Feature Selection and Predictive Analytics for Kidney Disease Risk

Ziyang Hao
Department of IEDA
HKUST
zhaoad@connect.ust.hk

Man Li
Department of IEDA
HKUST
mlicn@connect.ust.hk

Zihao Wang
Department of IEDA
HKUST
zihao.wang@connect.ust.hk

Abstract

High dimension and small sample size data is routinely acquired in healthcare. It is crucial to analyze medical data at an early stage. In this project, we use machine learning techniques to do feature selection and predictive analytics for kidney disease data. We build L_1 Logistic Regression model, Decision Tree and Random Forest classifiers for future medical prediction. The result shows that there exists a few dominating diagnosis indicators among high dimension features. Based on our results, we are able to design new indicators for early medical diagnosis.

1 Introduction

In healthcare, machine learning techniques have a lot of positive and life-saving outcomes, ranging from cost reduction to disease diagnosis[1, 2, 3]. However, most machine learning models are built based on big data. It is still challenging to analyse medical data with high dimensionality and small sample size, especially in the early stage of clinical research or drug discovery. For example, in the context of disease diagnosis and prognosis, patient's serum is quite useful but consists of hundreds of components, which are difficult to analyse for doctors. When this medical data stream is collected, the objective is to leverage this high dimensional data towards making diagnostic and prognostic decisions. From the view of machine learning, this can be treated as a classification problem for high dimension and small size data.

High dimensionality and small sample size pose a challenge to classification techniques[4], since they both decrease the accuracy of classifiers and increase the risk of over-fitting. Moreover, high dimensionality make the model training process time-consuming beyond reasonable computational limits, as classifiers usually do not scale or converge well for huge numbers of features. To deal with these medical problems, feature selection is the key to reducing data dimensionality. The main objective of feature selection is to obtain a reliable and robust list of predictive variables. Meantime, the selected feature should have good performance for the future prediction. A lot of studies in the literature show that in small-sample or high dimension settings, there is no unified feature selection method[5].

In this work, we tries to help find the feature importance in the context of kidney disease diagnosis. The diagnosis classifies the patients in three categories: control, low-risk and high-risk. The input of our algorithm is serum components from patients. Then we use Logistic-Regression, Decision Tree and Random Forest to predict the health condition. Our main contribution is to identify the dominated feature for kidney disease diagnosis. Based on our model, we are able to design new indicators for evaluating early kidney disease by collaboration with medical researchers.

2 Dataset and Features

The data for this report comes from a medical test which examines the correlation between the level of nephritis and a number of clinical measures. There are 68 samples in the dataset, 18 of them in control group, 28 of them in low risk group, 20 of them in high-risk group and 2 of them have no labels. There are also 75 features but 4 of them are all zeros. So after a necessary cleaning, we have 66 observations with 71features. Note that we do not separate data into training set and test set since samples are too few when compared with the number of features.

We try to discuss four classification cases: Control-high, control-low, low-high and Control-high-low. We sort the predictors according to their correlations to response variable and then create the heatmap of correlation matrix. Figure(1) shows strong correlations between predictors. Thus, the data suffers severe multi-collinearity, which means two or more explanatory variables are highly correlated.

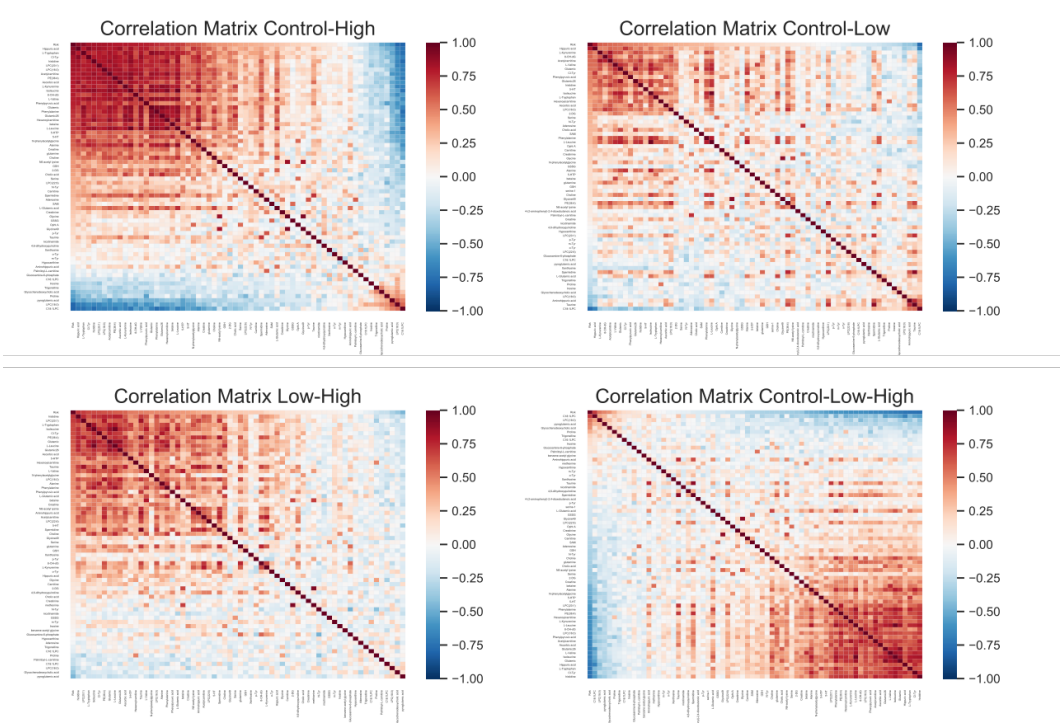


Figure 1: Correlation Matrix

Firstly, we try to find some important features. For every feature, we do one-way ANOVA test for Control-high-low case and some significantly important features are show in Figure(2).

Besides, we do K-S test pairwise to test whether the distributions in different groups are the same. Consequently, 43 features have significantly different distributions and here are scatter plots of some examples.

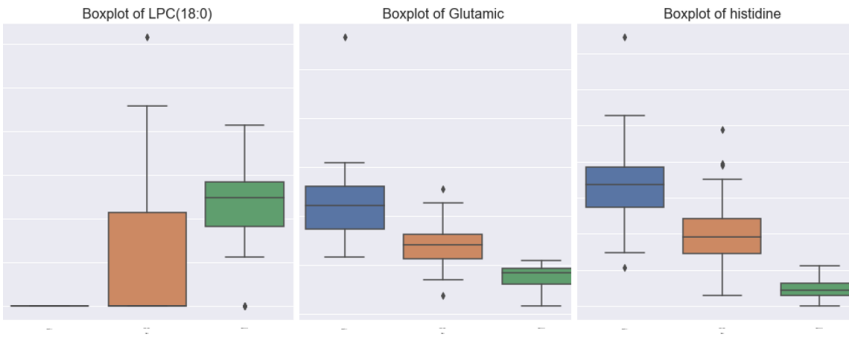


Figure 2: Box Plot of Some Features

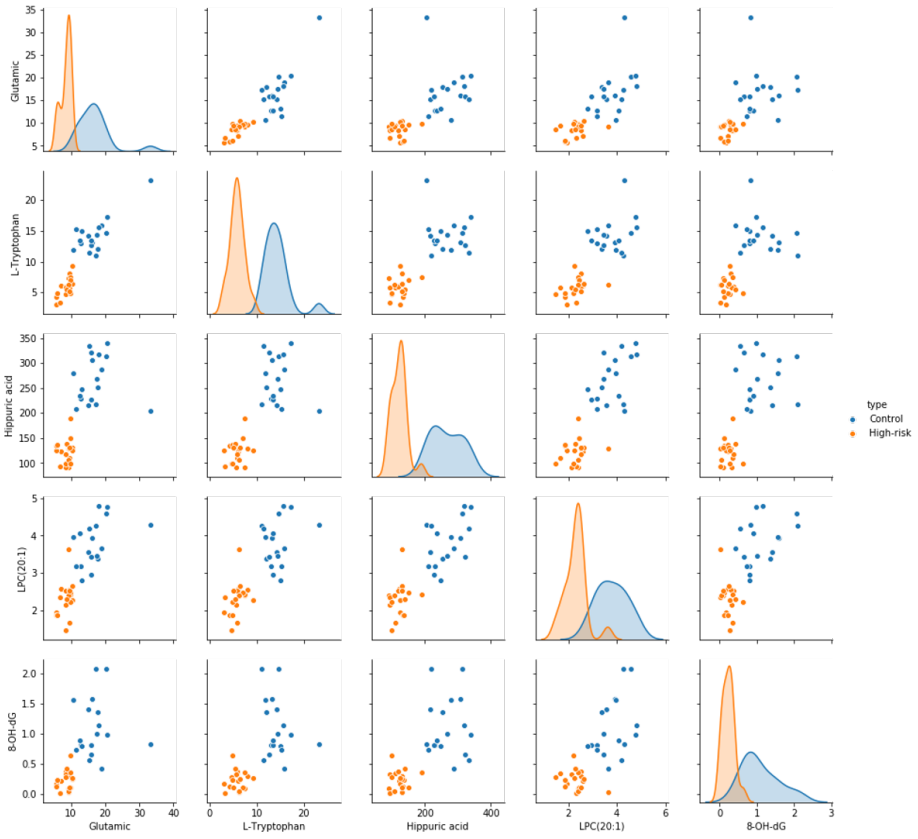


Figure 3: Scatter Plot of Some Features

3 Methods

3.1 L_1 Regularized Logistic Regression

As pointed out above, the data suffers severe multi-collinearity, the logistic regression problem is singular. Besides, it is not only important to be able to separate two data sets, but also to determine which variables are the most relevant for achieving this separation. Thus, L_1 penalty used in lasso can be used for shrinkage to make problem non-singular and also for variable selection. For the two-class case, the lasso estimate is defined by

$$\hat{\beta} = \arg\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i \left(\beta_0 + \beta^T x_i \right) - \log \left(1 + e^{\beta_0 + \beta^T x_i} \right) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Here λ is the complexity parameter that controls the amount of shrinkage[6]. We typically do not penalize the intercept term, and standardize the the predictors for the penalty to be meaningful. To display the results more visualizing, we apply coordinate descent methods to compute the coefficient profiles on a grid of values for λ. The **R** package **glmnet** can fit coefficient paths for our problems efficiently.

3.2 Decision Tree

The medical diagnosis procedure motivates us to consider decision tree model. Intuitively, the decision trees more closely mirror human decision-making. During the process of building decision tree classifiers, we perform feature selection and model complexity automatically. The appealing tree structure gives easily understandable and interpretable information regarding the predictive importance of the features. In particular, we implemented CART(Classification and Regression Trees) algorithm[7, 6]. CART constructs binary trees using the feature and threshold that yield the largest information gain at each node.

Given training predictor $x_i \in R^n, i = 1, 2, \cdots l$ and response $y \in R^l$, a decision tree recursively partitions the space such that the samples with the same labels are grouped together[6]. Let the data at node m be represented by Q. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_{\text{left}}(\theta)$ and $Q_{\text{right}}(\theta)$:

$$Q_{\text{left}}(\theta) = (x, y)|x_j <= t_m$$
$$Q_{\text{right}}(\theta) = Q \backslash Q_{\text{right}}(\theta)$$

Minimize the total impurity:

$$\theta^* = \arg\min_{\theta} \frac{n_{\text{left}}}{N_m} H\left(Q_{\text{left}}(\theta)\right) + \frac{n_{\text{right}}}{N_m} H\left(Q_{\text{right}}(\theta)\right) \tag{1}$$

where H is impurity function. Define $p_{mk} = 1/N_m \sum 1\left(y_i = k\right)$. For the classification problem, we usually have the following impurity measure:

$$H^{\text{Gini}}\left(X_m\right) = \sum_k p_{mk}\left(1 - p_{mk}\right)$$
$$H^{\text{Entropy}}\left(X_m\right) = -\sum_k p_{mk} \log\left(p_{mk}\right)$$

In general, we may obtained a quite complicated tree with too many nodes. To avoid over-fitting, we can further prune a tree by adding penalty, such as ℓ_1 penalty $\alpha|M|$ (M is total number of nodes).

3.3 Random Forest

Random forest improves the predictive performance of trees from two perspectives[8, 6]. First, it builds a number of decision trees on bootstrapped training samples and averages the tree outcomes to reduce variance. Second, at each split in the tree, the algorithm only considers a random sample of predictors to reduce the correlation among trees. Another accompanying advantage is that we can use forest to output the importance of each feature more stably. We evaluate the importance of a variable

X_j for predicting Y by adding up the weighted impurity decreases for all nodes m where X_j is used, averaged over all N_T trees in the forest

$$\text{Imp}\left(X_j\right) = \frac{1}{N_T} \sum_{T} \sum_{m \in T: v(Q_m) = X_j} p(m) \Delta i\left(Q_m, m\right)$$

and where $p(m)$ is the proportion $Q_{m\text{left}}/Q_m$ of samples reaching t_m and $v(Q_m)$ is the variables used in split Q_m. $\Delta i\left(Q_m, m\right)$ is the weighted impurity decreases which is equal to the cost function defined in (1). The random forest model shines when analysing high dimensional and small sample size data.

In the context of medical diagnosis, random forest are easier to understand. We can treat these bootstrapped trees as doctors with different abilities to analyse sample components. Because each tree can only split a random sample of features. Then we let these doctors vote for the most importance feature. This process will not only give us more accurate diagnosis result, but also select the most important medical indicators.

4 Results

4.1 L_1 Regularized Logistic Regression

4.1.1 Classification between Control and High-Risk

We choose the complexity parameter λ by 5-fold cross-validation. Figure 4 shows the misclassification errors for cross-validation on the data. We also use binomial deviance to measure errors and its curve is much smoother, as shown in Figure 5.

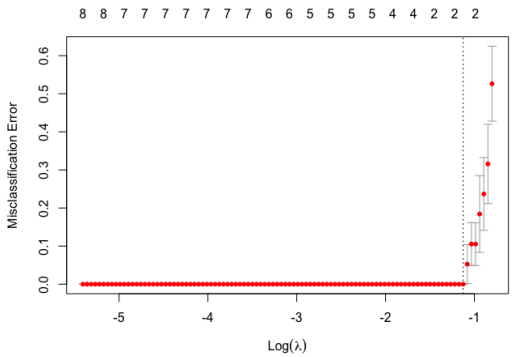


Figure 4: Misclassification Error

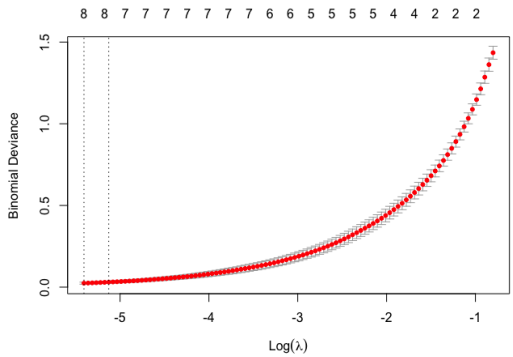


Figure 5: Deviance

Figure 4 suggests that $\ln \lambda = -1.13$ gives the minimum misclassification error and we choose 2 explanatory variables, which are L-Tryptophan and Hippuric acid, as shown in Figure 6. However, we conclude that $\ln \lambda = -5.40$ is more preferable when considering deviance. In latter situation, we select 8 explanatory variables, which are shown in Figure 7. Note that since the data are linearly separable, the solution is undefined at $\lambda = 0$, and degrades for very small value of λ. Hence the paths have been truncated as the fitted probabilities approach 0 and 1.

The reason why λ is different lies in the fact that L-Tryptophan and Hippuric acid can separate the data completely without misclassification error. So we only choose 2 variable in the former case. Note that 8 explanatory variables selected by deviance can also divide data into 2 groups without misclassification error, but the model has lower deviance.

From the point of deviance, we have select 8 explortary variables and we try to understand the relations between these 8 variables. So we apply factor analysis for these 8 variables and it suggests that the number of factors is 2. From Figure 8 and 9, if we only care about some variable whose coefficients are large, including 8-OH-dG, L-Tryptophan, Phenylpyruvic acid and C18:1LPC, then we can divide them into 2 groups. The first group includes 8-OH-dG, L-Tryptophan and Phenylpyruvic

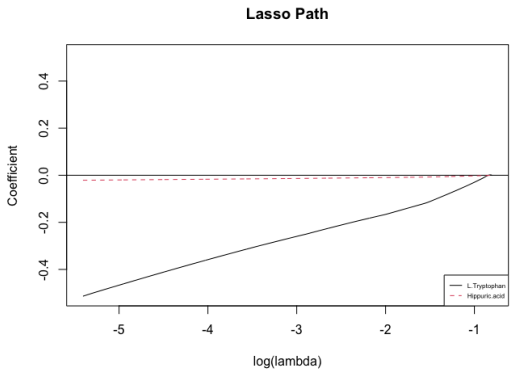


Figure 6: Misclassification Error

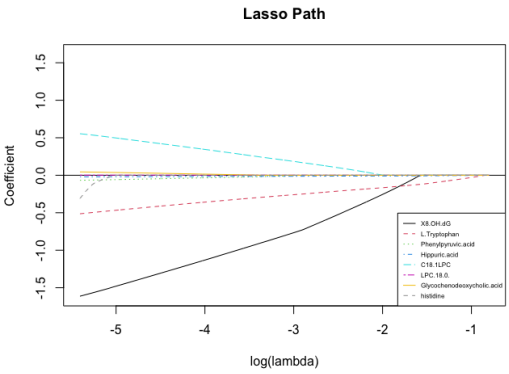


Figure 7: Deviance

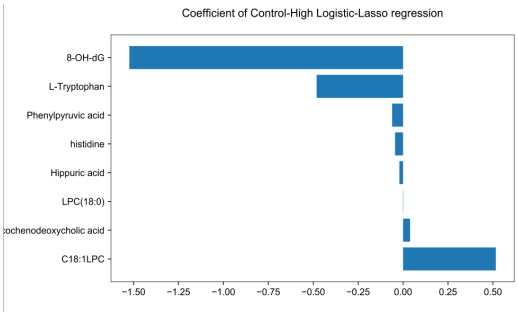


Figure 8: Coefficients of Deviance

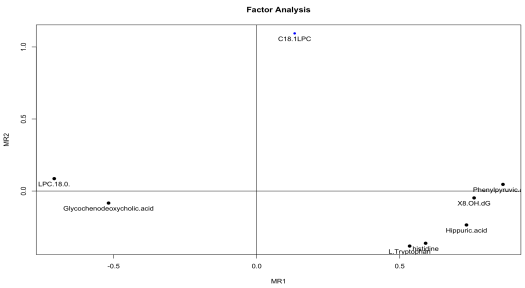


Figure 9: Factor analysis

acid, which lead people to high risk. The second group only includes C18:1LPC, which prevent people from high risk.

4.2 Classification between three groups

We choose the complexity parameter λ by 10-fold cross-validation. Figure 10 and 11 show the misclassification error and deviance respectively for 10-fold cross-validation on the data. Note that both two measures suggest that $\ln \lambda = -5.18$ is a good choice. And since the data are not linearly separable, the smallest misclassification error is not 0. Maybe we can use some nonlinear classification method.

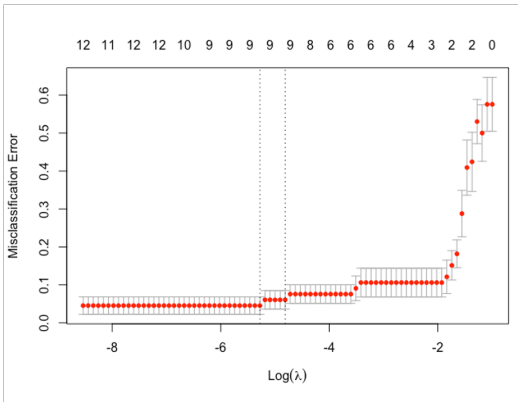


Figure 10: Misclassification Error

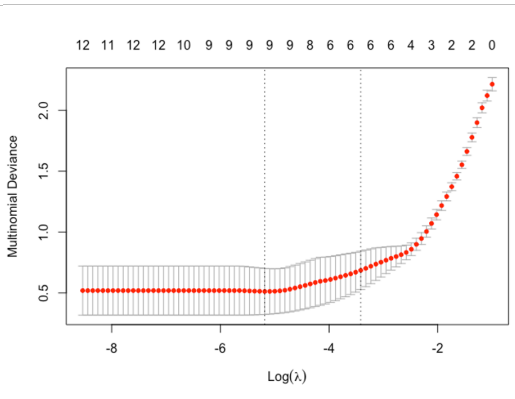


Figure 11: Deviance

The results are shown in the following. The results are not as good as outcomes in binary classification. There are 26 non-zero explanatory variables, which is a large number compared with 66 observations. So we may choose some significant variables. For example, 8-OH-dG, L-Kynurenine, Glycine, and Serine for control group; N-Tyr, Glucosamine-6-phosphate and nicotinamide for Low-risk group; LPC(20:1), PE(36:4), Cl-Tyr, L-Tryptophan, and pyroglutamic acid for High-risk group.

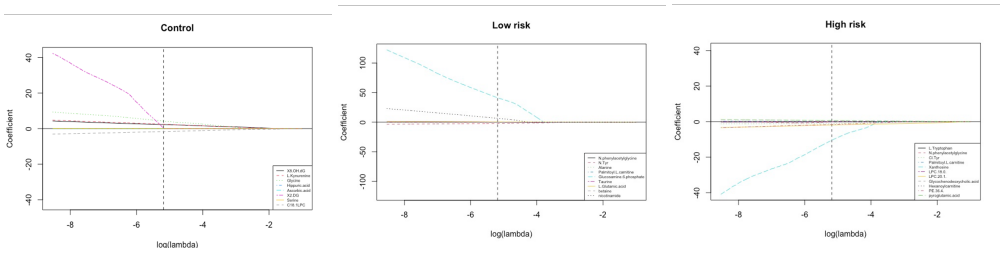


Figure 12: Coefficient Paths of Three Groups

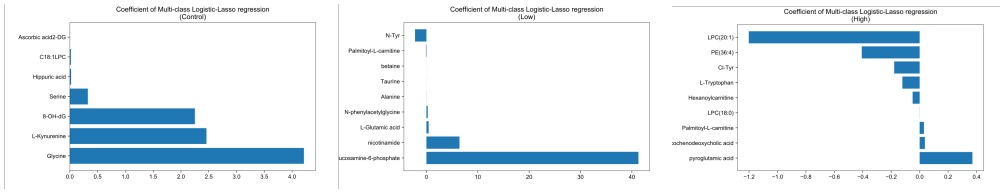


Figure 13: Coefficients of Three Groups

4.3 Decision Tree

4.3.1 Hyperparameters

- When building decision trees, we did 5-fold and 10-fold cross-validation to choose impurity function. Both results show that Gini index performs better than Entropy.
- Due to the existence of high quality features, in any classification case, we can always obtain simple and clean tree structure. There is no need to prune the tree structure.
- No maximal depth constraint.

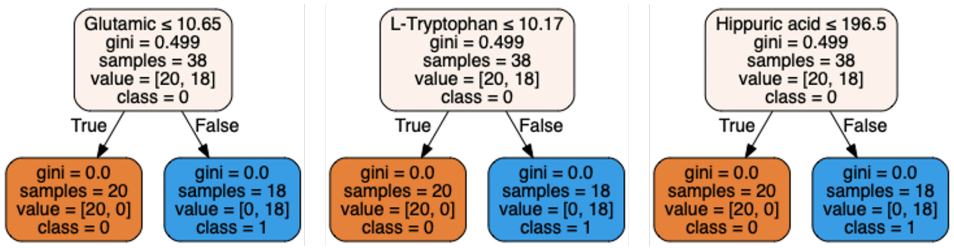
4.3.2 Visualization of tree structure

We visualize the tree structure by showing related information on each nodes in the tree. The results are shown in Fig 14. These tree visualization provides us clean and simple structure: we can perfectly classify the patients using only one or two features. Meanwhile, the disadvantage is also obvious. The tree did the work "too perfectly" on the training set. We can see that some nodes in the figure only include one sample. This will cause poor future predictive behavior.

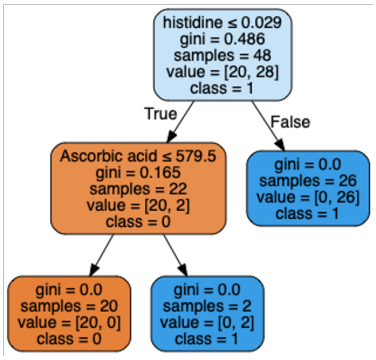
Although tree model suffers from high variance for prediction, it gives us important insight to the feature selection: a hierarchy structure of features. For example, Glutamic, L-Tryptophan and Hippuric acid are three dominating indicators in the classification of control and high-risk patient. Decision Tree also motivates us to use random forest to reduce the variance.

4.4 Random Forest

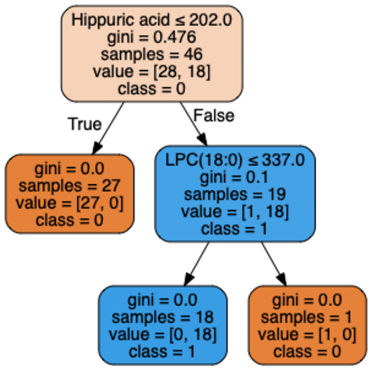
Compared with decision trees, the predictive performance of random forest is pretty good. In current dataset, we can achieve 100% accuracy for prediction. In this part, we mainly focus on feature importance which helps us do feature selection.



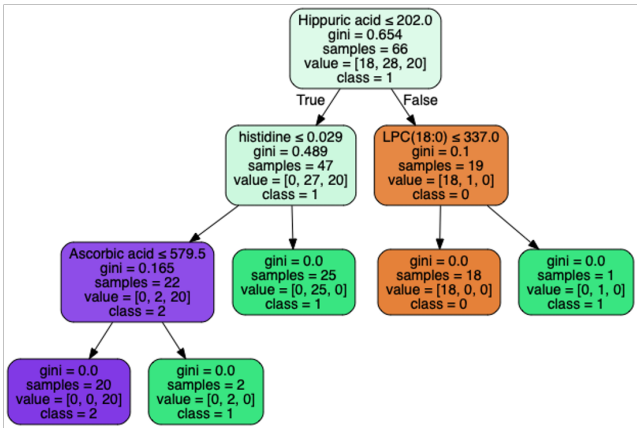
(a) All tree structures for control-high binary classification



(b) One possible structure for low-high binary classification



(c) One possible structure for control-low binary classification



(d) One possible structure for control-low-high classification

Figure 14: Decision Tree structure for classification

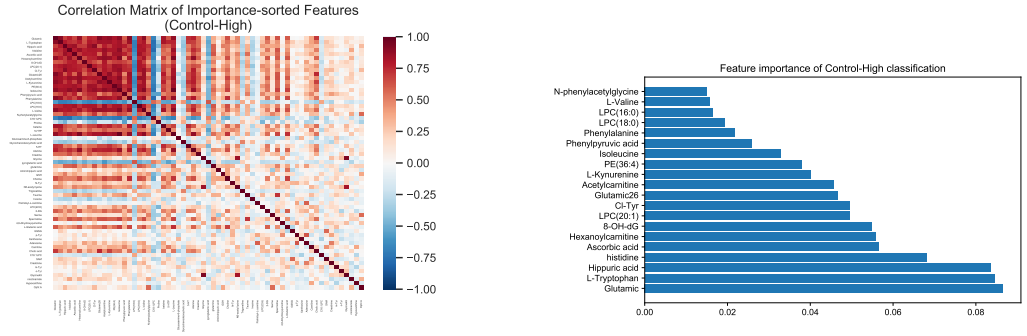
4.4.1 Hyperparameters

- We choose predictor subset size $m = \sqrt{p}$, where p is the total number of predictors. Here 8 out of 71 for our data.
- We build 10^4 trees in the forest. Because the data size is very small (less than 100), we can train such a huge size in the laptop easily.
- No maximal depth constraint.

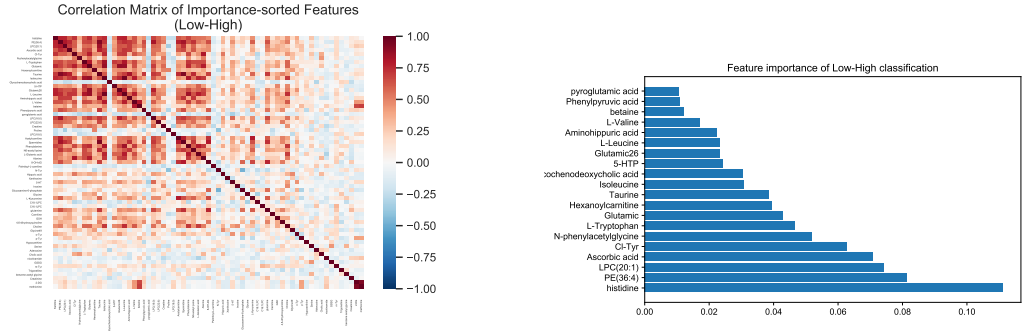
4.4.2 Feature importance

We visualize feature importance by creating horizontal bar graph. In addition, we visualize the correlation matrix as a heatmap in the order of importance. Compared with the correlation heatmap in EDA part, some features with high correlation to response have been shown less important. We conclude that such features can be discarded. In other words, when building tree-based classifier,

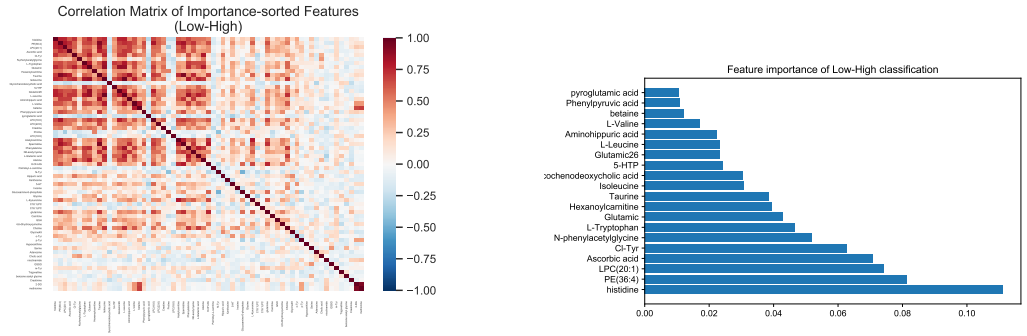
these features will not be considered due to their weak importance. This can be regarded as feature selection based on importance.



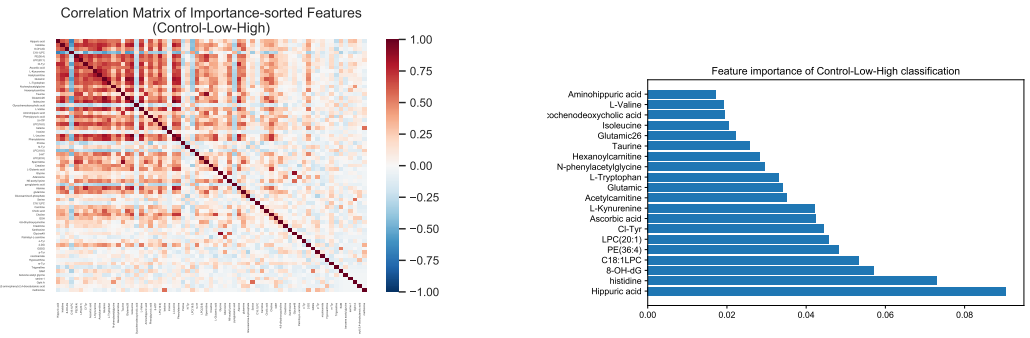
(a) control-high binary classification



(b) low-high binary classification



(c) control-low binary classification



(d) control-low-high multiclass classification

Figure 15: Random Forest feature importance information. Left column is heatmap of correlation matrix. Right column is horizontal bar graph with 20 most-importance features.

In the horizontal bar graph, we show 20 most important features. It is interesting to note that in control-high binary classification, the selected feature importance shows an appealing hierarchy structure. We can define the importance level of indicators. For example, Glutamic, L-Tryptophan and Hippuric acid belong to Level-I, Histidine is Level-II, 8OH-dG, Hexanoylcarntine and Ascorbic acid belong to Level-III, etc. In practice, we need to further collaborate with medical researchers to

5 Conclusion and Future Work

In regression, L_1 regularized logistic regression performs quite well in binary classification. By combining it with factor analysis, we can even find some relations hidden in the data for interpretation. However, when it comes to classification problem between three groups, it is not efficient as the former case. There are still plenty of explanatory variables, which is a large number when compared with samples. We will explore some methods such as splines to find some inherently nonlinear relations as future work.

In tree-based method, we are able to identify the feature importance for all the medical indicators. Random Forest does feature selection pretty good and provides us insights to design new medical diagnosis indicators for application.

We expect our results can provide insights and guidance for kidney disease research.

Acknowledgments

Upon the completion of this report, we are grateful to Prof. Jing, who have offered us lots of advice and support during the course of study. And special acknowledgment is given to Prof. Xie, who provided us with medical data and kind suggestions.

References

[1] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, 5:8869–8879, 2017.

[2] Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE journal of biomedical and health informatics*, 19(6):1893–1905, 2014.

[3] K Srinivas, B Kavihta Rani, and A Govrdhan. Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2(02):250–255, 2010.

[4] Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*, 2006.

[5] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[7] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

[8] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.