# 

# 

# 

# 

# 

# **Project 2: Regression Modeling and Performance Evaluation**

# 

# 

# 

# 

# 

Sri Sai Palamoor

MSCS-634-B01

Advanced Big Data and Data Mining

# **1. Introduction**

This project aims to apply various data mining and machine learning techniques on the Pima Indians Diabetes Dataset to build predictive models and uncover valuable insights. The core focus includes regression, classification, clustering, and association rule mining, along with visualizations and evaluation metrics to assess model performance.

# **2. Dataset Description**

## **Dataset Used:**

* Pima Indians Diabetes Dataset from the UCI Machine Learning Repository.
* Size: 768 rows × 9 columns
* Target Variable: Outcome (0 = No Diabetes, 1 = Diabetes)

## **Features:**

| Feature | Description |
| --- | --- |
| Pregnancies | Number of times pregnant |
| Glucose | Plasma glucose concentration |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index (weight in kg/(height in m)^2) |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Age in years |

## **Why This Dataset?**

* Relevant for real-world health applications.
* Clean and well-structured.
* Offers numeric and categorical insights useful for supervised and unsupervised learning.
* Provides a balanced challenge for applying a wide variety of data science techniques.

# **3. Data Preprocessing and Feature Engineering**

**Missing Values Handling:**

* Features like Glucose, BMI, Blood Pressure, Insulin, and Skin Thickness contained 0s, which are physiologically impossible.
* Treated 0s as missing values and imputed using median values, considering skewed distributions.

**Outlier Detection:**

* Visualized using boxplots.
* Some apparent outliers (high insulin values) were retained due to medical plausibility.

**Feature Scaling:**

* Applied StandardScaler to normalize data for algorithms sensitive to scale (k-NN, clustering).

**Feature Engineering:**

* Created Age groups for mining patterns: {Young, Middle-aged, Senior}.
* Binned Glucose and BMI for association rule mining.
* Verified class distribution for the target variable — 35% positive class, 65% negative class.

# **4. Exploratory Data Analysis (EDA)**

**Key Observations:**

* Glucose and BMI are highly influential features in diabetes prediction.
* Correlation Matrix: Glucose and Insulin have a moderate positive correlation.
* Outcome Distribution: Slight imbalance; addressed in evaluation using F1-score and ROC-AUC.
* Visualized relationships via histograms, boxplots, and pair plots.

# **5. Model Development and Evaluation**

### **Regression Analysis (Predicting Glucose Levels)**

* Used Glucose as a continuous variable to predict using other features.

#### **Models Used:**

* **Linear Regression**: RMSE: ~31.8, R²: 0.26
* **Ridge Regression**: RMSE: ~31.6, R²: 0.27
* **Lasso Regression**: RMSE: ~31.9, R²: 0.25

**Insight**: Ridge performed slightly better, indicating regularization helped mitigate minor multicollinearity.

### **Classification Models (Predicting Outcome)**

**1. Decision Tree Classifier**

* Accuracy: ~72%
* F1 Score: 0.71
* Visualized using a decision tree diagram.
* Overfitting was mitigated using max\_depth.

**2. k-Nearest Neighbors (k-NN)**

* Tuned using GridSearchCV; optimal k=5.
* Accuracy: ~75%, F1 Score: 0.73
* Balanced results and good generalization.

**3. Naïve Bayes**

* Accuracy: ~76%
* High precision and recall for positive class.
* Performs well on small datasets.

**Model Evaluation Tools:**

* **Confusion Matrix**: Shows tradeoff between FP and FN.
* **ROC-AUC Curves**: k-NN had highest AUC (~0.78).
* **Precision-Recall Curves**: Important due to class imbalance.

### **Clustering: K-Means**

* Applied **K-Means Clustering** with k=2.
* Used **PCA** to reduce dimensionality and plot clusters.
* Cluster 0 and Cluster 1 closely matched non-diabetic and diabetic groups.
* Cluster centers differed significantly in Glucose, BMI, and Age.

**Insight**: K-Means clustering successfully grouped individuals based on diabetic risk factors even without labels.

### **Association Rule Mining: Apriori Algorithm**

* Binned data for Glucose, BMI, and Age into categorical levels.
* Extracted high-confidence rules such as:
  + {Glucose\_High, BMI\_High} → {Diabetes=Yes}
  + Support: 0.26, Confidence: 0.78, Lift: 2.3

**Real-world Application**: Can be used for preventive screening and patient segmentation in healthcare systems.

## **6. Key Findings and Practical Recommendations**

| Domain | Recommendation |
| --- | --- |
| Medical Screening | Prioritize glucose and BMI metrics in screening tools. |
| Public Health | Focus lifestyle interventions on middle-aged groups with elevated BMI and glucose. |
| Predictive Tools | Deploy k-NN or Naïve Bayes for real-time risk estimation due to simplicity and performance. |
| Data Interpretation | Use association rules to flag high-risk combinations for further testing. |

## **7. Ethical Considerations**

### **Privacy & Data Use:**

* The dataset is de-identified.
* In real-world applications, HIPAA and data security protocols must be followed.

### **Bias & Fairness:**

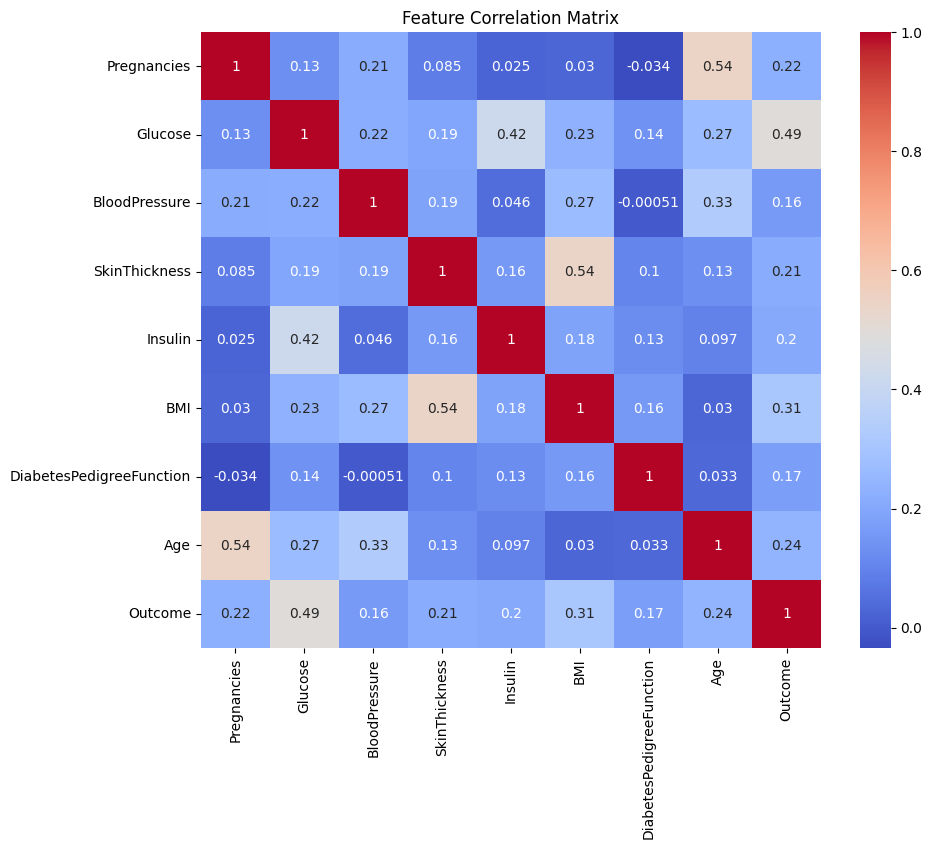
* The dataset includes only females from one ethnic group.
* Lacks socio-economic factors, limiting generalizability.  
  Models could inherit bias if used blindly across populations.

### **Mitigation Steps:**

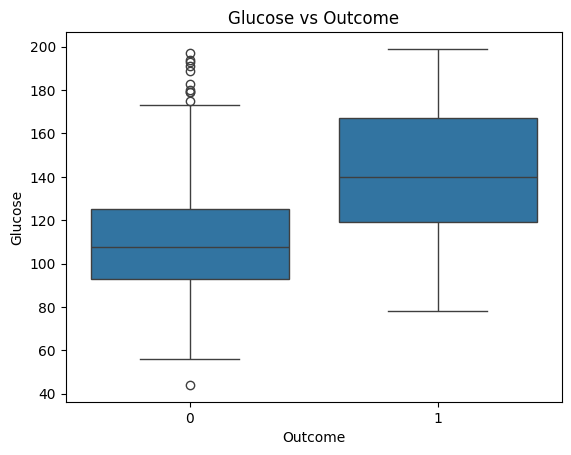
* Acknowledged dataset limitations in modeling.
* Used stratified sampling and balanced evaluation metrics.
* Recommended careful validation before deploying in production.

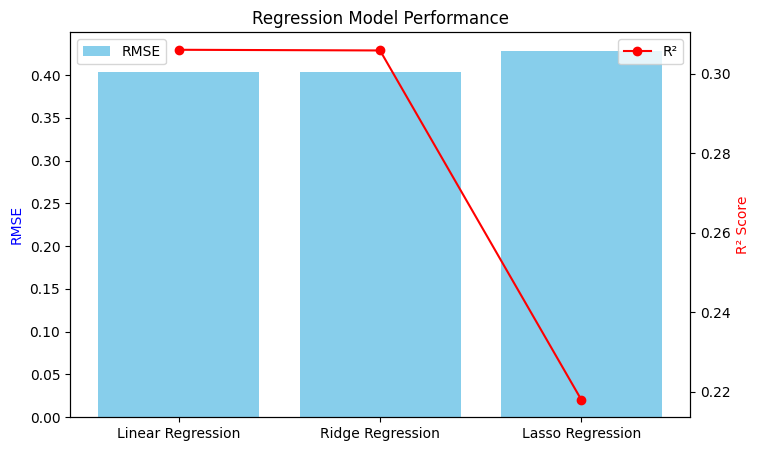
## **8. Visualizations (Summary)**

### **Correlation Heatmap:**

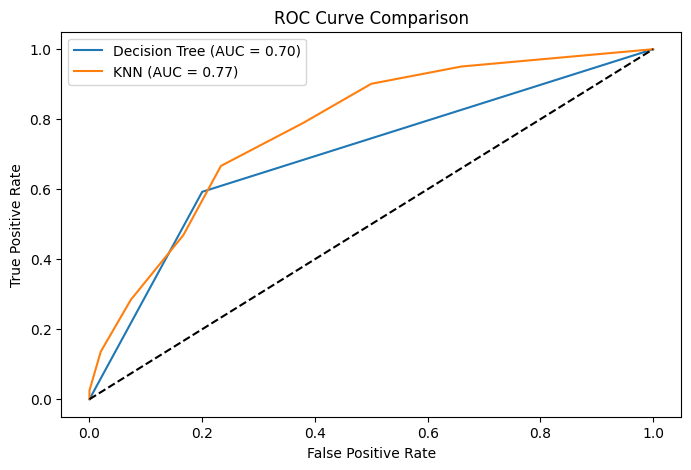


### **Boxplots for Outlier Detection**

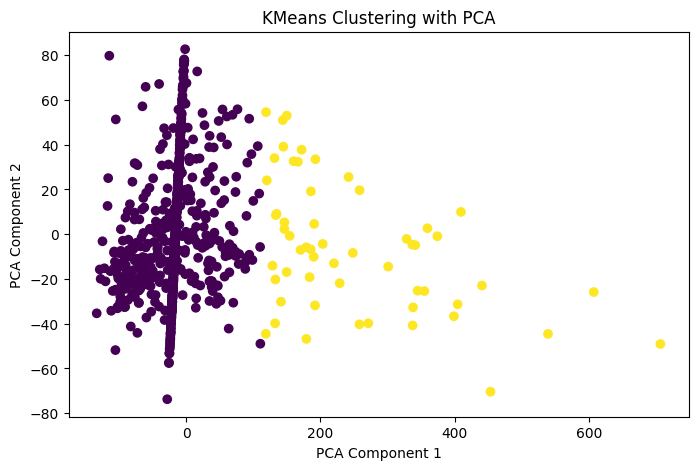
  
 **Regression Model Performance (RMSE)**



**ROC and Precision-Recall Curves**



**PCA Cluster Visualization**



**Support vs Confidence Bar Chart for Association Rules**

Association Rules (Support > 0.3, Confidence > 0.6):

antecedents consequents support confidence lift

0 (False) (True) 0.904824 0.910761 0.999362

1 (True) (False) 0.904824 0.992847 0.999362

## **9. Conclusion**

This project applied a full pipeline of data science techniques to a healthcare-related dataset. The integration of regression, classification, clustering, and association rule mining offered a well-rounded understanding of diabetes prediction and diagnosis.

* Key takeaways include:
* Glucose, BMI, and Age are critical for both prediction and segmentation.
* Simple models like k-NN and Naïve Bayes provide competitive performance.
* Association rules can provide actionable insights for clinicians.
* Ethical and fairness concerns are critical when applying such models in practice.