Pima Indians Diabetes Prediction Analysis

Khairul's Project Report

Project Overview

This project analyzed the Pima Indians Diabetes dataset to build predictive models for diabetes diagnosis using logistic regression. The dataset contains 768 samples with 8 medical predictor variables and a diabetes outcome (34.9% prevalence).

Key Findings

1. Most Predictive Single Variable: Glucose

- **Accuracy: 76.8%**
- **AUC: 0.767**
- Glucose level is the strongest individual predictor of diabetes
- Model: *Diabetes Risk = -0.778 + 1.245 × (Glucose)*

2. Most Predictive Variable Pair: Glucose + BMI

- **Accuracy: 71.4%**
- **AUC: 0.779**
- Adding BMI to glucose improves prediction reliability
- Model: *Diabetes Risk = -0.634 + 1.167 × (Glucose) + 0.641 × (BMI)*

3. All Features Model Performance

- **Accuracy: 71.4%**
- **AUC: 0.823**
- Using all 8 variables provides the most robust predictions
- Best overall model despite similar accuracy to the pair model

---

Feature Importance Ranking

1. **Glucose** (1.144) - Blood sugar level
2. **BMI** (0.714) - Body mass index
3. **Pregnancies** (0.373) - Number of pregnancies

4. **DiabetesPedigree** (0.256) - Family history factor
5. **Blood Pressure** (0.198) - Diastolic blood pressure
6. **Age** (0.184) - Patient age
7. **Insulin** (0.127) - Insulin level
8. **Skin Thickness** (0.067) - Triceps skin fold

Model Comparison

Model Type	Variables Used	Accuracy	AUC Score
Single Variable	Glucose only	**76.8%**	0.767
Best Pair	Glucose + BMI	71.4%	**0.779**
All Features	All 8 variables	71.4%	**0.823**

Conclusions

1. **Glucose is the dominant predictor** - As expected, blood glucose level is the most important single factor for diabetes prediction
2. **Diminishing returns from additional variables** - While the full model has the best AUC, the accuracy improvement over simpler models is minimal
3. **Practical recommendations**:
    - For **quick screening**: Use glucose level alone (76.8% accuracy)
    - For **balanced approach**: Use glucose + BMI (good accuracy with simplicity)
    - For **comprehensive diagnosis**: Use all features (highest reliability)
4. **Clinical insight**: The model confirms that glucose monitoring remains the cornerstone of diabetes diagnosis, with BMI as an important secondary factor

---

Technical Notes

- Used logistic regression (appropriate for binary classification)
- Applied feature scaling for optimal performance
- Maintained 80/20 train-test split with stratification
- All models show good discriminatory power (AUC > 0.75)