Mini Project：Diabetes Classification

41047027S 簡暐軒

Dataset：Pima Indians Diabetes Database

Number of Instances: 768

Number of Attributes: 8 plus class (Not contain p_id)

Attribute description：

Number of Times Pregnant: Describes the subject's number of pregnancies.

Plasma Glucose Concentration: This is the plasma glucose concentration measured after 2 hours in an oral glucose tolerance test. High blood sugar concentrations may be an indicator of diabetes.

Diastolic Blood Pressure: A diastolic measurement of blood pressure in millimeters of mercury (mm Hg). Blood pressure that is too high or too low can be associated with health problems.

Triceps Skin Fold Thickness: This is a measurement used to estimate the proportion of body fat in millimeters (mm). This can serve as an indicator of body fat and weight-related issues.

2-Hour Serum Insulin: Serum insulin level measured 2 hours after an oral glucose tolerance test, measured in mu U/ml. Reflects insulin resistance and pancreatic beta cell function.

Body Mass Index (BMI, Body Mass Index): Calculated as weight (kilograms) divided by height (meters) squared. Used to assess whether your weight is within a healthy range.

Diabetes Pedigree Function: This is a function based on family history that reflects the probability of developing diabetes.

Age: The age of the subject in years.

Class Variable: Target variable, indicating whether the subject has been diagnosed with diabetes (1 means diagnosed with diabetes, 0 means no).

Feature analysis and pre-processing:

We found no information missing, so we checked that all information was valid. And we found that there are 0 values in some features (such as blood glucose concentration, blood pressure, skin fold thickness, serum insulin, and BMI), which are medically unreasonable, that is, invalid. Treat these 0 values as invalid data and replace them with the **mean** of the corresponding feature.

Feature selection:

To understand which features are more important for the prediction results of the Random Forest model, we can view the **feature importance** of the model. Feature importance is a metric used to evaluate the relative importance of each feature in making predictions in a model's decision tree. These values are usually normalized to sum to 1.

So we selected the four most important features: blood glucose concentration, BMI, family history of diabetes, and age.

Reasons for choosing these features:
Blood glucose concentration: directly reflects the glucose level in the body and is a key indicator for diagnosing diabetes.

BMI: The ratio of weight to height, which can indirectly reflect the degree of obesity. Obesity is an important risk factor for diabetes.

Family history of diabetes: Provides information about genetic risk, which plays an important role in the development of diabetes.

Age: The risk of developing diabetes increases with age.

Model selection：

When dealing with the Pima Indians Diabetes dataset, we chose the **Random Forest Classifier** as our model.

This choice is based on the following reasons:
Powerful performance and accuracy: Random forest is a powerful machine learning algorithm that improves the accuracy and stability of predictions by combining multiple decision trees. It works across a variety of data types and complexities, and is particularly suitable for medium-sized data sets like this case.

Handling interactions between features: Random forests can handle interactions and non-linear relationships between features well, which are very common in medical data.

Avoid overfitting: Compared with a single decision tree, random forest reduces the risk of overfitting through integrated learning, allowing the model to have better generalization ability to new data.

Evaluation indicator and observations：

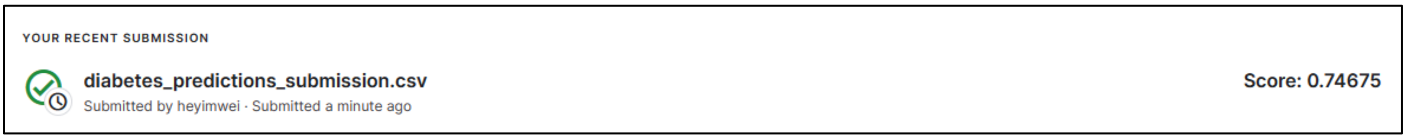
Accuracy: In our case, the model achieved approximately 76.42% accuracy on the validation dataset. This means that the model is able to correctly predict most instances, but still has a certain error rate.

Precision: For class 0 (non-diabetic), the precision is 78%; for class 1 (diabetic), the precision is 73%. This shows that the model is more accurate in predicting non-diabetic instances.

Recall: For category 0, the recall rate is 87%; for category 1, the recall rate is 59%. This means that the model performs better at identifying non-diabetic instances, but performs worse at identifying instances that are actually diabetic.

Observe:
The model performed better at predicting non-diabetic instances, with high precision and recall. For diabetes instances, the model's recall is lower, which may mean that it misses some instances that are actually diabetic, which may be an important consideration in medical diagnosis.

Screenshot of my test scores：



YOUR RECENT SUBMISSION

diabetes_predictions_submission.csv

Submitted by heyimwei · Submitted a minute ago

Score: 0.74675