Diabetes Prediction Using Machine Learning

- Presented by: Medha Chawla
- MIT Manipal | CSE (AI-ML)
- Roll No: 55 | Reg No: 220962340

Abstract

- Diabetes is a growing global health concern.

- Goal: Use ML to predict diabetes from health data.

- Dataset: 2,000 records, 9 features.

- Algorithms used: LR, DT, SVM, RF.

- Best: Random Forest (95.25% accuracy).

Introduction

- Diabetes Mellitus: Chronic high blood sugar.

- Causes heart, kidney, eye issues.

- ML helps in early detection and quicker diagnosis.

- Goal: Analyze medical data using ML models.

Objectives

Understand and clean

- Understand and clean dataset.

Rename

- Rename confusing features.

Balance

- Balance data and correct invalid values.

Perform

- Perform EDA and visualize distributions.

Train, tune, and compare

- Train, tune, and compare models.

Select

- Select the best performing model.

Literature Review - Part 1

Initial research used statistical models:

- *Logistic Regression* (Smith et al., 1988) on the Pima Indian Diabetes dataset
- *Linear Discriminant Analysis* (LDA) for basic classification tasks

- Limitations of early approaches:

- Poor at modeling non-linear relationships
- Depended on linear assumptions that don't reflect complex medical data

- Importance:

- These foundational methods paved the way for ML-based diagnostic systems by establishing baseline expectations for performance and feasibility.

Literature Review - Part 2

Machine learning adoption grew in the 2010s:

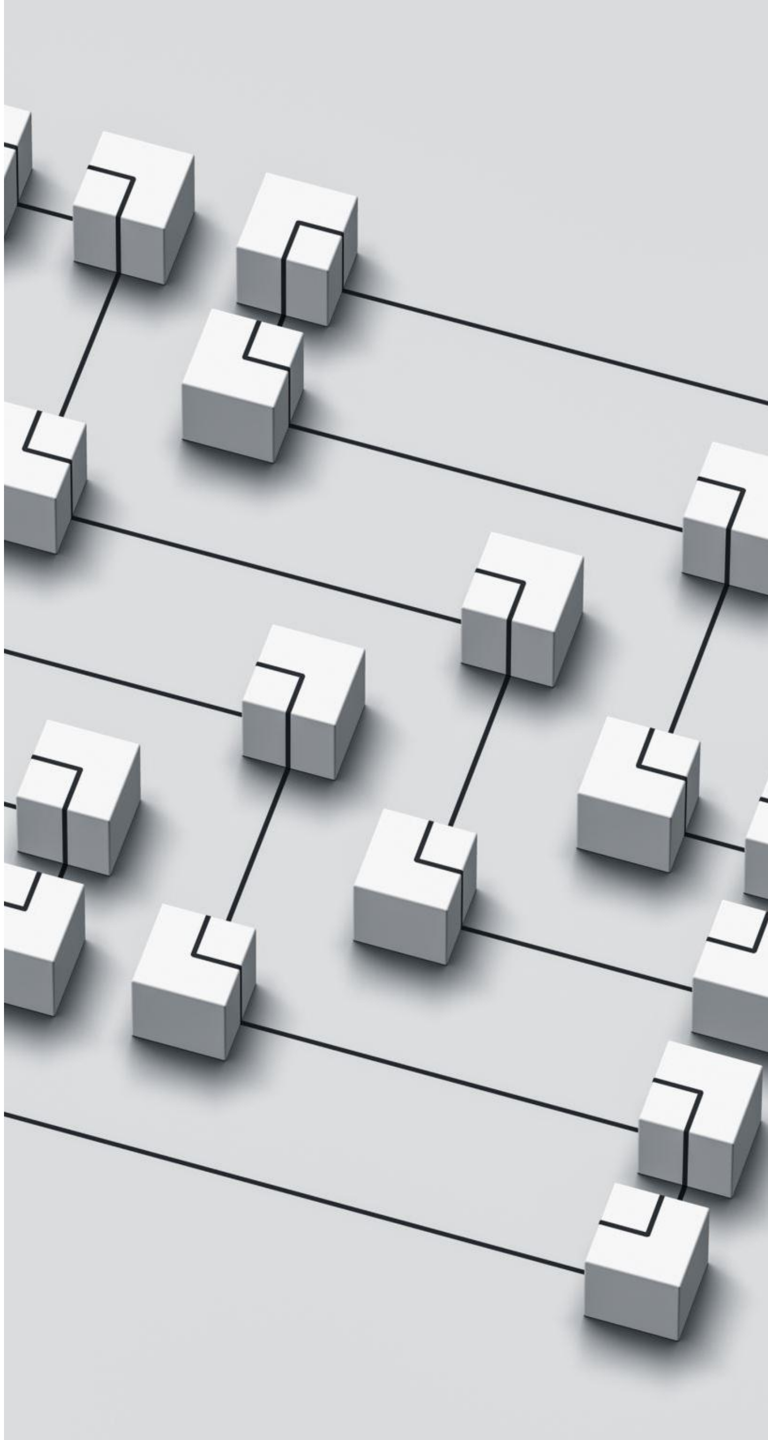
- *Patil et al. (2010)* built a decision support system combining Decision Trees and probabilistic techniques
- *Kumari & Chitra (2013)* applied Support Vector Machines (SVMs) with kernel functions for better handling of non-linearity

- Ensemble methods gain popularity:

- *Sisodia & Sisodia (2018)* used Random Forests, achieving improved accuracy and robustness
- *Jiang et al. (2020)* optimized Random Forests using hyperparameter tuning and cross-validation, reaching around 94% accuracy

Research Gaps

- - Overuse of single (Pima) dataset.
- - Poor data cleaning.
- - No hyperparameter tuning.
- - Weak validation (no cross-validation).
- - Models act as 'black boxes'.



Methodology Overview

1. Load and inspect dataset.
2. Handle missing/unrealistic values.
3. Rename features for clarity.
4. Visual and statistical EDA.
5. Train-test split (80/20).
6. Train and tune models.
7. Final model selection.

Dataset Overview

Source:
• Kaggle (Pima Indian Diabetes dataset)
• 2,000 patient records, all numeric features

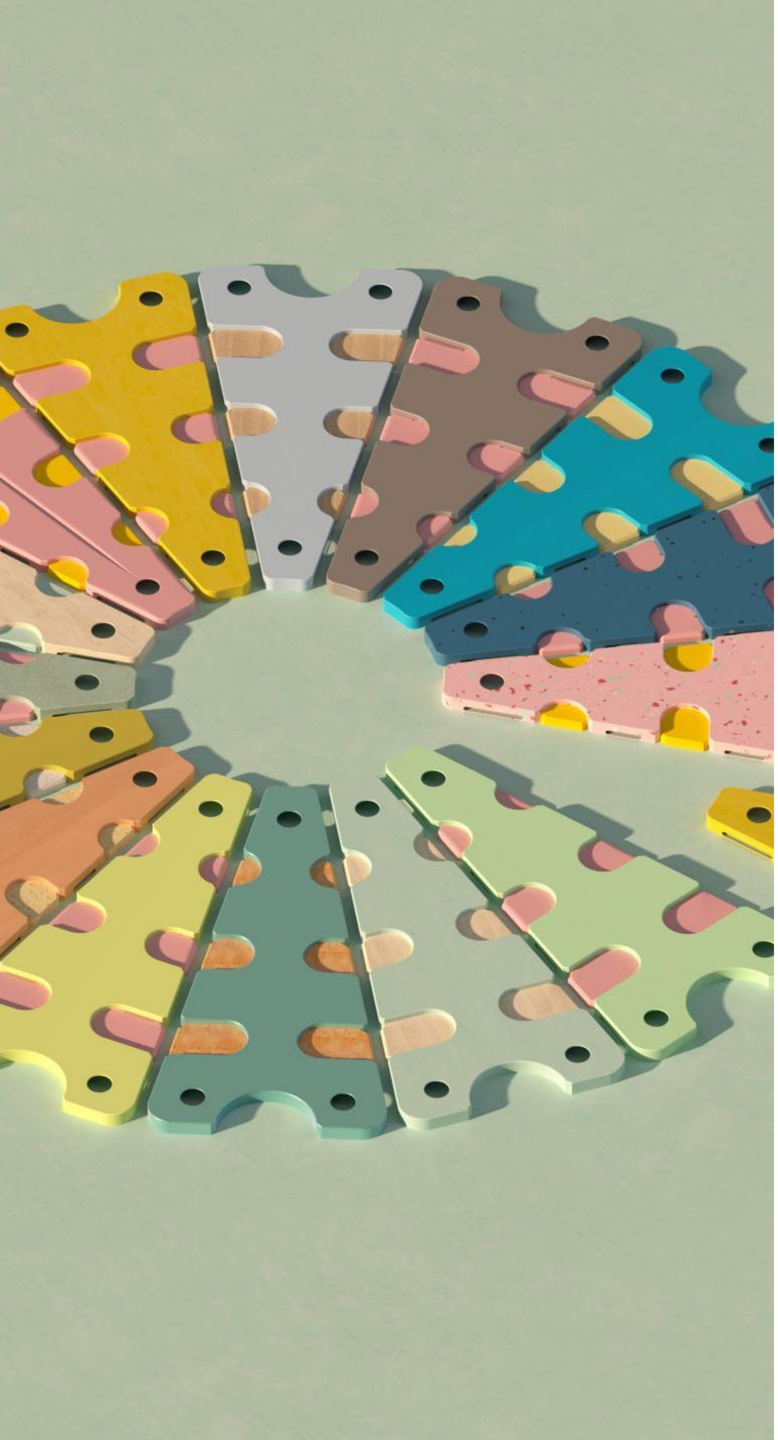
Features (8 Predictors):
• *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*
• *Insulin*, *BMI*, *DPF* (family history), *Age*

Target (Outcome):
• 0 – Non-diabetic
• 1 – Diabetic

Key Notes:
• Invalid values (e.g., 0 in BMI/Glucose) were replaced using mean imputation
• Class imbalance observed (~2/3 non-diabetic)

Data Preprocessing

- Zero values marked as NaN.
- Imputed using mean/median.
- Renamed features for clarity.
- Normalized for sensitive algorithms.



Models Used

- - **Logistic Regression**: Baseline model.
- - **Decision Tree**: Captures non-linear patterns.
- - **SVM:** Good class separation.
- - **Random Forest**: Ensemble of decision trees.

Hyperparameter Tuning

- - Used GridSearchCV.
- - Tuned depth, estimators, C, kernels.
- - Used cross-validation to avoid overfitting.

Model Performance

- Logistic Regression: 76.31%

- Decision Tree: 90.50%

- SVM: 86.93%

- Random Forest: 95.25% (best)

Model Input & Output

Input: 8 health features (e.g. glucose, BMI).

Output:

- 0 = Non-diabetic
- 1 = Diabetic

Optional: Probability score

Conclusion

- ML successfully used to predict diabetes.
- Random Forest = best model.
- Clean data + tuning = high accuracy.
- Practical for early screening in real world.

Thank you!

Name: Medha Chawla

Reg No. 220962340

Btech.(CSE-AIML)

medha.mitmpl2022@learner.manipal.edu