Diabetes Analysis Project

Overview

This collaborative data science project analyzes diabetes health indicators to build a predictive model for diabetes diagnosis. The project uses the Pima Indians Diabetes dataset and involves a team of six members, each focusing on specific aspects of the machine learning pipeline.

Team Structure

Our team consists of six members with the following roles:

1. **Data Engineer** - Handles data cleaning, preprocessing, and preparation
2. **Exploratory Data Analyst** - Performs exploratory data analysis and visualization
3. **Feature Engineer** - Creates, selects, and transforms features
4. **Model Developer** - Builds and evaluates different machine learning models
5. **Model Optimizer** - Fine-tunes models and creates ensemble solutions
6. **Project Manager/Technical Writer** - Coordinates team efforts and prepares documentation

Project Structure

diabetes-analysis-project/

├── data/

│ └── diabetes_raw.csv

├── notebooks/

│ ├── 1_data_cleaning.ipynb

│ ├── 2_eda.ipynb

│ ├── 3_feature_engineering.ipynb

│ ├── 4_model_development.ipynb

│ └── 5_model_optimization.ipynb

├── visuals/

│ ├── static/

```
│   └── interactive/
├── outputs/
│   ├── cleaned_data.csv
│   ├── features_engineered.csv
│   ├── features_engineered_scaled.csv
│   ├── models/
│   │   ├── random_forest.pkl
│   │   ├── random_forest_optimized.pkl
│   │   └── ensemble_model.pkl
│   └── executive_summary.pdf
├── docs/
│   └── data_dictionary.md
│   └── requirements.txt
└── README.md
```

Workflow

The project follows a structured data science workflow:

1. **Data Cleaning** - Process raw data, handle missing values, and prepare for analysis
2. **Exploratory Data Analysis** - Uncover patterns, correlations, and insights in the data
3. **Feature Engineering** - Create new features and select the most relevant ones
4. **Model Development** - Build and evaluate various machine learning models
5. **Model Optimization** - Fine-tune hyperparameters and develop ensemble models

Key Findings

- The dataset contains health metrics for Pima Indian heritage females with approximately 35% diabetes cases.
- Glucose level is the strongest predictor of diabetes.
- BMI, Age, and Insulin also show strong correlations with diabetes outcomes.

- Engineered features like Diabetes_Risk_Index (BMI * Glucose / 100) improve model performance.
- The best-performing model achieved:
    - Accuracy: ~85%
    - Precision: ~77%
    - Recall: ~68%
    - F1 Score: ~72%

Installation & Setup

Prerequisites

- Python 3.8 or higher
- Required Python packages are listed in `requirements.txt`

Installation

1. Clone this repository
2. git clone https://github.com/your-username/diabetes-analysis-project.git
3. Navigate to the project directory
4. cd diabetes-analysis-project
5. Create and activate a virtual environment (optional but recommended)
6. python -m venv venv
7. source venv/bin/activate # On Windows: venv\Scripts\activate
8. Install dependencies
9. pip install -r docs/requirements.txt

Running the Notebooks

The notebooks should be run in the following order:

1. `1_data_cleaning.ipynb`
2. `2_eda.ipynb`
3. `3_feature_engineering.ipynb`
4. `4_model_development.ipynb`

Interactive Visualizations

The project includes several interactive visualizations located in the `visuals/interactive/` directory:

- Scatter plot matrix of key features
- Interactive boxplots comparing diabetic vs non-diabetic patients
- Feature radar chart for comparison

- ROC and PR curves for model evaluation

Model Performance

We evaluated multiple models to identify the best approach for diabetes prediction:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.82	0.74	0.63	0.68
Gradient Boosting	0.85	0.77	0.68	0.72
Logistic Regression	0.81	0.72	0.59	0.65
SVM	0.79	0.69	0.55	0.61
Ensemble Model	0.85	0.78	0.67	0.72

The optimized Gradient Boosting Classifier and Ensemble Model show the best overall performance.

Future Work

- Collect additional features like HbA1c levels and family history details
- Implement a web application for real-time diabetes risk assessment
- Explore deep learning approaches for feature extraction
- Incorporate time-series data for diabetes progression analysis
- Expand the study to include male patients and different ethnic groups

Contributing

Each team member should follow these guidelines:

1. Create a branch for your specific role/task
2. Follow the established code and documentation standards
3. Submit pull requests for review

4. Participate in code reviews for other team members
5. Document your process and findings in the appropriate notebook

License

……

Acknowledgments

- The original Pima Indians Diabetes dataset providers
- All team members for their valuable contributions
- Our project mentor for guidance and feedback