Diabetes Prediction: Multi-Class Classification

Ibrahim Alghrabi

2025-05-06

Experimental Setup

Dataset Overview

The dataset used for this assignment contains medical data for classifying 13 distinct types of diabetes. It comprises 70,000 observations with 34 features (13 numerical and 21 categorical) and no missing values. The target classes are relatively balanced, with approximately 5,000 observations per class. The dataset includes various medical indicators such as genetic markers, autoantibodies, insulin levels, BMI, blood pressure, and lifestyle factors.

Preprocessing Steps

The preprocessing pipeline consisted of the following steps:

68%, 17%, 15% 1. **Data Splitting**: The data was divided into training (68%), validation (17%), and testing (15%) sets.

2. **Feature Scaling**: Numerical features were standardized using `StandardScaler` to ensure all features contribute equally to the model training process.
3. **Categorical Encoding**: Categorical features were transformed using `OneHotEncoder` to convert them into a format suitable for machine learning algorithms. This was appropriate as the categorical features had relatively low cardinality.
4. **Feature Selection**: After initial model training, feature importance analysis was performed using Random Forest's built-in feature importance metric. The top 20 most important features were selected for the final models, which improved both efficiency and performance.

Model Configurations

Five classification models were implemented and evaluated:

1. **Logistic Regression**:
   - Penalty: L2 regularization
   - Solver: SAGA (suitable for multiclass problems)
   - Max iterations: 5000
2. **Support Vector Machine (Linear SVC)**:
   - Dual: False (optimized for cases where n_samples > n_features)
   - Max iterations: 2000
3. **Decision Tree**:

- Max depth: 5
- Min samples per leaf: 5
- Cost-complexity pruning alpha: 0.01

4. **Random Forest** (bagging ensemble):
    - Initial configuration:
        - Max depth: 7
        - Min samples per leaf: 5
    - Optimized configuration (after grid search):
        - Bootstrap: False
        - Max depth: 15
        - Feature selection: sqrt
        - Min samples split: 6
        - Min samples leaf: 1
        - Number of estimators: 600
5. **AdaBoost** (boosting ensemble):
    - Default configuration with base estimator

All models were evaluated using 5-fold cross-validation to ensure robust performance assessment.

Model Comparison

Performance Metrics

The models were evaluated using two key metrics:

1. **F1-Score (Macro)**: This metric provides a balanced measure of precision and recall, especially important for multi-class classification with potentially imbalanced classes.
2. **Accuracy**: The proportion of correctly classified instances, providing an overall measure of model performance.

Comparative Analysis

The performance of all implemented models is summarized in the table below:

Model	F1-Score (Train)	F1-Score (Test)	Accuracy (Train)	Accuracy (Test)
Logistic Regression	0.760	0.754	0.761	0.756
Linear SVC	0.708	0.705	0.711	0.708
Decision Tree	0.463	0.463	0.538	0.538
Random Forest (initial)	0.844	0.845	0.845	0.847
Random Forest (with feature selection)	0.854	0.857	0.855	0.858
Random Forest (optimized)	0.926	0.906	0.926	0.908
AdaBoost	0.161	0.161	0.267	0.267

The Random Forest model consistently outperformed other algorithms across all evaluation metrics. The performance improved significantly after feature selection and hyperparameter optimization, achieving a final test F1-score of 0.906 and accuracy of 0.908.

Feature Importance Analysis

Feature importance analysis from the Random Forest model revealed that certain medical indicators were particularly influential in predicting diabetes types:

1. Blood Glucose Levels
2. Insulin Levels
3. Age
4. BMI
5. Genetic Markers

These findings align with medical knowledge about diabetes risk factors. The feature selection process, which retained only the top 20 features, not only improved model performance but also enhanced interpretability by focusing on the most relevant predictors.

Model Optimization Results

The grid search for Random Forest hyperparameters explored various combinations of tree depth, number of estimators, and splitting criteria. The optimal configuration achieved a cross-validated F1-score of 0.904, representing a substantial improvement over the initial model.

The final optimized Random Forest model showed excellent generalization, with only a small gap between training (F1: 0.926) and testing (F1: 0.906) performance, indicating minimal overfitting despite the model's complexity.

Conclusion

This project successfully developed a multi-class classification model for predicting 13 different types of diabetes based on medical data. The Random Forest algorithm, after feature selection and hyperparameter optimization, demonstrated superior performance with an F1-score of 0.906 and accuracy of 0.908 on the test set.

The performance gap between simpler models (Logistic Regression, SVM) and ensemble methods highlights the complex, non-linear relationships in the data. The feature importance analysis provides valuable insights into the key factors associated with different diabetes types, which could inform clinical decision-making.

Future work could explore more advanced ensemble techniques, deep learning approaches, or the incorporation of additional medical data to further improve classification performance.