Executive Summary: Diabetes Risk Prediction Model

Project Overview

The Diabetes Analysis Project aimed to develop a predictive model for diabetes diagnosis using health indicators from the Pima Indians Diabetes dataset. This six-member collaborative project followed a comprehensive machine learning workflow from data cleaning to advanced model optimization. The resulting model provides valuable insights for healthcare professionals to identify patients at risk of diabetes.

Key Business Questions Addressed

1. **What factors are most predictive of diabetes risk?**
2. **How accurately can we predict diabetes using available health metrics?**
3. **What combination of models yields the most reliable predictions?**
4. **Can derived features improve prediction accuracy?**
5. **What decision threshold balances sensitivity and specificity optimally?**

Data Insights

The analysis revealed several important patterns in the data:

- **Disease Prevalence**: The dataset contains approximately 35% diabetic cases and 65% non-diabetic cases.
- **Age Correlation**: Diabetes prevalence increases significantly with age, with the highest rates in the 60+ age group.
- **BMI Impact**: Obese individuals (BMI > 30) show a substantially higher diabetes risk compared to those with normal BMI.
- **Glucose Levels**: Plasma glucose concentration shows the strongest association with diabetes diagnosis.
- **Feature Relationships**: Strong correlations exist between certain features (e.g., Age and Pregnancies, BMI and Insulin).

Feature Engineering Contribution

The feature engineering phase substantially improved model performance through:

1. **Creation of derived metrics** like Diabetes_Risk_Index (BMI * Glucose / 100)
2. **Feature transformation** using various scaling techniques
3. **Feature selection** to identify the most predictive variables
4. **Dimensionality reduction** to capture complex patterns in the data

The most impactful engineered feature was the Diabetes_Risk_Index, combining the strong predictive power of both BMI and glucose levels.

Model Performance

After evaluating multiple machine learning algorithms and optimization techniques:

Model	Accuracy	Precision	Recall	F1 Score
Optimized Gradient Boosting	85%	77%	68%	72%
Ensemble Model	85%	78%	67%	72%
Random Forest	82%	74%	63%	68%
Logistic Regression	81%	72%	59%	65%

The **Optimized Gradient Boosting** and **Ensemble Model** demonstrated the best overall performance, balancing accuracy with the ability to identify true diabetes cases.

Threshold Optimization

Analysis of different decision thresholds revealed:

- The default 0.5 threshold favors precision over recall
- A threshold of 0.35 optimizes the F1 score
- A threshold of 0.30 provides better balanced sensitivity and specificity

Healthcare providers can select the appropriate threshold based on their specific clinical priorities.

Implementation Recommendations

Based on our analysis, we recommend:

1. **Deploy the Ensemble Model** for diabetes risk assessment
2. **Focus on monitoring** the top five predictive features:
    - Plasma glucose concentration
    - BMI
    - Diabetes_Risk_Index (our engineered feature)
    - Age
    - DiabetesPedigreeFunction (genetic score)
3. **Implement different thresholds** for different clinical scenarios:
    - Screening (lower threshold): Maximize recall to identify all potential cases
    - Diagnosis confirmation (higher threshold): Maximize precision to reduce false positives

Limitations & Future Work

While the model performs well, some limitations should be noted:

- The dataset only includes female patients of Pima Indian heritage
- Several important diabetes risk factors are not included (HbA1c, diet, exercise)
- The sample size is relatively small for machine learning applications

Future development should focus on:

1. **Expanding the dataset** to include more diverse populations
2. **Incorporating additional variables** like HbA1c levels and lifestyle factors
3. **Developing a real-time application** for clinical decision support
4. **Investigating time-series analysis** to track diabetes progression

Team Contributions

This project succeeded through the specialized efforts of six team members:

- **Data Engineer**: Cleaned and processed raw data, ensuring high-quality inputs
- **EDA Analyst**: Uncovered critical patterns and visualized key relationships
- **Feature Engineer**: Created innovative derived metrics that improved prediction
- **Model Developer**: Built and evaluated various machine learning approaches
- **Model Optimizer**: Fine-tuned models and created performance-enhancing ensembles
- **Project Manager**: Coordinated efforts and documented findings

Conclusion

The diabetes prediction model developed through this collaborative project demonstrates strong potential as a clinical decision support tool. With 85% accuracy and the ability to identify high-risk patients, the model can help healthcare providers prioritize interventions and improve diabetes management strategies. The insights gained about key risk factors also contribute valuable knowledge to diabetes research and prevention efforts.