STAT6128：Data-Driven Disease Prognosis Analysis

Yu-hang Huang

2024 年 1 月 12 日

目录

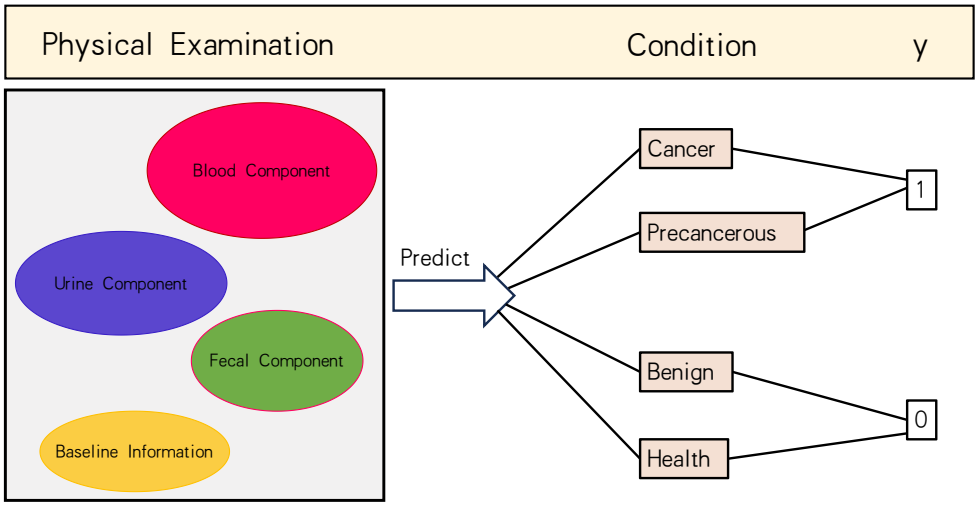

图 1: Basic Structure of the dataset.

1 Pre-sayings

In this course, I think I have gained more proficiency in the use of R language. During the course, I not only learned some basic R programming techniques, some machine learning methods, but also gasping how to create a R package and use several ways to accelarate the speed of the codes. As a student in math institute, this class just opened a new world to me, and I really appreciate the teaching style of Professor Wang Cheng and the work by teaching assistant. And in the report below, I try to do data-analysis work on a real-world dataset based on R.Thanks!

2 Introduction

Disease Prognosis is important in today's medical system. When individuals feel discomfort or illness, seeking medical attention is a common recourse. However, diagnosis based solely on verbal communication during consultations does not always provide a comprehensive understanding of the patient's condition. In such cases, additional diagnostic measures, particularly physical examinations, seems natural and common. This is especially crucial in the context of serious illnesses like cancer, where early detection is often challenging.

This paper focuses on leveraging machine learning techniques to assist in the diagnosis of patients based on the data obtained from physical examinations. The aim is to train a model using one real data-set comprising various physiological indicators derived from these examinations. The significance of this research lies in the potential to enhance diagnostic accuracy and the help to find the most significant medical examinations for patients.

3 About dateset

This dataset includes 26615 rows of data of the physical examination, the original source are attached after data anonymization. Also there are over seventy variables which are likely to explain the possibility of diseases.

3.1 Response variable

The dependent variable within the dataset pertains to the predictive condition of the disease, categorized into four dummy variables: "cancer," "precancerous lesion" (indicating benign conditions at risk of developing cancer), "benign disease" and "healthy or non-colonic disease." During the preprocessing phase, these variables were consolidated into a single dependent variable column. Specifically, "cancer" and "precancerous lesion" were coded as 1, while "benign disease" and "healthy or non-colonic disease" were coded as 0

3.2 Predictor variable

The predictor variable contains a variety of physiological indicators from physical examinations, mainly divided into blood components, urine components, fecal components, and personal baseline information. Specifically, blood components include counts of various blood cells (the data of red blood cells and platelets are more specifical) and levels of proteins, enzymes, sodium, potassium, urea, etc. in the blood. Urine components cover specific gravity, PH, and glucose content. Other components in urine and feces are marked as normal or abnormal. During the pre-processing phase, these variables were consolidated into a single dependent variable column, with normal marked as 1 and abnormal as 0. Fields are left blank if the relevant data is missing. Additionally, the predictor includes the gender with males marked as 1 and females as 0, weight, height, and BMI of the examinees.

3.3 Longitudinal

Check if there are multiple examinations conducted on the same patient in this dataset. After checking the dataset,every patient only has one record, so this project doesn't account for longitudinal data. We can assume each row of data is independent.

3.4 Data Balance

The balance of dependent variables is crucial in data analysis and statistical modeling, particularly in classification and predictive modeling. Imbalanced datasets can lead to model bias towards the majority class, reducing accuracy and reliability, especially in critical fields like medical diagnostics and fraud detection. Hence we check the shape of the dataset, and output its proportion in the Figure[4], and we are happy to find that the data is sort of balanced.

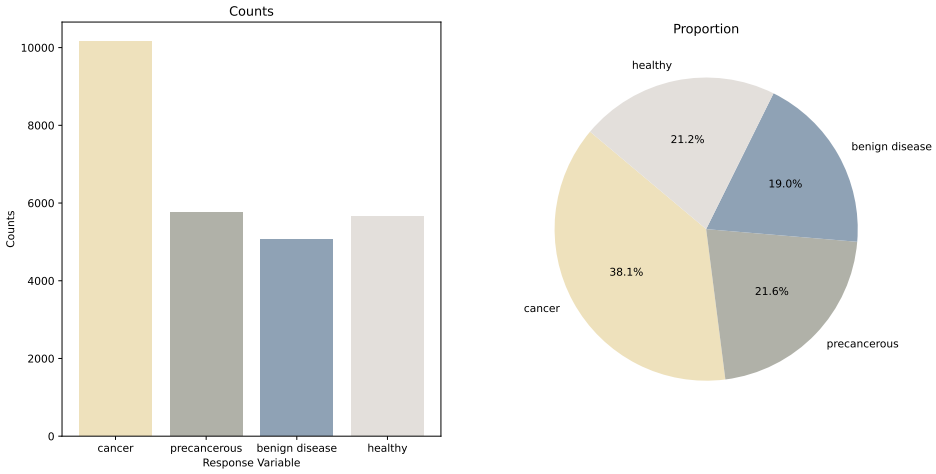

图 2: We count all positive samples(Left) and see the proportion of each dummy variable(Right).

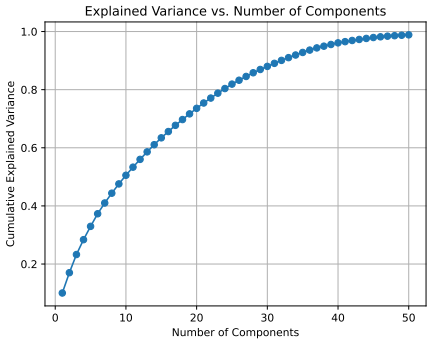

图 3: We apply PCA to the data and find it needs over 40 PCs to reach 95% variance.

4 Data Processing

4.1 Variance Analysis

As there are lots of columns in this dataset, we hence apply Principle Component Analysis method hoping to find the more significant variables to explain the phenomenon. However, we draw the accumulated explained-variance plot and find that it needs over 40 PCs to reach 96% variance. Which indicates that there might be much noise and unimportant features in the data. Therefore, we might need to do some variable-selection then. For example, we can choose features by the importance score in RandomForest, which will be discussed further in[5].

4.2 Missing Value Imputation

Our goal is to train a prediction model by the data provided, however, the patient sometimes does not participate in all means of physical examination or some data are accidentally

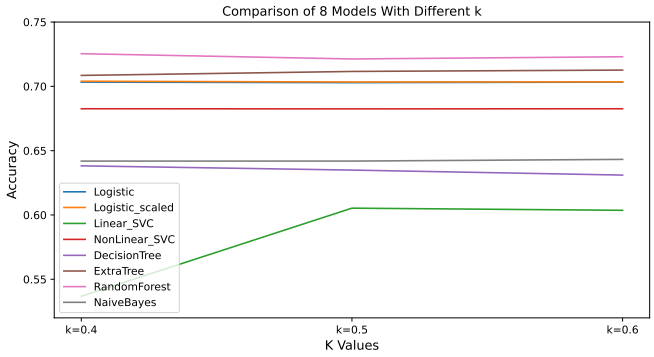

图 4: We adjust benchmark(used in KNN) and apply to 8 classification models

gone. Moreover, after checking the dataset, we find that there are over a half of the rows which have some data missing. In this way, we have to pay significant attention to the filling activities.

In this passage, we decide to use KNN(K Nearest Neighbors) method to fill the missing blanks. There are numerical variables and categorical variables: For numerical one, we use the average of its neighbours; For categorical variables, we filled it with the mean values first, then set a benchmark to decide the value to be one or zero. In this section, 7 different machine learning classifiers were tested: Logistic Regression, a Linear Support Vector Classifier, the standard nonlinear Support Vector Classifier, a Decision Tree, an Extra-Tree, a Random-Forest and a Naive-Bayes Classifier. These models were trained with different benchmarks(k) in KNN completion[**??**].We trained algorithms independently, and chose the one showing the best accuracy. We found that Random Forest provided the most accurate and significant performance.

4.3 Parameter Optimization

RandomForest is decision tree models based on the bagging framework, so the optimization of RandomForest parameters includes two parts: (1) Optimization of RF framework parameters; (2) Optimization of RF decision tree parameters. Therefore, we mainly try to find the best parameters of the RF model from two sides.

In RF bagging framework, there are not many parameters. The one worth optimizing is the parameter"n-estimator", which refers to the number of trees.

Another part about the parameters of the decision trees. Here as we only pursue the accuracy, we do not limit the depth of the tree or the number of leaf nodes. We think the parameters that count should be the features each split considers. So we compared several conditions with changing this two parameters, and draw the result as figure[5].

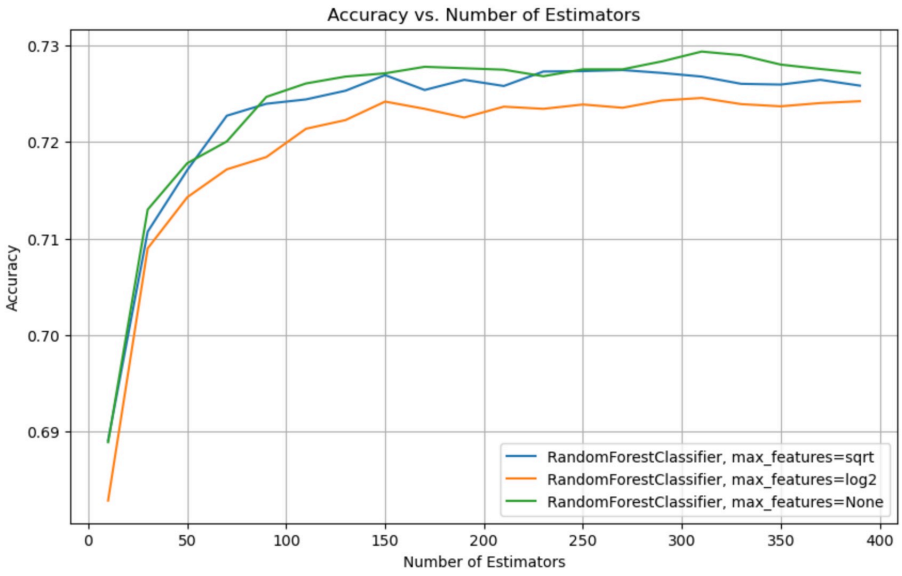

图 5: Select parameters in RandomForest Classifier.

5 Feature Engineering

The training results show that the importance of numerical variables in the dataset dominates over categorical variables. Therefore, feature engineering mainly focuses on numerical variables. Since numerical variables are quantified results of laboratory tests, they are compared with the normal values in medicine to obtain corresponding binary variables. Training the new binary variables together with the original data does not significantly improve the accuracy. In the ranking of variable importance, numerical variables generally have higher importance than the newly generated binary variables, while the importance of the newly generated binary variables is generally higher than that of the original binary variables. This suggests that the random forest algorithm itself truncates continuous variables, so this approach does not greatly improve the results.

In order to explore the interaction between numerical variables, logarithmic transformation, exponential transformation, addition, and subtraction between pairs of standardized variables are performed. These operations result in new numerical variables, which are applied to the random forest model. However, there is no significant improvement in accuracy. By observing the ranking of variable importance, it is found that the importance of variables is averaged, indicating that this approach introduces too much noise.

Therefore, only 9 variables with importance greater than 1.8% are selected. Then, these 9 variables are combined through addition, subtraction, and multiplication, and the obtained new variables are trained together with all standardized numerical variables. It is found that the accuracy improves.

6 Future Prospect

In this study, we applied machine learning technology to analyze patients' physical examination data and predict the likelihood of cancer. This research demonstrates the immense potential of machine learning in the diagnosis and treatment of cancer. With the continuous advancement of technology, we anticipate the emergence of more efficient and accurate techniques for predicting cancer risk. Currently, our analysis primarily focuses on individual health data of patients. However, in reality, the likelihood of cancer may not only be related to personal health conditions but fundamentally influenced by genetic and environmental factors. Future research could take these factors into account, combining genetic information and environmental factors with patients' physical data to further predict cancer onset and to understand the mechanisms of cancer more comprehensively.The analysis of these data is also significant for the development of cancer treatment methods. Through biomedical means, we can explore ways to reduce the increase of certain substances in the body or to promote the secretion of specific materials, thereby alleviating the patient's condition. In the future, the combination of data analysis with fields such as biology and medicine, we hope to achieve this goal.

Overall, machine learning technology has a broad application prospect in the prediction, prevention, and treatment of cancer. With the development of machine learning technology and its integration with other disciplines, the early diagnosis and later treatment of cancer will see significant improvement, thereby enhancing the survival rate and quality of life of patients.