Predicting Diabetes Outcomes

David Robertson

Introduction

What is Diabetes?

- Diabetes is a chronic medical condition where the body cannot properly regulate blood sugar (glucose) levels.
    - **Type 1 Diabetes**: The body produces little to no insulin (an autoimmune condition).
    - **Type 2 Diabetes**: The body becomes resistant to insulin or doesn't produce enough.
    - **Gestational Diabetes**: Occurs during pregnancy and can lead to Type 2 diabetes later.

Introduction

How Does Diabetes Affect People?

- **Physical Health**
  - High blood sugar can damage blood vessels, nerves, and organs over time.
  - Leads to complications such as:
    - Heart disease
    - Stroke
    - Kidney failure
    - Vision problems (diabetic retinopathy)
    - Amputations due to poor circulation.
- **Mental and Emotional Health**
  - Managing diabetes can lead to stress and anxiety.
  - Risk of depression is higher among individuals with diabetes.

Introduction

Prevalence

- Over **37 million Americans** have diabetes.
- About **96 million** adults have prediabetes.

Economic Burden

- The total cost of diabetes care in the U.S. exceeds **$327 billion annually**.
  - $237 billion in direct medical costs.
  - $90 billion in lost productivity.

Health Disparities

- Higher prevalence in underserved communities, due to limited access to healthcare and nutritious food.

Early Prediction is Key: Detecting and managing diabetes early can reduce complications and improve quality of life.

Personal and Societal Impact: Better predictions allow healthcare providers to allocate resources effectively, improve outcomes, and reduce costs.

Mission

- **Mission Statement**
  - The goal is to leverage machine learning techniques to explore and address a critical healthcare challenge of predicting the progression of diabetes in patients. By analyzing clinical data, the aim is to uncover insights that can enhance early intervention strategies and improve patient outcomes.

- **Research Question**
  - How accurately can machine learning models predict the progression of diabetes in patients one year after their baseline assessment based on clinical and serum measurement data?

- **Hypothesis**
  - There is a significant relationship between body mass index (BMI), blood pressure (BP), and other measurements and the progression of diabetes. Specifically, machine learning models can predict disease progression with statistically significant accuracy using these predictors.

- **Rationale**
  - By identifying patterns in key predictors like BMI, BP, and serum measurements, this project aims to contribute to the development of predictive tools that support healthcare professionals in early diagnosis and personalized treatment planning, ultimately improving the quality of life for individuals living with or at risk of diabetes.

Data Overview

- **Source**: Diabetes dataset from scikit-learn.
- **Size**:
  - **442 samples**.
  - **10 features** capturing clinical and serum measurements.

- **Features**: Includes quantitative predictors such as:
  - Age
  - Sex
  - Body Mass Index (BMI)
  - Blood Pressure (BP)
  - Serum measurements (S1, S2, S3, S4, S5, S6).

- **Target Variable**
  - **Diabetes Progression**: Quantitative measure of disease progression one year after baseline assessment.

- **Mean-Centered and Standardized**:
  - All features were transformed to have a mean of 0 and a standard deviation of 1.
  - Ensures that features are on the same scale, improving model performance and stability.

- **Why Standardization?**
  - Allows machine learning models to treat all features equally.
  - Essential for algorithms sensitive to feature scales, like Linear Regression.

Tableau Exploration

- **Objective**
  - Understand the relationship between BMI and Target and how this relationship varies by sex
- **Key Variables**
  - BMI: Independent variable (predictor)
  - Target: Dependent variable (outcome of interest)
  - Sex: Categorical factor (male/female) influencing the relationship
- **BMI-Target Relationship**
  - A statistically significant positive relationship exists between BMI and Target ($p < 0.0001$)
  - As BMI increases, Target also increases for both males and females
- **Sex-Specific Differences**
  - **Males:** For every unit increase in BMI, Target increases by 54.75
  - **Females**: For every unit increase in BMI, Target increases by 39.10
  - This indicates BMI has a stronger effect on males than on females

Tableau Exploration

- **Model Performance**
  - A general upward trend is visible for both sexes, confirming the positive relationship between BMI and Target
  - The model explains about 35.4% of the variability in Target (R-squared = 0.354), indicating a moderate fit
- **Remaining Variability**
  - The remaining variability suggests there are other factors influencing Target that we did not capture, such as lifestyle, age, or medical conditions
- **Standard Error**
  - The model has a standard error of 62.17, reflecting the variability in actual Target values

Methodology

Step 1: Data Acquisition and Exploration

- **Initial Exploration**:
  - Examined dataset size, features, and target variable.
  - Visualized data distributions and relationships between predictors and the target.

Step 2: Data Preprocessing

- **Standardization**:
  - Mean-centered and scaled all features to have a mean of 0 and standard deviation of 1.
  - Ensured uniform feature scales for effective model training.
- **Train-Test Split**:
  - Split the data into **80% training** and **20% testing** sets to evaluate model performance.

Step 3: Model Selection and Training

- **Chosen Models**:
  - Linear Regression (baseline model for interpretability).
  - Random Forest Regressor (to capture nonlinear relationships).
  - Gradient Boosting (for enhanced predictive performance)
- **Hyperparameter Tuning**:
  - Used Grid Search to optimize model parameters.

Methodology

Step 4: Model Evaluation

- **Performance Metrics**
  - **R^2 (Coefficient of Determination):** Measures how well the model explains variance in the target variable.
  - **Mean Absolute Error (MAE):** Evaluates prediction accuracy by measuring average error.
  - **Mean Squared Error (MSE):** The overall size of the errors your model makes.
- **Insights**:
  - Compared models to determine which provided the best balance of interpretability and accuracy.

Step 5: Visualization

- Presented key findings using scatterplots, error distributions, and model comparison charts.
- Tableau for an interactive dashboard.

Results

Model	MAE	MSE	R^2
Baseline (Linear Regression)	42.79	2900.19	0.45
Default Random Forest	44.38	3004.35	0.43
Optimized Random Forest	43.78	2975.17	0.44
Gradient Boosting	42.84	2886.36	0.46

- **Gradient Boosting** had the best overall performance, achieving the lowest MSE and highest R^2.
- **Baseline Linear Regression** was a close second, showing comparable performance but less flexibility with nonlinear data.
- **Random Forest** models, though slightly less accurate, provided robust performance with lower variance in predictions.

Results

Feature Importance

- **Key Predictors**:
  - **BMI**: Most influential predictor of diabetes progression.
  - **S5:** Strong correlation with blood sugar regulation.
  - **BP**: Moderate contribution.

Insights from Visualizations

- **Scatterplots**:
  - Gradient Boosting showed the closest alignment between actual and predicted values.
- **Feature Importance Chart**:
  - Gradient Boosting highlighted BMI, S5, and BP as critical features.
- **Error Distributions**:
  - Gradient Boosting achieved the most consistent prediction accuracy, with fewer outliers compared to other models.

Conclusion

- **Key Takeaways**
  - Gradient Boosting demonstrated the best performance, achieving the highest R^2 (0.46) and the lowest MSE (2886.36).
  - Key predictors like BMI, blood pressure, and S5 play a significant role in predicting diabetes progression.
  - Standardizing features and iterative model improvement were critical for success, especially given the small dataset size.

- **Improvements**
  - **Expand the Dataset**
    - Incorporate more samples from diverse populations to improve model generalizability.
    - Explore additional features, such as genetic markers or lifestyle factors.
  - **Experiment with Advanced Techniques**
    - Implement neural networks or ensemble methods.
    - Use feature engineering to create new predictors.
  - **Deploy the Model**
    - Create an interactive web application or dashboard to provide real-time diabetes progression predictions for clinicians.
  - **Collaborate with Healthcare Experts**
    - Validate the model’s findings and predictions with medical professionals.

References

- American Diabetes Association. (2022). *Economic costs of diabetes in the U.S. in 2022*. Retrieved from https://diabetes.org

- Centers for Disease Control and Prevention. (2022). *National diabetes statistics report, 2022*. U.S. Department of Health and Human Services. Retrieved from https://cdc.gov/diabetes

- Mayo Clinic. (n.d.). *Diabetes*. Retrieved January 8, 2025, from https://mayoclinic.org

- National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). *Diabetes statistics*. National Institutes of Health. Retrieved January 8, 2025, from https://niddk.nih.gov

- World Health Organization. (2022). *Diabetes fact sheet*. Retrieved from https://who.int