|  |  |  |
| --- | --- | --- |
| **PROJECT OVERVIEW** | **Project Name:**  Stroke interpretation | **Project Manager:** |

**STATEMENT (POS)**

**Problem/Opportunity:**

According to the World Health Organization (who), stroke is the second leading cause of death in the world, accounting for about 11% of all deaths. Stroke is a sudden cerebrovascular event. If we can predict in advance before the incident, we can take effective preventive measures against high-risk individuals, which is of great significance for reducing the incidence rate of stroke. Prediction model is the main means to predict the risk of stroke. At present, most of the existing stroke early warning tools at home and abroad are mathematical models and simple early warning tools based on traditional risk factors. Because of its complicated calculation, its clinical application is limited. According to the stroke patient information data set provided by medical institutions, this study constructs a stroke prediction model by analyzing the characteristics of the data, accurately predicts the onset time of patients, and helps medical personnel to take protective measures in time.

**Goal:**

* Specific Goal:

In this study, we plan to use random forest method to build a model to predict whether there is stroke or not, and use relevant medical knowledge and feature engineering to select appropriate variables for model training, so as to get a reliable and accurate model. The results of this project are used for stroke prediction. The data of a person's physical condition such as age, variable diseases, and smoking status are input into the model, and the model can accurately predict whether the person has a stroke. Stroke prediction can be used in the early warning of personal physical condition and the evaluation of customers' physical condition by entertainment projects, which has high application value.

The data set used in this study is downloaded from the data platform kaggle, and the programming language is Python 3. Tools mainly include jupyter, notebook, mosaic, etc. Python is a high-level programming language which is easy to learn, portable, readable, extensible and embeddable. Python is also one of floss, with rich functions and standard syntax. It supports both process oriented functional programming and object-oriented Abstract Programming. Jupyter notebook, the tool used in this study, is a convenient, flexible and interactive development tool, while mosaic is a very powerful image processing package.

* Measurement：

Accuracy, recall, ROC curve and confusion matrix were used to analyze and evaluate the performance of training model and cross validation. The accuracy of the model is higher than 90%, and the recall rate is good.

* Task arrangement：

1.  data cleaning

2.  Analyze numerical features and categorical features

3.  Construction of stroke prediction model

4.  Training and parameter adjustment using cross validation model

5.  Show the prediction effect of evaluation model

* Feasibility：

Random forest is a supervised learning algorithm, which is an ensemble learning algorithm based on decision tree. Random forest is easy to implement and has low computational cost, but it shows amazing performance in classification and regression. Because the random forest adopts the integration algorithm, its accuracy is better than most single algorithms, so the accuracy is high. Due to the introduction of two randomness, the random forest is not easy to fall into over fitting, and has a certain anti noise ability, which has certain advantages compared with other algorithms. Because of the combination of trees, random forest can deal with nonlinear data and has strong adaptability to data sets. It can deal with both discrete trees and continuous data. In the training process, the interaction between features can be detected, and the importance of features can be obtained, which has a certain reference significance.

The ROC curve is simple and intuitive. The accuracy of the analysis method can be observed through the curve, and can be judged by naked eye. ROC curve can accurately reflect the relationship between the specificity and sensitivity of an analytical method, and is a comprehensive representative of the accuracy of the test. ROC curve does not fix the classification threshold and allows the existence of intermediate state, which is beneficial for users to combine professional knowledge, weigh the impact of missed diagnosis and misdiagnosis, and select a better cut-off point as the diagnostic reference value.

So it is feasible to use the ROC curve to evaluate the model.

* Timeline.

February 11 - February 25:

Consult the relevant information, preprocess the data set, learn the relevant knowledge of prediction model construction, and complete the literature review part.

From February 25 to March 11:

The numerical features and categorical features of the data set are analyzed in a visual way to provide direction for the next construction of stroke prediction model.

March 11/25:

Build the model, train, analyze and evaluate the model, get the preliminary experimental results, and complete the first draft of the paper.

From March 25 to April 15:

Perfect the experiment, modify the paper according to the tutor's advice, and complete the final paper.

April 15/29:

Finish the PPT and prepare for the defense.

**Objectives:**

* February 25th

The data cleaning part processes the null value and missing value in the data, and other processing noise. Ensure that there is no redundant and missing data after data cleaning to ensure data quality.

* March 11th

Analyze numerical features and categorical features and analyze the impact of these features on stroke incidence by visualization, so as to find the variables suitable for stroke prediction

* March 25th

Using cross validation model training and parameter adjustment, the distributed random forest model is defined and trained, and cross validation method is used to measure the model. The grid search method is used to find the parameter setting with the minimum loss of verification set

* April 1th

The accuracy, recall, ROC curve and confusion matrix are used to evaluate the model. The prediction effect of the model is visualized, and the performance of the model is intuitively displayed, which is convenient for analysis

**Success Criteria:**

This paper uses the random forest method and other stroke prediction methods or machine learning methods to carry out stroke prediction comparative experiments, and analyzes the accuracy, recall rate, ROC curve and confusion matrix of the experimental results obtained by different methods. This method obtains the highest classification effect and the best classification effect, which indicates that this method is suitable for stroke prediction. Input a person's physical condition data to the training knife model, and the model can accurately predict whether he has a stroke, which shows that the method in this paper is successful.

**Assumptions, Risks, Obstacles:**

* Experimental analysis: first, due to the lack of relevant professional knowledge, in the feature engineering stage, the selected features may not be the best; second, I am not a professional medical staff, so the analysis of the experimental results may be one-sided.
* Experimental data: the data of this study is from the data integration platform, only for learning research, does not represent the actual situation. The number of data sets used in this experiment is not rich enough, the data coverage is not wide enough, can not cover all ages, occupations, living habits of the population, data reliability is not strong enough, there may be errors and missing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Prepared By** | **Date** | **Approved By** | **Date** |
|  |  |  |  |