The report of the project "Predicting diabetes using machine learning"

Introduction:

Diabetes mellitus is one of the most common diseases in the world, and its early detection and treatment can significantly improve the quality of life of patients. In this project, we will use machine learning techniques to develop a model that can predict the likelihood of developing diabetes based on patient clinical data.

Setting the task:

The goal of our project is to create a model that can predict the likelihood of developing diabetes in patients based on their clinical characteristics. We strive for high accuracy, completeness and reliability of the model to help doctors and patients identify the risk of diabetes at an early stage and take appropriate measures.

Methodology:

1. **Data preprocessing**: We uploaded a collection of diabetes data and started the study. Then we conducted a data study, including an analysis of the distribution of variables and correlation analysis. We have also standardized the functions to bring them to the same scale.

2. **Model selection**: We have reviewed several machine learning models such as logistic regression, k-nearest neighbors, reference vector machines, and decision trees. For each model, we adjusted the parameters using the cross-validation method.

3. **Training and evaluation**: We divide the data into training and test sets, train each model in the training set and evaluate their performance in the test set. To assess the quality of the models, we used indicators of accuracy, completeness and reliability.

Results and analysis:

For each model, we have obtained the following results:

- **Logistic regression**: accuracy-0.77, recall-0.61, accuracy-0.72.
- **k-nearest neighbors**: accuracy-0.74, recall-0.49, accuracy-0.66.
- **Vector machine support**: accuracy-0.77, recall-0.58, accuracy-0.75.
- **Decision trees**: accuracy-0.70, recall-0.60, accuracy-0.63.

From these results, it can be seen that the logistic regression model has achieved the best performance in terms of accuracy, recall and reliability.

Conclusions and further work:

In this project, we have successfully developed a machine learning model for predicting the likelihood of developing diabetes based on clinical data from patients. However, more research needs to be done to improve the performance of the model, including experiments with other machine learning algorithms, as well as collect more data to improve forecasting.