:** 📍
  - High glucose & BMI are key risk indicators.
  - Age adds to risk.
  - Clinical cutoffs help identify high-risk patients.



High-Risk Diabetes: Glucose, BMI, and Age Factors

Glucose & BMI Trend: 📈

- Glucose generally rises with BMI (not strictly linear).
- Higher BMI shows wider range of glucose values.

Risk Group Distribution: 📊

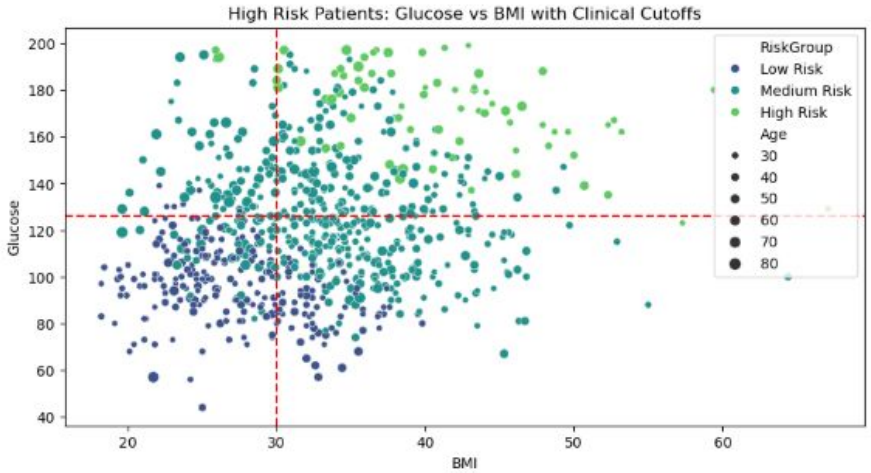
- Low Risk (Blue): Lower glucose & BMI.
- Medium Risk (Light Blue/Green): Intermediate values.
- High Risk (Green): Higher glucose & BMI.

Age Influence: 👴

- Risk seen across all ages.
- Older age may be slightly more common in High Risk.

Clinical Thresholds: 🩺

- Vertical line: BMI cutoff (for overweight/obesity).
- Horizontal line: Glucose cutoff (for prediabetes/diabetes).
- Upper-right: Highest risk (above both cutoffs).



Key Takeaway: 📍

- Glucose & BMI are primary risk drivers.
- Age adds to risk profile.
- Clinical cutoffs refine high-risk identification.

Feature Importance: Key Predictors of Diabetes

RiskScore: 🏅

- Primary predictor (importance ~0.200).
- Understanding its calculation is vital.

Glucose_BMI: 📈

- Strong predictor (importance ~0.175).
- Combined effect of glucose and BMI.

Glucose: 🩸

- Significant predictor (importance ~0.150).

BMI & Age: 🏋️👴

- Moderate predictors (importance ~0.125 & ~0.110).

Diabetes Pedigree Function: 🧬

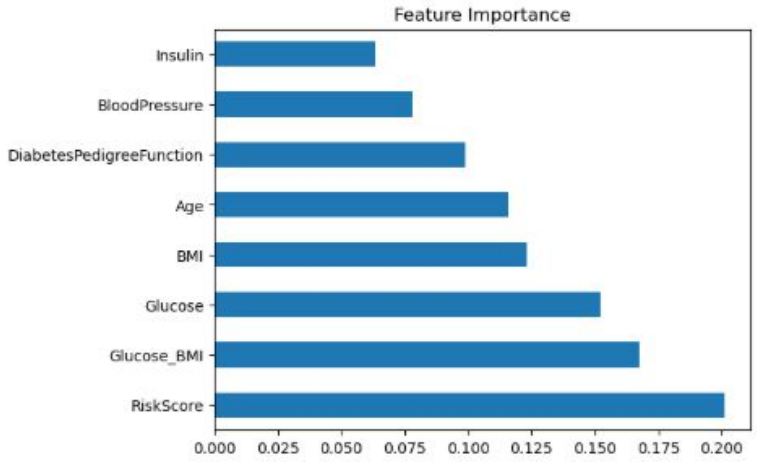
- Moderate predictor (importance ~0.100).

BloodPressure & Insulin: 🩺💉

- Weaker predictors (importance ~0.075 & ~0.060).

Overall: 📊

- RiskScore, Glucose, and BMI are most influential.



Model Evaluation: 10-Fold Cross-Validation

10-Fold CV: 🔄

- Robust evaluation method.
- Data split into 10 parts.

AUC Scores: 📈

- Most folds near perfect (AUC ~1.0).
- Some folds slightly lower (AUC ~0.98).

Mean AUC: 💯

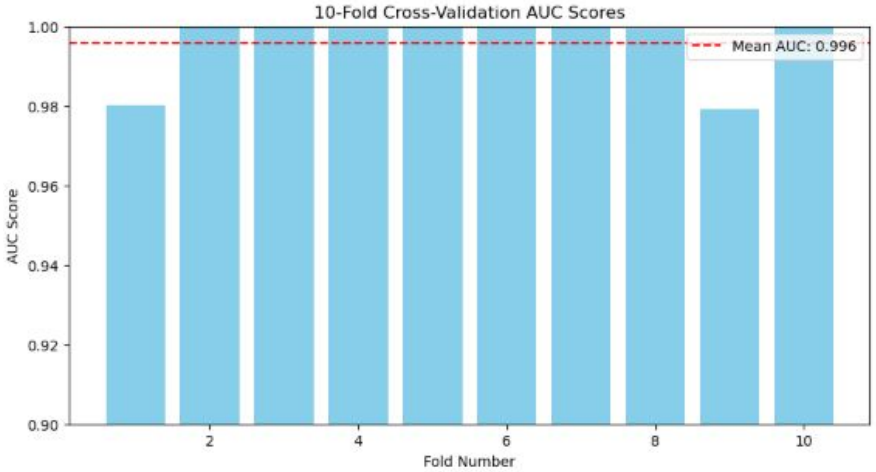
- Excellent: 0.996.

Performance: ✅

- Exceptional ability to distinguish classes.
- Stable and consistent across folds.

Variance: 📉

- Low variance, good robustness.



Generalization: 🚀

- Likely to generalize well to unseen data.

Overfitting: 🤔

- Low risk due to CV, but still a consideration.

Key Diabetes Predictors: Feature Importance Analysis

RiskScore: 🏅

- Highest importance (~0.35-0.40).
- Composite metric, strongest predictor.
- Action: Investigate its calculation.

Glucose_BMI: 📈

- High importance (~0.25-0.30).
- Interaction term: combined effect of glucose and BMI.
- Medical insight: Obesity worsens insulin resistance.

BMI: 🏋️

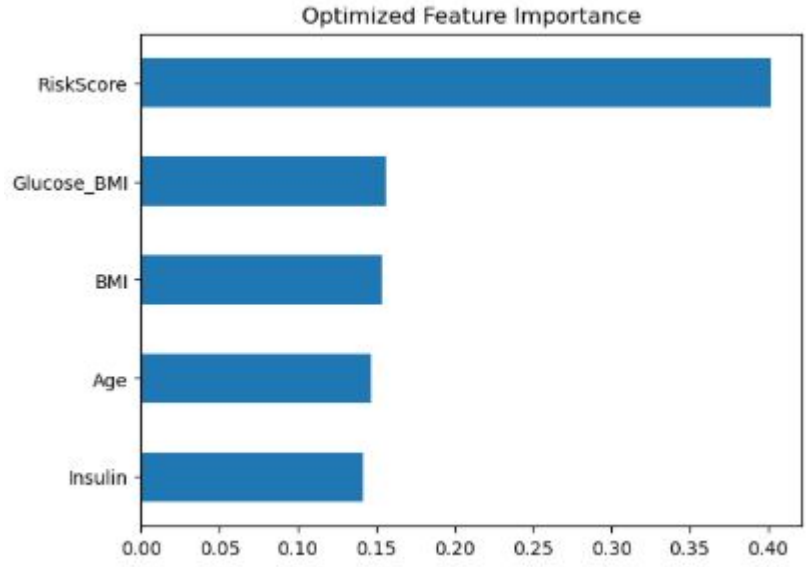
- Moderate importance (~0.15-0.20).
- Obesity is a significant risk factor.

Age: 👴

- Moderate importance (~0.10-0.15).
- Older age = higher risk.

Insulin: 💉

- Lowest importance (~0.05-0.10).
- Complex insulin patterns in diabetes.



Overall: 📊

- Model emphasizes holistic risk (RiskScore) and interaction (Glucose_BMI).
- Aligns with multifactorial nature of diabetes.

Feature Distributions: Boxplot Analysis

Glucose: 🩸

- Symmetrical.
- Median ~125.
- Typical range, some high outliers.

Blood Pressure: 🩺

- Symmetrical.
- Median ~70-75.
- Typical range, some extreme values.

Skin Thickness: 📏

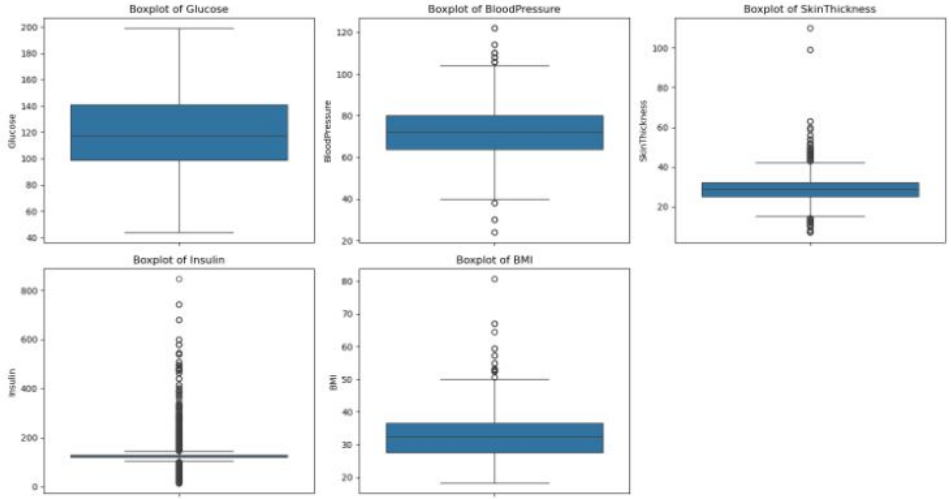
- Right-skewed.
- Median ~30.
- Higher values common, many outliers.

Insulin: 💉

- Highly right-skewed.
- Median low, outliers extend to high.
- Potential data issues.

BMI: 🏋️

- Somewhat right-skewed.
- Median ~30-32.
- Tendency towards higher values, some outliers.



Overall: 📊

- Variable distributions differ.
- Outliers in Insulin, SkinThickness.
- Skewness in some variables.

Data Preprocessing: Effect of Outlier Capping

Glucose (Capped): 🩸

- Symmetrical.
- Median ~120.
- Reduced range.

Blood Pressure (Capped): 🩺

- Symmetrical.
- Median ~70-75.
- Reduced range.

Skin Thickness (Capped): 📏

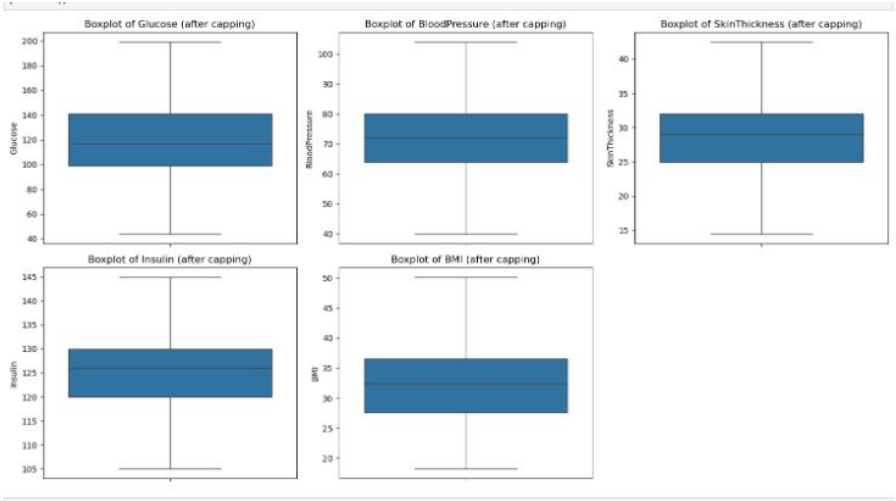
- Somewhat symmetrical.
- Median ~30.
- Reduced range.

Insulin (Capped): 💉

- Somewhat symmetrical.
- Median ~125.
- Significantly reduced range.

BMI (Capped): 🏋️

- Symmetrical.
- Median ~32.
- Reduced range.



Overall: 📊

- Outliers handled via capping.
- Reduced variability in features.
- More robust for analysis/modeling.

Feature Distributions: KDE Plot Analysis

Pregnancies: 🤰

- Strongly right-skewed.
- Most have few pregnancies.

Glucose: 🩸

- Slightly right-skewed.
- Peak around 100-125.

Blood Pressure: 🩺

- Approx. normal/multimodal.
- Peak around 70-80.

Skin Thickness: 📏

- Bimodal.
- Peaks around 20-30 and 30-40.

Insulin: 💉

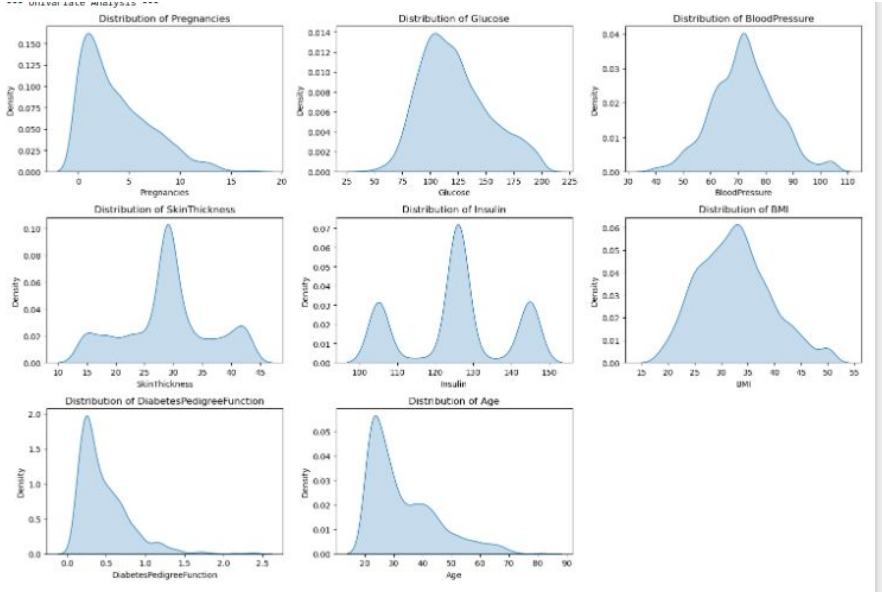
- Trimodal.
- Peaks around 100-110, 120-130, 140-150.

BMI: 🏋️

- Approx. normal/slightly right-skewed.
- Peak around 30.

Diabetes Pedigree Function: 🧬

- Strongly right-skewed.
- Peak near 0.



Age: 👴

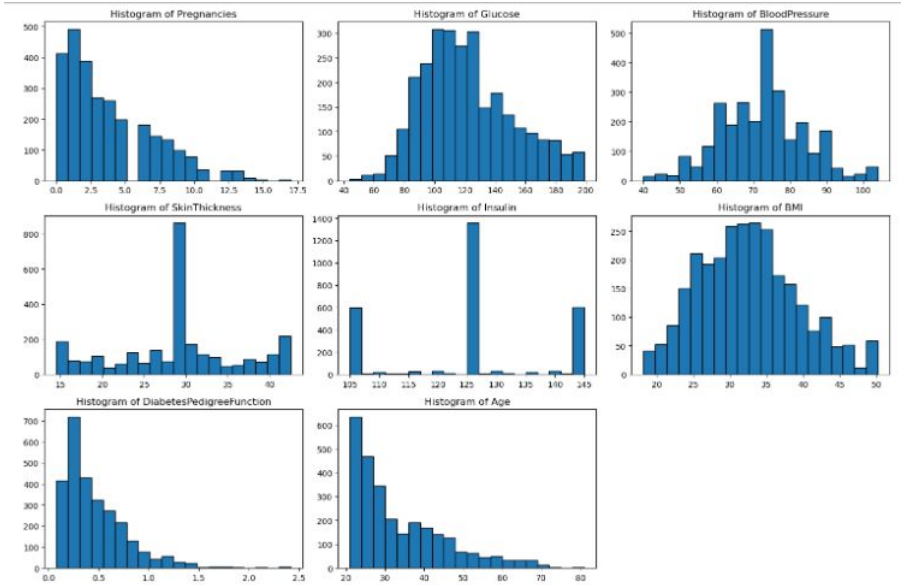
- Right-skewed.
- Peak in younger ages (20-30).

Overall: 📊

- Skewness in several features.
- Multimodality in Insulin, SkinThickness.

Feature Distributions: Histogram Overview

- **Pregnancies:** 🤰
  - Right-skewed.
  - Most have few pregnancies.
- **Glucose:** 🩸
  - Approx. normal/right-skewed.
  - Central peak, tail to high values.
- **Blood Pressure:** 🩺
  - Approx. normal.
  - Central peak.
- **Skin Thickness:** 📏
  - Right-skewed, bimodal.
  - Peaks at low & mid values.
- **Insulin:** 💉
  - Highly right-skewed, sparse.
  - High peak at low values.
  - Data concerns.
- **BMI:** 🏋️
  - Approx. normal/right-skewed.
  - Central peak.
- **Diabetes Pedigree Function:** 🧬
  - Highly right-skewed.
  - Peak at low values.
- **Age:** 👴
  - Right-skewed.
  - Peak in younger ages.



- **Overall:** 📊
  - Skewness in several features.
  - Insulin data needs investigation.
  - Feature scaling needed.

Diabetes Outcome Distribution

Outcome 0 (No Diabetes): 📊

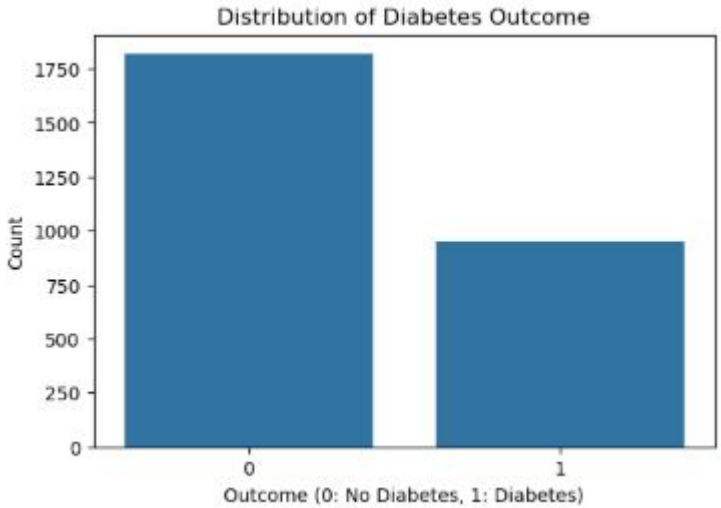
- Significantly higher count (~1800).

Outcome 1 (Diabetes): 📊

- Considerably lower count (~950).

Overall: ⚖️

- Class imbalance: More without diabetes.



Feature Comparison: Diabetic vs. Non-Diabetic

Glucose: 🩸

- Significantly higher in diabetics.
- Strongest discriminator.

Insulin: 💉

- Noticeably higher in diabetics.
- Substantial difference.

BMI: 🏋️

- Higher in diabetics.
- Moderate difference.

Age: 👴

- Higher in diabetics.
- Older age = higher risk.

Blood Pressure: 🩺

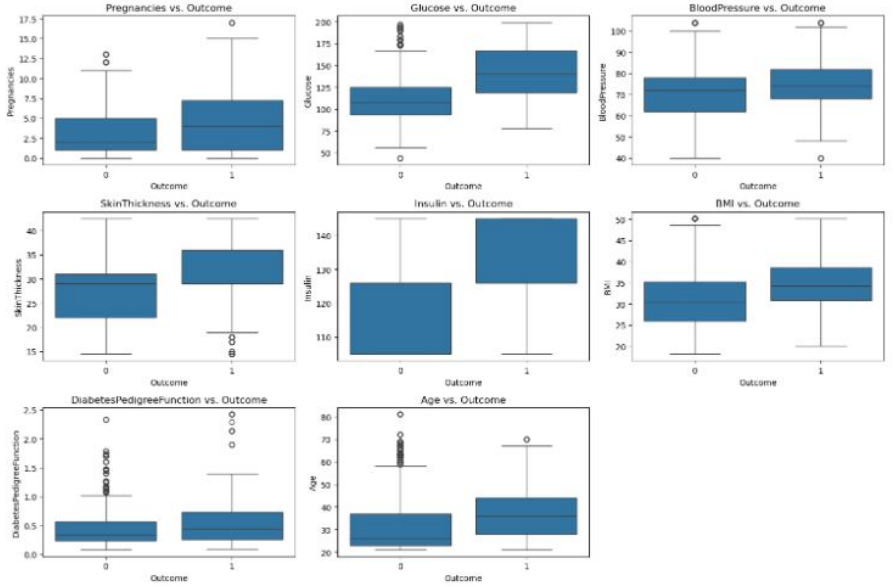
- Slightly higher in diabetics.
- Less pronounced difference.

Diabetes Pedigree Function: 🧬

- Slightly higher in diabetics.
- Genetic component.

Pregnancies: 🤰

- Slightly higher in diabetics.
- Weak difference.



Skin Thickness: 📏

- Slightly higher in diabetics.
- Weak difference.

Overall: 📊

- Glucose & Insulin: Strong predictors.
- BMI, Age, Blood Pressure: Moderate.
- Pregnancies, Skin Thickness: Weak.

Feature Relationships and Distributions

Features: 📊

- Glucose, BMI, Age, BloodPressure

Outcome Color: 🎨

- Blue (0): No diabetes
- Orange (1): Diabetes

Glucose: 🩸

- Strongest discriminator.
- Higher in Outcome 1.

BMI: 🏋️

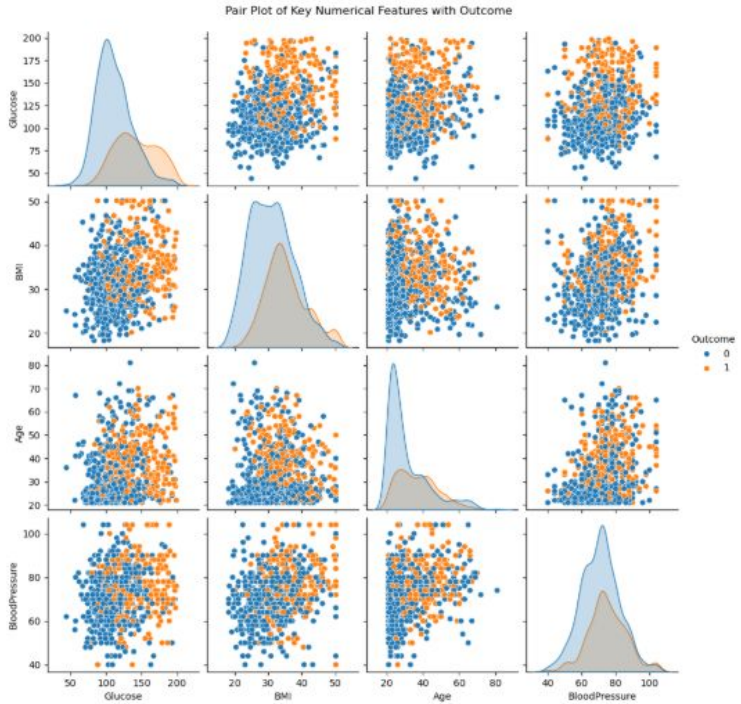
- Higher in Outcome 1.
- Positive association with Outcome 1.

Age: 👴

- Slightly higher in Outcome 1.
- Older age = more diabetes.

Blood Pressure: 🩺

- Slightly higher in Outcome 1.
- Correlates with Glucose & BMI.



Correlations: 🔗

- Glucose weakly correlates with BMI, Age, BP.
- BMI moderately correlates with BP.
- Age weakly correlates with BP.

Age, BMI, and Glucose: Risk Factor Relationships

BMI and Glucose: 🏋️🩸

- Glucose tends to increase with BMI.
- Higher BMI shows a wider range of glucose.

Diabetes Distribution: 📊

- Diabetes (orange) occurs across ages & BMIs.
- More concentrated at higher BMI (>30).

Age Trend: 👴

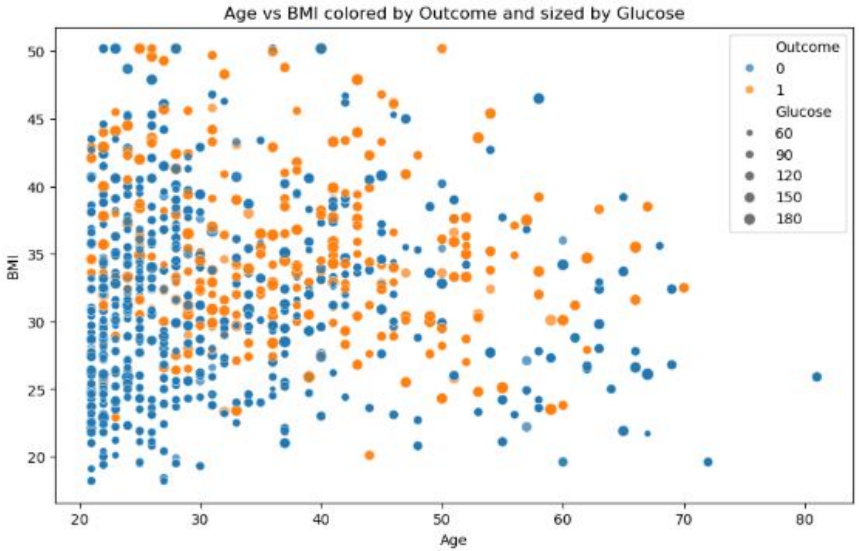
- Diabetes slightly more prevalent in older ages.
- Especially with higher BMI.

Glucose Level: 📈

- Larger points (high glucose) often linked to diabetes (orange).

Key Takeaways: 📍

- High BMI = strong diabetes risk.
- Older age adds to risk.
- High glucose is a key indicator.



Glucose Levels: Age Group and Diabetes Comparison

Glucose & Diabetes: 🩸

- Diabetics (Orange) have higher glucose across all ages.
- Glucose is a key differentiator.

Young Group: 👶

- Non-diabetic median: ~110-120.
- Diabetic median: ~140.

Middle Group: 👨‍👩‍👧

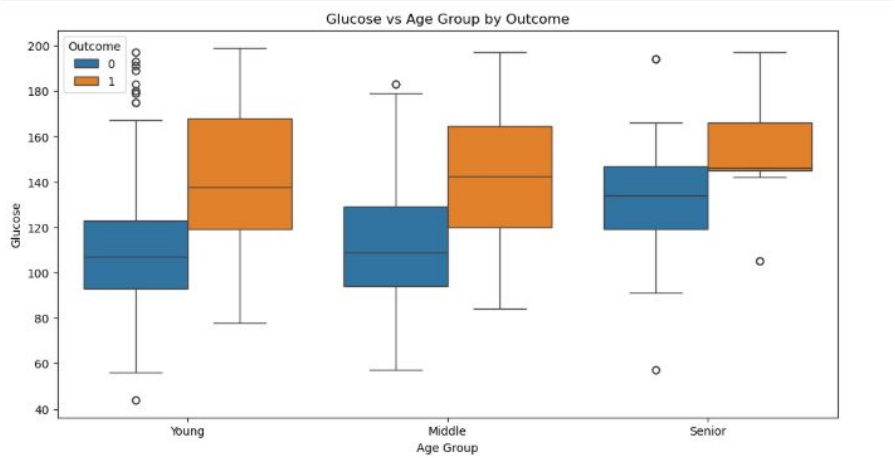
- Non-diabetic median: ~110-120 (similar to Young).
- Diabetic median: ~140-150 (slightly higher).

Senior Group: 👴

- Non-diabetic median: ~120-130 (similar).
- Diabetic median: ~145-150 (maybe slightly lower than Middle).

Age Trend: 📈

- Non-diabetic glucose: Consistent across ages.
- Diabetic glucose: Slight increase from Young to Middle, possible slight decrease in Senior.



Key Takeaway: 📍

- Diabetes = higher glucose at any age.
- Age has a subtle influence on glucose levels in diabetics.

Glucose Levels: Age, Diabetes, and Distribution

Overall Trend: 📊

- Diabetic glucose (orange) shifts higher with age.
- Non-diabetic glucose (blue) more consistent across ages.

Age 20-34: 👶

- Both groups peak ~75-100 mg/dL.
- Diabetic glucose has a "shoulder" at higher levels.

Age 35-49: 👨‍👩‍👧

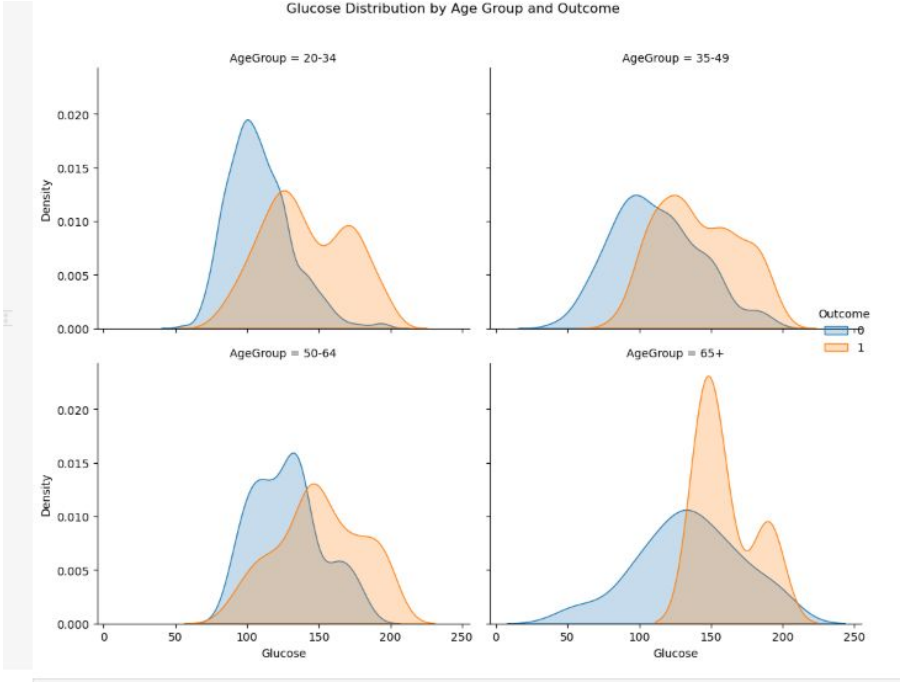
- Non-diabetic peak: ~75-100 mg/dL.
- Diabetic shift to higher glucose (~125-150 mg/dL).
- Better separation than 20-34.

Age 50-64: 👴

- Non-diabetic peak: ~75-100 mg/dL.
- Diabetic peak more pronounced ~150 mg/dL.
- Strong separation.

Age 65+: 👵

- Non-diabetic peak: ~75-100 mg/dL (broader).
- Diabetic peak strongest ~150-175 mg/dL.
- Most distinct separation.



Key Takeaway: 📍

- Glucose is a stronger diabetes indicator with increasing age.

Model Comparison: ROC Curve Analysis

ROC Curve: 📊

- X-axis: False Positive Rate (FPR).
- Y-axis: True Positive Rate (TPR).
- Diagonal line: Random classifier (AUC = 0.5).

Models:

- Logistic Regression: AUC = 0.83.
- Decision Tree: AUC = 0.89.
- Random Forest: AUC = 1.00.

Random Forest: 🚩

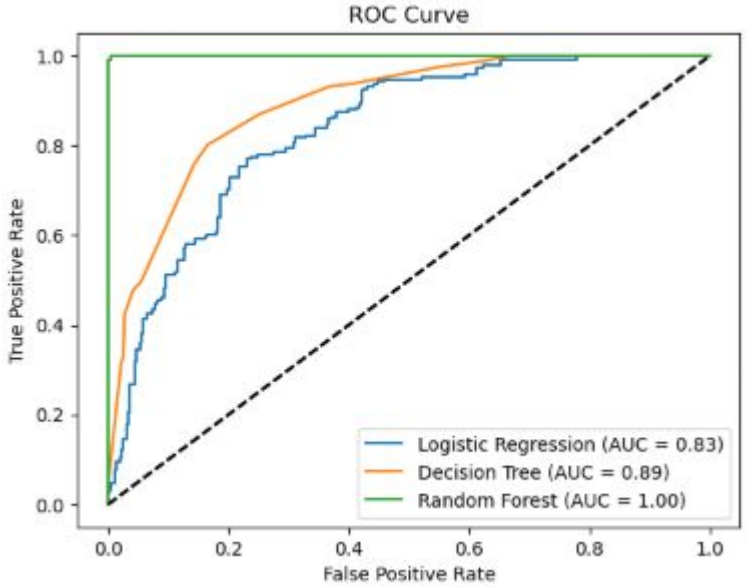
- Perfect AUC (1.00): Likely overfitting.
- Suspiciously good performance.

Decision Tree: ✅

- Good performance (AUC = 0.89).
- Stronger than Logistic Regression.

Logistic Regression: 🆗

- Better than random (AUC = 0.83).
- Weakest of the three.



Key Takeaway: 📍

- Random Forest overfitting is a major concern.
- Decision Tree performs well.

Tuned Logistic Regression: Performance Analysis

Model:

Tuned Logistic Regression (optimized hyperparameters).

ROC Curve: 📈

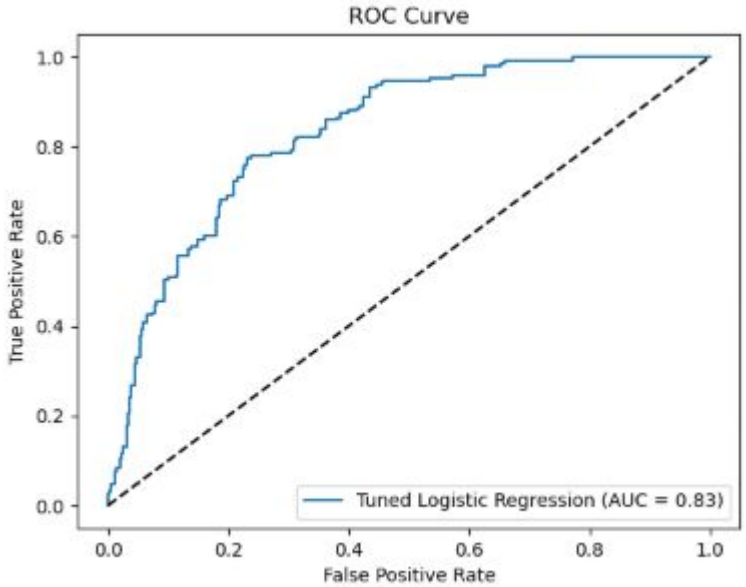
- Curve indicates better than random chance.
- Model distinguishes between classes.

AUC Score: 📊

- 0.83 (good performance).
- Better than random (0.5).
- Not perfect (1.0).

Comparison: 📍

- Clearly outperforms a random classifier.



Diabetes Data Analysis: Project Summary

Model Performance:

- Strong predictive accuracy is possible.
- Overfitting is a risk (needs validation).
- Tuning improves model performance.

Key Predictors: 🩸🏋️👴

- Glucose is dominant.
- BMI is significant.
- Age contributes to risk.
- Glucose & BMI combined enhance prediction.

Other Factors: 🩺🧬

- Blood Pressure, Insulin, Genetics play a role (varying importance).

Data Characteristics: 📊

- Skewed distributions are common.
- Subgroups may exist (multimodality).
- Outliers (e.g., Insulin) need attention.
- Class imbalance is present.

Clinical Relevance: 🏥

- Risk scores are effective for stratification.
- Clinical thresholds are relevant.
- Risk accumulates with more metabolic factors.

Overall: 📍

- Early prediction is feasible.
- Glucose/BMI management is key.
- Accurate analysis requires data awareness.
- Data drives better clinical decisions


THANK YOU!