Predicting Diabetic Patients Health Outcomes

Authors: Rohan Khandelwal & Rupesh Sasmal

Department of Computer Science, Georgia State University

Course: Data Mining

Instructor: Prof. Jingyu Liu

Why Diabetes Prediction Matters

Diabetes is a global health crisis, affecting millions and leading to severe complications if undetected. Traditional screening often lacks personalization, missing critical opportunities for early intervention. Machine learning offers a transformative approach to identify at-risk individuals sooner and more accurately.

537M Adults Affected

Worldwide prevalence of diabetes underscores the urgency.

Preventing Complications

Early detection can significantly reduce long-term health issues.

Personalized Risk

Moving beyond generic methods to individualized risk assessment.

ML for Early Diagnosis

Leveraging AI to enhance diagnostic capabilities and stratification.

Project Objectives: Research Goals

Our project aims to harness the power of data mining and machine learning to improve diabetes prediction. We seek to understand the underlying medical factors, build robust predictive models, and uncover natural groupings within patient data.

Feature Analysis

Analyze key medical features and their correlation with diabetes outcomes.

Model Training

Develop and train classification models for accurate diabetes prediction.

Clustering

Apply unsupervised learning to identify distinct diabetes risk groups.

Performance Comparison

Evaluate and compare models to determine the most effective approach.

Dataset Overview & Preprocessing

We utilized the PIMA Indian Diabetes dataset, a foundational resource in diabetes research. Rigorous preprocessing steps were crucial to prepare the data for accurate model training and analysis.

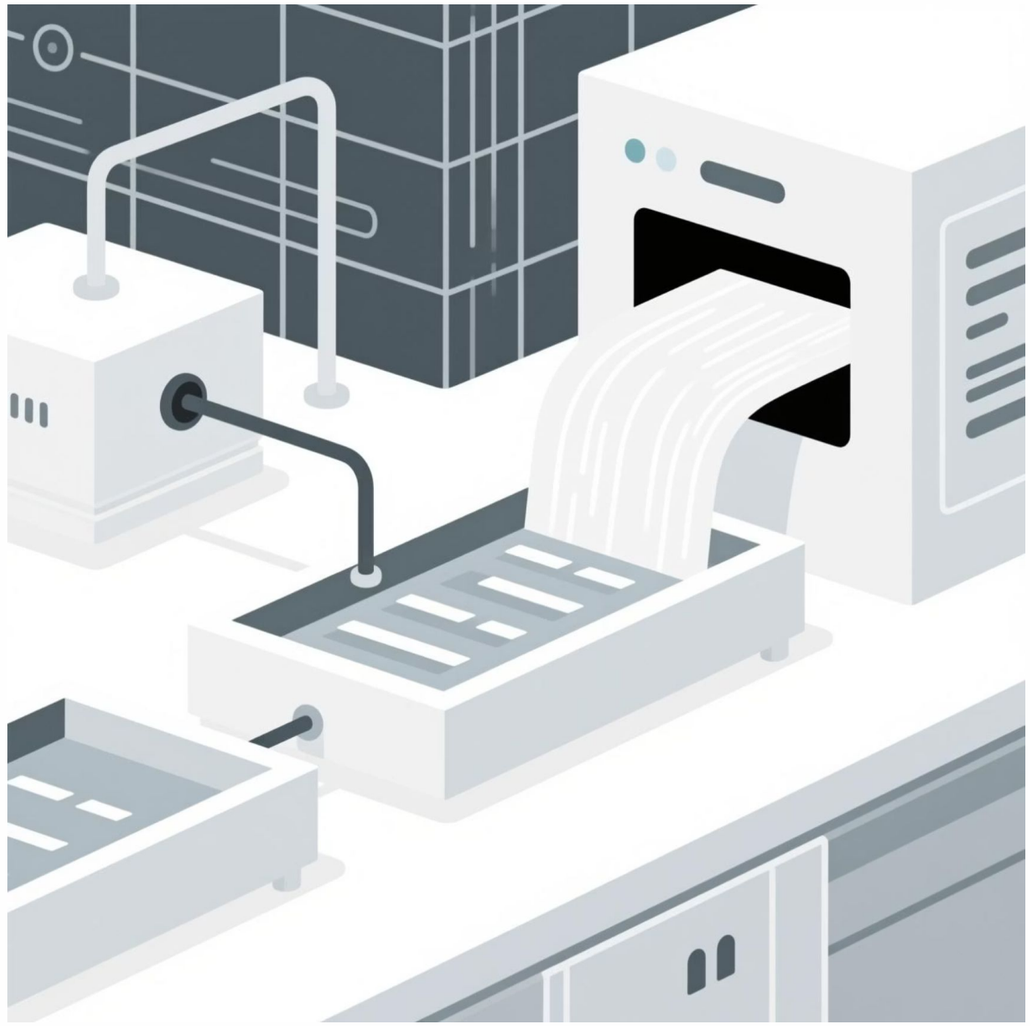
Dataset: PIMA Indian Diabetes (768 records)

Features: 9 clinical features + outcome variable

Class Imbalance: 65.1% Non-Diabetic, 34.9% Diabetic

Preprocessing Steps:

- Zero placeholders for vital metrics (e.g., Blood Pressure, BMI) were intelligently replaced with median values to prevent data loss and maintain statistical integrity.
- MinMax scaling normalized feature ranges, ensuring all variables contributed equally to the models and preventing features with larger values from dominating the learning process.
- An 80/20 train-test split (614 training samples / 154 testing samples) was applied to evaluate model generalization effectively.



Exploratory Data Analysis: Uncovering Key Insights

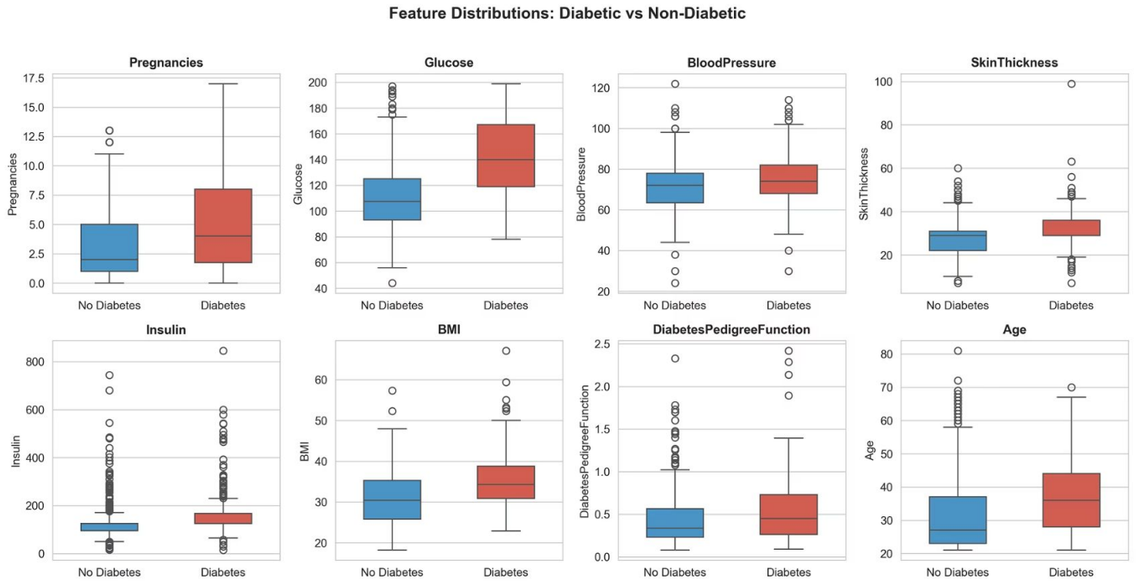
Through visual analysis of histograms and boxplots, we identified crucial patterns within the PIMA dataset, highlighting features strongly associated with diabetes presence.

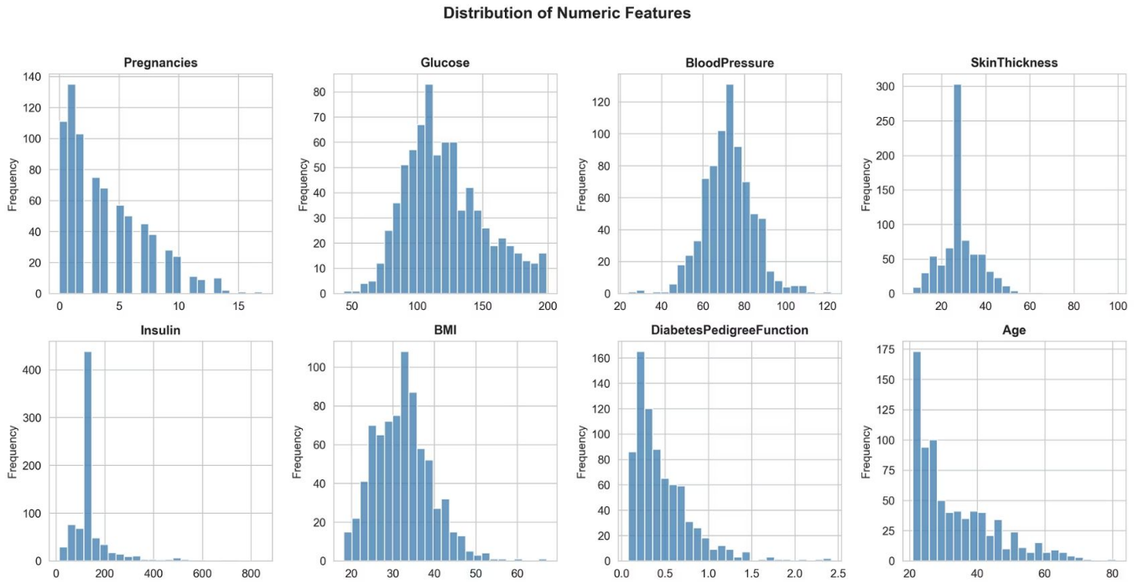
Key Insights:

Glucose & BMI: Significantly higher values in individuals with positive diabetes outcomes, indicating their strong predictive power.

Insulin & Pedigree Function: Exhibited high skewness, suggesting non-normal distributions that might require specific handling in modeling.

Age & Pregnancies: A noticeable correlation between age and the number of pregnancies, reflecting demographic patterns within the PIMA population.

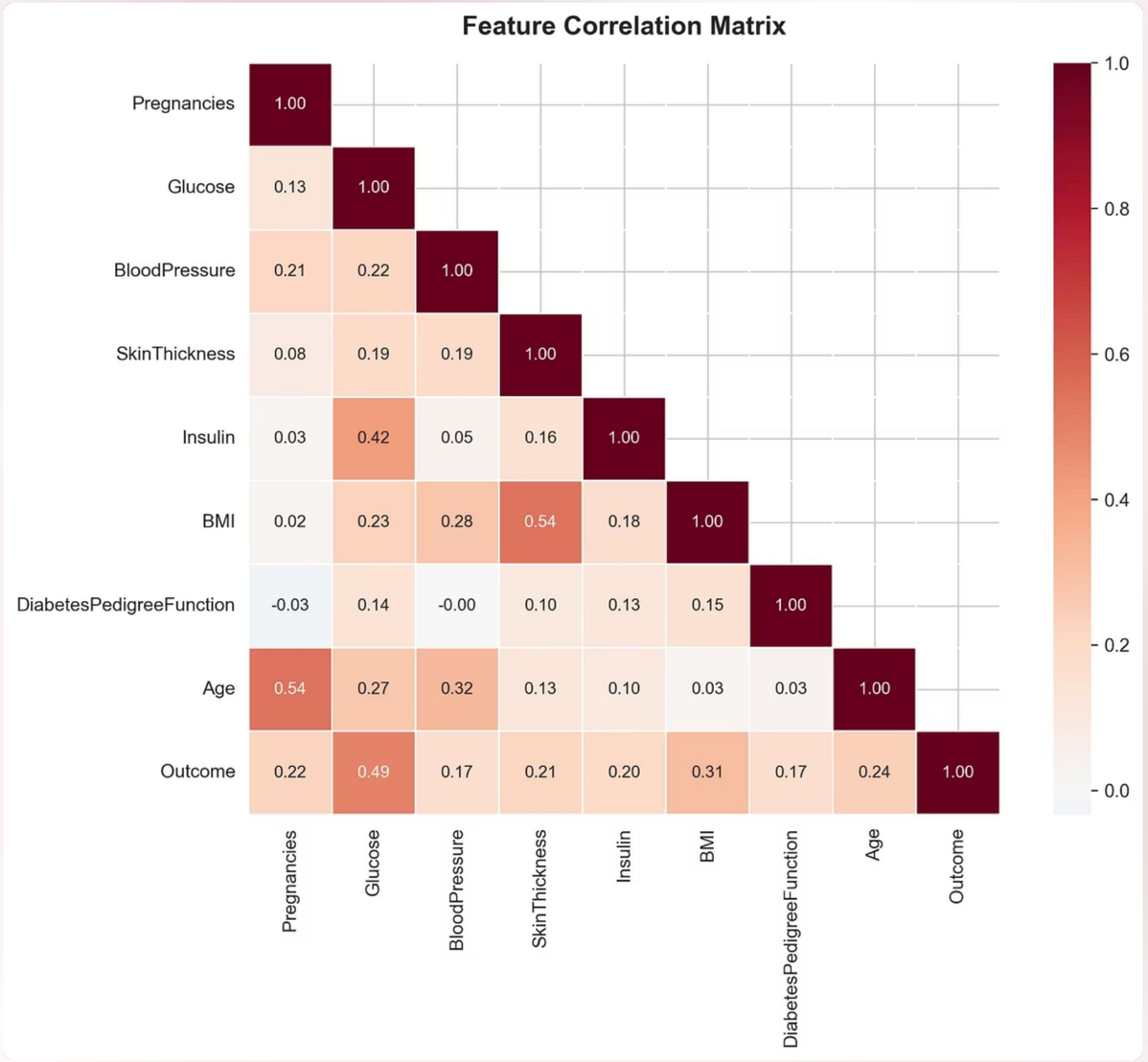




Feature Correlation Analysis

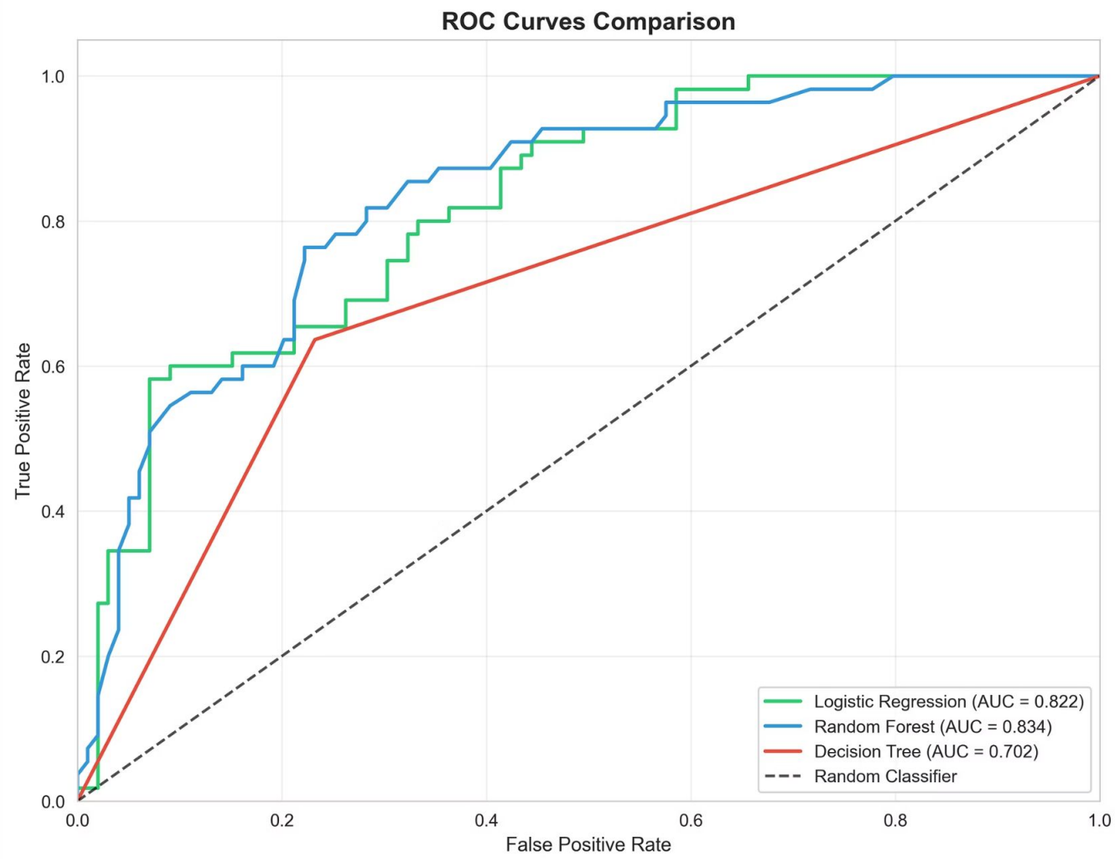
A correlation heatmap vividly illustrates the relationships between different clinical features and the diabetes outcome, guiding our understanding of feature importance.

> **Glucose** emerged as the most strongly correlated feature with diabetes outcome, underscoring its primary role in diagnostic assessment.
>
> **BMI** (Body Mass Index) and **Age** also demonstrated significant positive correlations, reinforcing their established importance as contributing factors to diabetes risk.



Model Training & Evaluation: Predictive Performance

We trained and evaluated several classification models, assessing their ability to predict diabetes outcomes based on F1-Score and ROC-AUC metrics. The ROC Curve visually represents the trade-off between true positive rate and false positive rate.



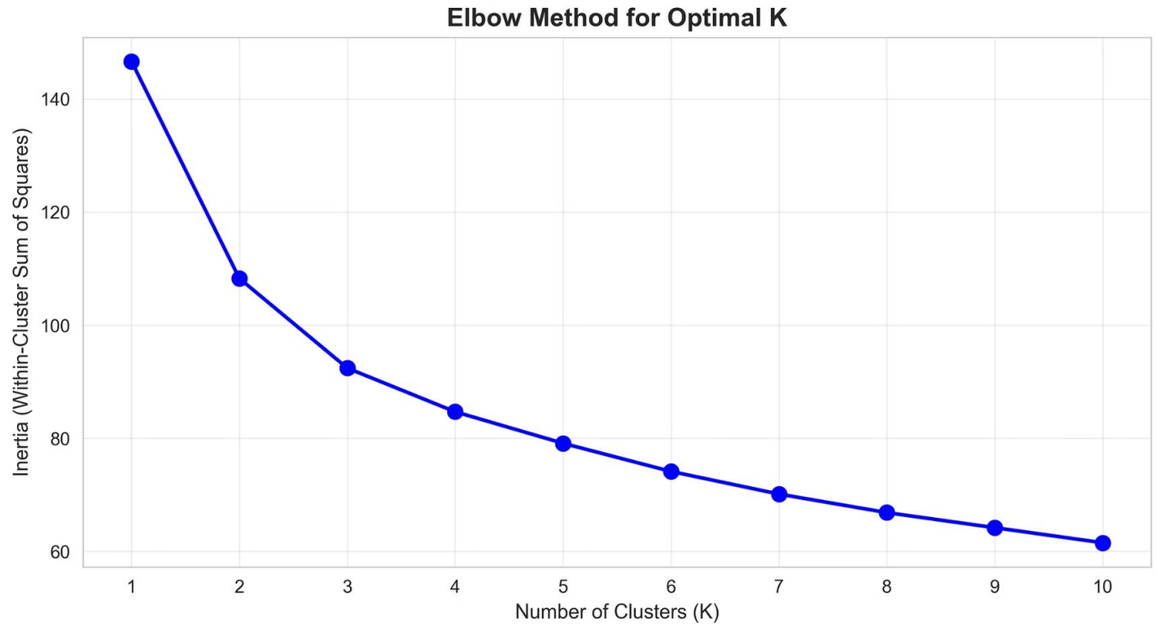
Logistic Regression	0.66	0.83
Random Forest	0.61	0.79
Decision Tree	0.55	0.71

Best Model: Logistic Regression

Logistic Regression outperformed other models, primarily due to the linear separability of the dataset's features related to diabetes, making it a robust choice for clinical prediction.

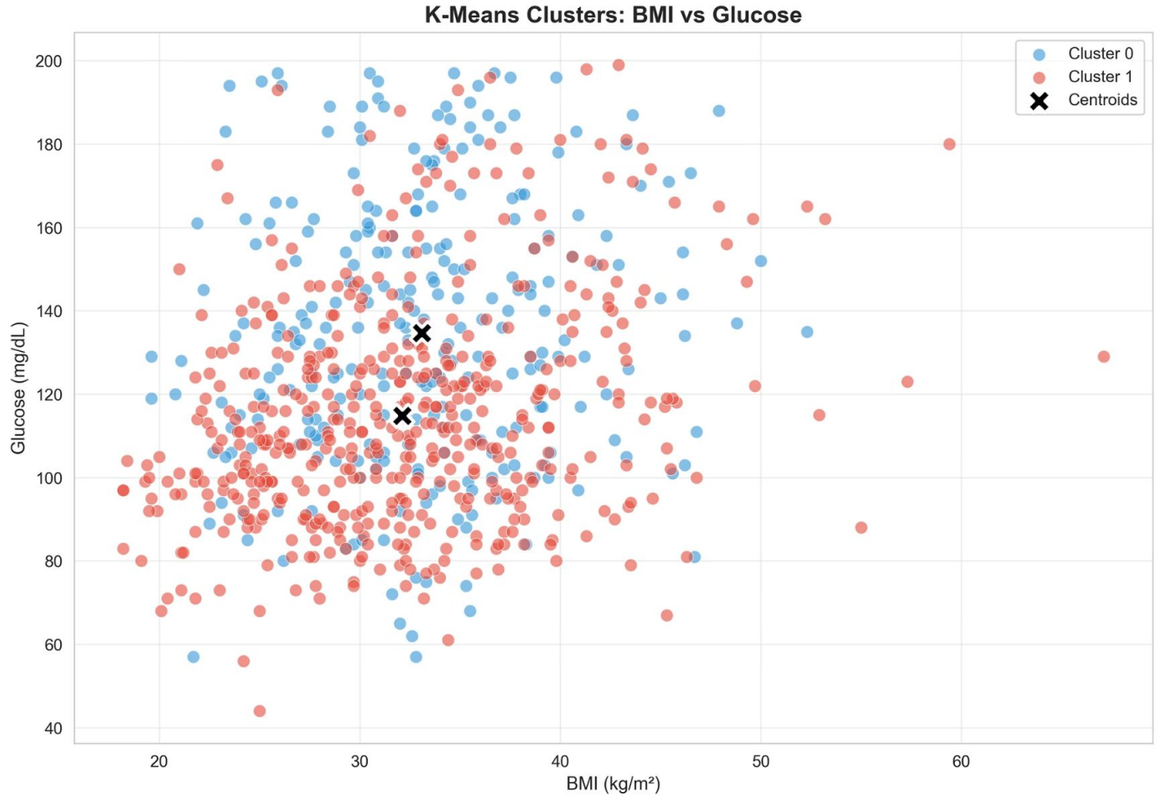
Clustering (Unsupervised Learning): Discovering Patient Sub-Populations

Unsupervised clustering revealed inherent groupings within the patient data, offering valuable insights into different diabetes risk profiles. The Elbow Method determined the optimal number of clusters.



Optimal Cluster Size: K=2

The Elbow Method, a widely-used technique for determining optimal cluster count, plots the within-cluster sum of squares (WCSS) against the number of clusters. The resulting curve exhibits a characteristic 'elbow' point where the rate of WCSS decrease significantly diminishes. In our analysis, this elbow point clearly manifested at K=2, indicating that two clusters provided the most meaningful separation of the data while minimizing information loss and avoiding over-segmentation.



Identified Risk Clusters:

High-Risk Cluster: Showed a significant diabetes rate of 54%, identifying individuals requiring closer monitoring and early intervention.

Low-Risk Cluster: Exhibited a diabetes rate of 25%, representing a population with lower immediate concern but still needing awareness.

Conclusions & Future Work

Key Conclusions

- Machine learning models significantly enhance diabetes risk screening capabilities.
- Logistic Regression proved to be the most effective model for clinical prediction, offering a balance of accuracy and interpretability.
- Clustering successfully identified distinct patient sub-populations, crucial for targeted intervention strategies.

Future Enhancements

- Incorporate additional clinical features, such as **HbA1c levels, dietary habits, and exercise routines**, for a more comprehensive predictive model.
- Explore advanced algorithms like **XGBoost** and rigorous **cross-validation** techniques to further boost accuracy and model robustness.

Acknowledgment & Questions

We extend our sincere gratitude to:

- Professor **Jingyu Liu** for invaluable guidance and support throughout the coursework.
- The **UCI / Kaggle Diabetes Dataset Source** for providing the essential data for our project.

Thank you for your attention!

Questions?