Journal of Data Science 14(2016), 97-116

# Prediction of Hypothyroidism Disease by Data Mining Technique

Kim-Fa KHIEW[1], Tsuey-Lan WANG[1], Mark Y.S. LIN1, Yefei JIANG[1,*]

[1] *Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan*

*Abstract:* **Background:** Hypothyroidism is one of the endocrine diseases found in human being, it is not a immediate fatal disease, but progress in chronic status that lead to other diseases. Several machine learning techniques were applied to hypothyroidism for the prediction of hypothyroid medical diseases.
**Methods:** The dataset used for UCI repository which has 3163 observations with 151 belongs to category hypothyroid disease. Considered with two missing data method, four imbalance data method and two kinds of classification models by using serval evaluation indexes to find a better model.
**Results:** After comparison of these models, query_on_thyroxine and lithium were ignored and TBG was reconsidered in the model. A new RF imputation method was used. Finally, all variables model have 100% accuracy but cost time and money, only use TSH, FTI, TBG, TT4 and age variables model also can keep 0.9988 AUC, which will more useful in real case.
**Conclusion:** Both all variables mode and five variables model have very high accuracy than previous studies, a nonparametric ensemble model is suggest in this case. But there also have some limitations like out-of-date data, hypothyroidisms type were not considered.

## 1. Introduction

Disease prediction can be applied to different domains such as risk management, tailored health communication and decision support systems. Risk management plays an important role in health insurance agencies, mainly in the underwriting process[14]. Most crucial and challenging task in the field of medical science is to identify or diagnose disease at correct time. Disease can be cured if diagnosed correctly on time. It helps doctors in proper treatment of patient. Only doctors and physicians can identify the disease using their experience, medical reports. During earlydays diseases were determined by the symptoms exhibited by the patients and various medical tests but now doctors are taking assistance of various advance systems as well[29].

Hypothyroidism is one of the most common diseases found in human being, it is not a deadly disease, but it is chronic disease which can give rise to other diseases. Thyroid is a butterfly-shaped gland, which is located at the bottom of the throat responsible of producing two active thyroid hormones, levothyroxine (T4) and triiodothyronine (T3) that affect some functions of the body as shown on Figure1. These functions include stabilizing body temperature, blood pressure

and regulating the heart rate. Hypothyroidism and hyperthyroidism are a result of an imbalance of thyroid hormone. Hypothyroidism is simply not enough thyroid hormone. The imbalance affects the metabolism in the body. Hypothyroidism causes a reduction in stroke volume and heart rate causing lowered cardiac output with a decrease in heart sounds. The symptoms of hypothyroidism shown on Figure 2.

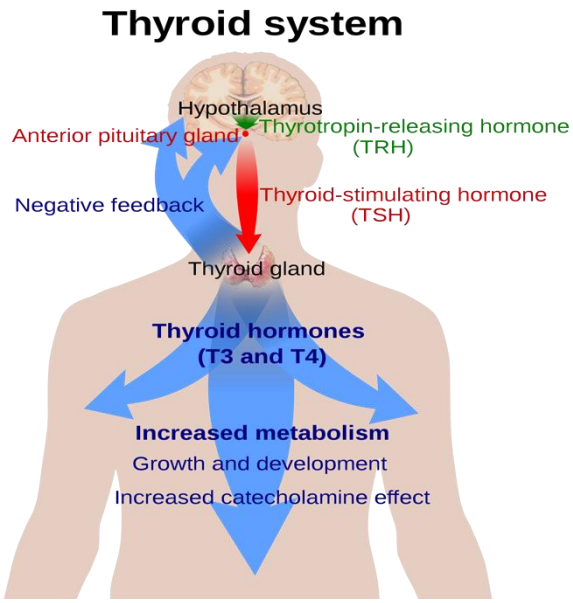


Figure 1: Thyroid System

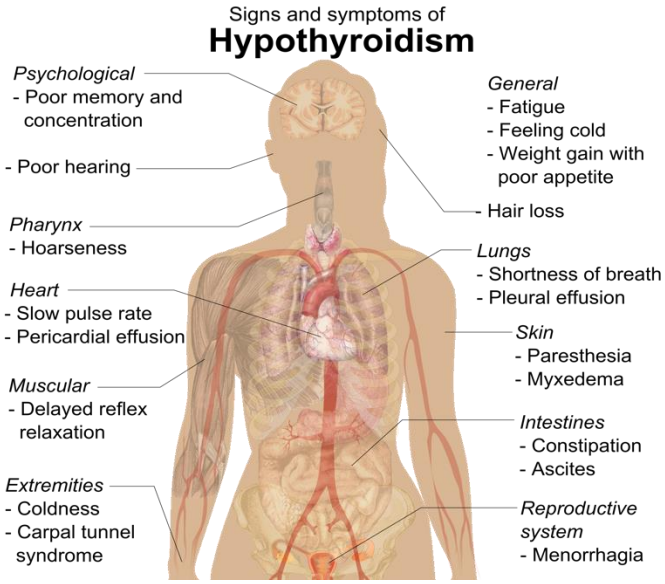


Figure 2: Symptoms of Hypothyroidism

There  Diagnosis of congenital hypothyroidism (HT) from year 1997 to 2008 in Taiwan National Health Insurance Research Database (NHIRD), it was 5.02 per/10000 births were diagnosed with development delay, congenital anomalies of the heart and mental retardation[28]. The epidemiology of congenital hypothyroidism is higher in Asia than in Europe. However treatments of hypothyroidism also increase medical cost.

Several machine learning techniques were applied to hypothyroidism datasets for the prediction of hypothyroid medical diseases, including Hashimoto's thyroiditis, hyperthyroidism and hypothyroidism. These diseases have a large range of symptoms and affect all ages. Gaikwad and Pise (2014) focused on hypothyroid medical diseases caused by under active thyroid glands[15]. Classification of this thyroid disease is a considerable task. An experimental study is carried out using rotation forest using features selection methods to achieve better accuracy. Margret et al. (2012) using decision tree attribute splitting rules helps to classify the data in dataset according to aforesaid disorders [25]. This method provides five different splitting criteria for the construction of decision tree. The splitting criteria are Information Gain, Gain Ratio, Gini Index, Likelihood Ratio Chi-Squared Statistics, Distance Measure. Among, the aforementioned splitting rules three rules belong to Impurity based splitting criteria and other two are Normalized Impurity based splitting criteria. As a result, the decision tree classifies the thyroid data-set into three classes of thyroid disorders. Akbaş et al. (2013) aims performance improvement in diagnosis of thyroid cancer with machine learning techniques [2]. They found that by using ensemble approaches performance improvement has been achieved in diagnosis of thyroid cancer. Kousarrizi et al. (2012) proceed an experimental comparative Study on Thyroid Disease Diagnosis based on feature subset selection and classification [22]. The method they proposed has two stages. In the first stage, feature selections are utilized as a pre-processing step. The main purpose of feature selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy [6]. In this study sequential forward selection (SFS), sequential backward selection (SBS) and Genetic Algorithm are used as feature selection methods. In the second stage, SVM is used to classify thyroid data. Radwan and Assiri (2013) study the Thyroid Diagnosis based technique on Rough Sets with Modified Similarity Relation [30]. According to their research, various neural network methods including Multi-Layer Perception with Back Propagation method (MLP), Radial Basis Function (RBF) and adaptive Conic Section Function Neural Network (CSFNN) are used to help diagnosis of thyroid disease, their classification accuracies are separately 88.3%, 81.69% and 85.92%.

## 2. Method

### Data Collection and Pre-processing

The study was approved by a research ethics review committee at Fu Jen Catholic University, Taiwan. The dataset used for experimental purpose is downloaded from University of California of Iravin (UCI) repository site [3]. The dataset has 3163 observations from which 3012 belongs to category negative, 151 belongs to category hypothyroid disease. The first variable is the class

of hypothyroid disease, hence there are 25 features in all, which will be used to classify the data. The detail of data set is shown in Table 1[11].

Table 1: Variable Description

| Variable Name | Value | Variable Name | Value |
|---|---|---|---|
| Hypothyroid | True, False | goitre | True(t), False(f) |
| age | Numeric | TSH_measured | True(t), False(f) |
| sex | Male(M), Female(F) | TSH | Numeric |
| on_thyroxine | True(t), False(f) | $T_3$_measured | True(t), False(f) |
| query_on_thyroxine | True(t), False(f) | $T_3$ | Numeric |
| on_antithyroid_medication | True(t), False(f) | $TT_4$_measured | True(t), False(f) |
| thyroid_surgery | True(t), False(f) | $TT_4$ | Numeric |
| query_hypothyroid | True(t), False(f) | $T_4U$_measured | True(t), False(f) |
| query_hyperthyroid | True(t), False(f) | $T_4U$ | Numeric |
| pregnant | True(t), False(f) | FTI_measured | True(t), False(f) |
| sick | True(t), False(f) | FTI | Numeric |
| tumor | True(t), False(f) | TBG_measured | True(t), False(f) |
| lithium | True(t), False(f) | TBG | Numeric |

Data explore found that the dataset should be cleaned in several aspects. Such as observations do not meet the common sense, lose lots of information and missing value hard to be imputed. Finally, 2011 observations were recorded with the proportion for 1867: 144.

### Missing Data

The Missing data is a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. Appropriately dealing with missing can be challenging as it requires a careful examination of the data to identify the type and pattern of missingness, and also a clear understanding of how the different imputation methods work. Sooner or later all researchers carrying out empirical research will have to decide how to treat missing data. In a survey, respondents may be unwilling to reveal some private information, a question may be inapplicable or the study participant simply may have forgotten to answer it[35].

There are many methods to hand the missing data, two famous methods to treat the missing data were used on this case, and compare their results performance. In this case has missing data for any of the variables, then simply exclude that case from the analysis. It is usually the default in statistical packages (Briggs et al.,2003). The positive side is that it can be used with any kind of statistical analysis and no special computational methods are required. But the limitations are that it will exclude a large fraction of the original sample [17].

MCMC method originated in physics as a tool for exploring equilibrium distributions of interacting molecules. In statistical applications, it is used to generate pseudorandom draws from

multidimensional and otherwise intractable probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends on the value of the previous one. In MCMC, one constructs a Markov chain long enough for the distribution of the elements to stabilize a common 3 distribution. This stationary distribution is the distribution of interest. By repeatedly simulating steps of the chain, it simulates draws from the distribution of interest [32].

**Imbalanced Data**

Many classification models built on the basis of equilibrium data. For imbalanced data, these models will be subject to the more severely affected, especially for a rare type of recognition capability has been greatly weakened. But most of frauds, breach of contract, customer loss and cancer, the concerned target is always the weak side.

A modified approach is to consider the misclassification costs, for example, Cost Sensitive Learning algorithms, it is through the rare class of misjudgment given a higher cost to correct the class of rare ability to identify[12, 40]. Another idea is corrected through a variety of resampling methods to improve data collection from unbalanced data into equilibrium data, and then subjected to various classification model or classification algorithms. In both ideas, the resampling method has more research priorities. For example, Random Over-Sample and Random Under-Sample are most frequently used [6]. SMOTE (Synthetic Minority Over-sampling Technique) algorithm, sampling to the rare class of pairs of adjacent samples inserted to the synthesis of new, non-repeating rare class[7].

**Lasso-Logistic Regression**

The Lasso (Least Absolute Shrinkage and Selection Operator), which was originally proposed by Tibshirani for linear regression models, has become a popular model selection and shrinkage estimation method[37]. The motivation comes from Breiman's NNG (Non-negative Garrote)[4] in 1995 as

$$\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}c_j\hat{\beta}_j x_{ij}\right)^2 \qquad \text{s.t. } c_j\geq 0,\ \sum c_j\leq t \tag{1}$$

Suggest that we have a data $(X^i,y_i)$,$i$=1,2,…$n$, where $X^i=(x_{i1},\ldots,x_{ip})$ and $y_i$ are design matrix and response variable respectively. Assume that $\frac{1}{n}\sum_i x_{ij}=0, \frac{1}{n}\sum_i x_{ij}^2=0$, Let$\hat{\beta}=(\hat{\beta}_1,\ldots,\hat{\beta}_p)'$, then the lasso estimator is then defined as

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta}\left\{\sum_{i}^{n}\left[\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right]\right\} \\ &= \arg\min_{\beta}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda\sum_{j=1}^{p}|\beta_j|\end{aligned} \tag{2}$$

where $\lambda \geq 0$ . The first part of equation(2) represent a performance of fit and the second part represent a l 1 - type penalty. For large values of the penalty parameter $\lambda$ , some components of $\beta\hat{}$ are set exactly to 0. The l 1 - type penalty of Lasso can also be applied to other model as for example Cox Regression[38], Logistic Regression[24, 31, 34, 16] or Multinomial Logistic Regression[23] by replacing the residual sum of squares by the corresponding negative log-likelihood function.

Consider the $y_i$ changes into binary variables as $y_i \in \{0,1\}$ . The conditional probability of Logistic type on linear regression model can be expressed as

$$\log\left\{\frac{p\left(y_i=1 \mid X^i\right)}{1-p\left(y_i=1 \mid X^i\right)}\right\}=\eta_\beta\left(X^i\right) \tag{3}$$

where $\eta\beta(Xi) = \beta_0 + \sum x_{ij}\beta_j$. The parameter estimates in Lasso-Logistic model $\hat{\beta}$ can be minimize the convex function by the equation (4)

$$S_\lambda(\beta)=-l(\beta)+\lambda\sum_{j=1}^{p}|\beta_j| \tag{4}$$

where $l(\cdot)$ is log-likelihood function, and the $l(\beta)$ in equation (4) is

$$l(\beta)=\sum_{i=1}^{n}\left\{y_i\eta_\beta\left(X^i\right)-\log\left\{1+\exp\left[\eta_\beta\left(X^i\right)\right]\right\}\right\} \tag{5}$$

then we can get the estimators as

$$\hat{\beta}=\arg\min_{\beta}\sum_{i=1}^{n}\left\{y_i\eta_\beta\left(X^i\right)-\log\left\{1+\exp\left[\eta_\beta\left(X^i\right)\right]\right\}\right\}+\lambda\sum_{j=1}^{p}|\beta_j| \tag{6}$$

**Random Forest**

Random forest (Breiman, 2001) is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process[5]. The main idea is supposed that we have a dataset $D$, then using bootstrap resampling method sample $k$ sample datasets $D_i$,$(i=1,2,\cdots,k)$.In every sample datasets $D_i$ , random choice $p$ variables and training a Decision Tree model $\{h_i(X_i), i=1,2,\cdots,k\}$ respectively. After we $k$ Decision Tree model built up, then combine the results $\{h_1(X_1), h_2(X_2),\cdots,h_k(X_k)\}$ and get a final ensemble model, the final classification results was using simple majority voting method as

$$H(x)=\arg\max_{Y}\sum_{i=1}^{k}I\left(h_i\left(x_i\right)=Y\right) \tag{7}$$

where $H(x)$ is ensemble classification model, $h_i$ is single Decision Tree model, $Y$ is the response variable, $I(\cdot)$ is indicator function.

One of the most important features of RF is the output of the variable importance. Variable importance measures the degree of association between a given variable and the classification

result. RF has four measures for the variable importance: raw importance score for class 0, raw importance score for class 1, decrease in accuracy and the Accuracy and Gini index. To estimate variable importance for some variable *j* , the Out-of-Bag (OOB) samples are passed down the tree and the prediction accuracy is recorded. Then the values for variable *j* are permuted in the OOB samples and the accuracy is measured again. These calculations are carried out tree by tree as the RF is constructed. The average decrease in accuracy of these permutations is then averaged over all the trees and is used to measure the importance of the variable *j* . If the prediction accuracy decreases substantially, then it suggests that the variable *j* has strong association with the response[19]. After measuring the importance of all the variables, RF will return a ranked list of the variable importance[21].

Formally, let $\beta_t$ be the OOB samples for tree *t* , $t \in \{1,\cdots,ntree\}$ ,$y_{i,a}^{\prime t}$ is the predicted class for instance *i* before the permutation in tree *t* and $y_{i,a}^{\prime t}$ is the predicted class for instance *i* after the permutation. The variable importance VI for variable *j* in tree *t* is given by

$$VI_j^t = \frac{\sum_{i=1}^{N} \beta_t I\left(y_i = y_t^{\prime t}\right)}{|\beta_t|} - \frac{\sum_{i=1}^{N} \beta_t I\left(y_i = y_{i,\alpha}^{\prime t}\right)}{|\beta_t|} \tag{8}$$

The raw importance value for variable *j* is then averaged over all trees in the RF.

$$VI_j = \frac{\sum_{t=1}^{ntree} VI_m^t}{ntree} \tag{9}$$

**Model Performance Evaluation**

Considered about the dataset which has missing data, imbalanced with response variable and lots of factor variables. Serval dealing methods was compared in order to find the best model. Three sets of analyses were conducted. First of all, for missing data, two kinds of dealing methods such as directly remove missing samples, using MCMC (Markov Chain Monte Carlo) imputation were considered. Secondly, provide three data balancing method (Random Over-Sample, Random Under-Sample and SMOTE) compare with using imbalanced data directly. Finally, two kinds of classification model Logistic for parametric and Random Forest for nonparametric were conducted, since the Logistic model can not select the variables, a l 1 - type penalty called Lasso was added into Logistic model. Moreover, lots of papers do data balance one the whole datasets, but the balanced validation/test data is unbiased of the real situation. Thus, 5-fold cross validation was conducted. In every fold, one subset treated as validation data directly, the rest four dataset as training data, just doing data balance algorithms on training data and get the new training data to build the model.

To evaluate the performance of each model, all the process was repeated 50 times. The metrics such as average Confusion Matrix, average AUC, average TPR, average FPR and average Best Cut-off for each classifiers were calculated and compared[18]. All the metrics are unctions of the confusion matrix as shown in Table 2. The rows of the matrix are actual classes, and the columns are the predicted classes. Based on Table 2, the performance metrics are defined as:

$$\text{TPR (True Positive Rate)} = \frac{TP}{TP + FN}$$
$$\text{FPR (False Positive Rate)} = \frac{FP}{FP + TN} \quad (10)$$

Table 2: Confusion Matrix

| | **Predicted Negative Class** | **Predicted Positive Class** |
|---|---|---|
| **Actual Negative Class** | TN (True Negative) | FP (False Positive) |
| **Actual Positive Class** | FN (False Negative) | TP (True Positive) |

To combine the FPR and the TPR into one single metric, we first compute the two former metrics with many different threshold (for example 0.00,0.01,0.02,⋯,1.00) for the model, then plot them on a single graph, with the FPR values on the abscissa and the TPR values on the ordinate. The resulting curve is called ROC curve, and the metric we consider is the Area Under Curve of this curve, which we call AUC. The whole research process flow chart of the hypothyroidism data was shown on Figure 3 . Analysis was conducted using R Version 3.2.2.

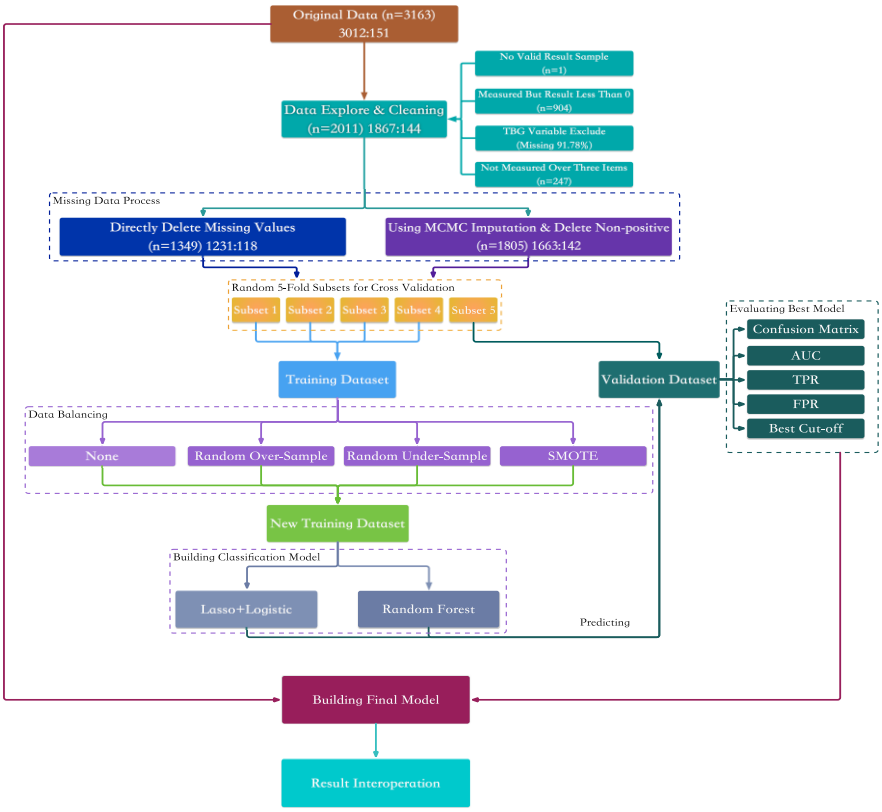


Figure 3: Research Process of Hypothyroidism Data

## 3. Results

The study collected data on 3163 observations. All the records are classified as negative and hypothyroid. In medical view, the key attributes to determinate a person who is suffering hypothyroidism is T3, TT4 and TSH. As the distribution shown on Figure 4, the distribution of T3, TT4 and TSH in different classes of hypothyroid are significantly different. But it also can be found that age, FTI and TBG have the same pattern like T3, TT4 and TSH.

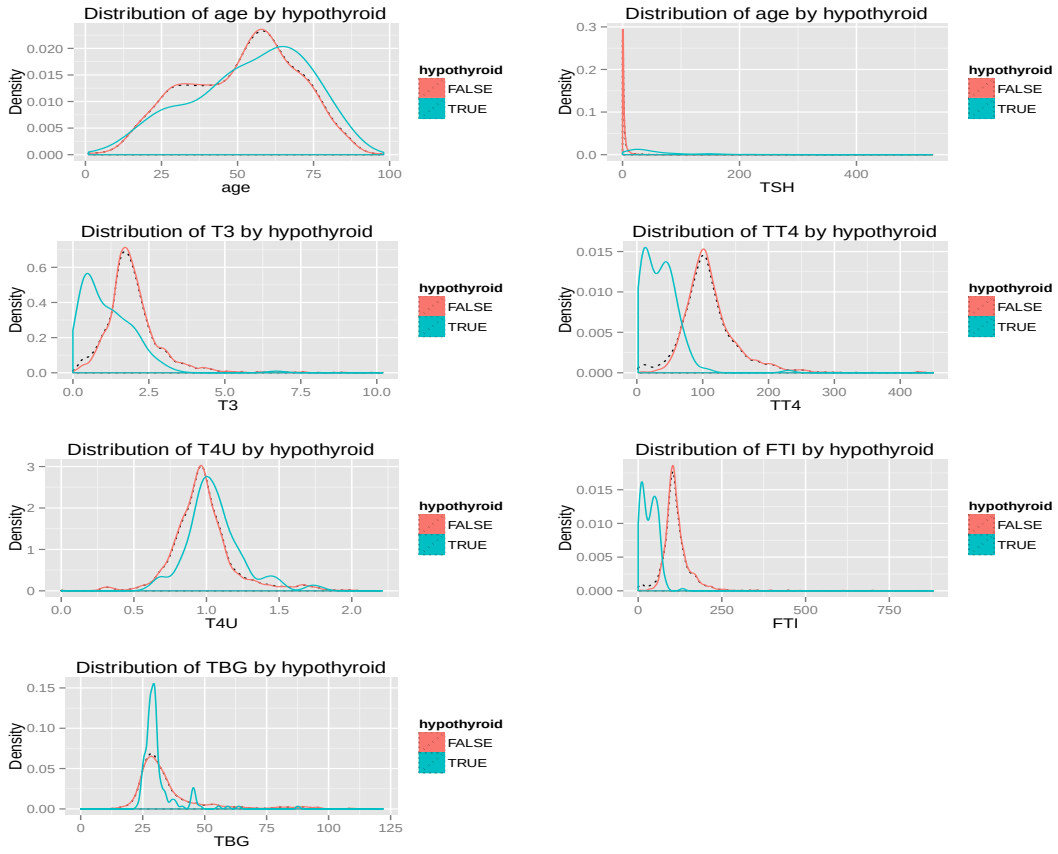


Figure 4: Distribution of Numeric Data

### Model Selection

As the consideration of missing data, imbalance data and classification methods, a comparison of model performance shown in Table 3. Specific details of the operation also on Figure 3. Every model is average of repeated 50 times in order to reduce the random errors.

### Comparison of Missing Data Methods

Imputation sometimes make more noise, but it also saves information than omit. As the result in Table 3, imputation have good performance of control the errors of negative hypothyroid and FPR, vice versa.

## Comparison of Imbalance Dat Methods

As the result in Table 3, imbalance will cause Lasso-Logistic more impact than RF, imbalance data always give much more positive hypothyroid error than balanced data, especially in RUS. But under sample must lose lots of information of original data, luckily sample is very important. SMOTE taking into account the advantages of ROS and RUS, but the result have no significant improvement. That's because the training dataset is the balanced data, and testing, or real situation is imbalance. The distribution of two dataset is different, so that it will give more misclassification.

Table 3: Performance Comparison on Hypothyroidism Data

| Missing Data Method | Data Balancing Method | Classification Method | Confusion Matrix(0.5) | | Error | AUC | TPR | FPR | Best Cut-Off |
|---|---|---|---|---|---|---|---|---|---|
| Imputed | None | Lasso-Logistic | 329.2 | 2.8 | 0.008 | 0.950 | 0.950 | 0.104 | 0.056 |
| | | | 16.0 | 12.4 | 0.562 | | | | |
| | | Random Forest | 330.0 | 2.6 | 0.008 | 0.989 | 0.964 | 0.024 | 0.233 |
| | | | 4.2 | 24.2 | 0.150 | | | | |
| | ROS | Lasso-Logistic | 303.8 | 28.8 | 0.084 | 0.922 | 0.916 | 0.142 | 0.420 |
| | | | 7.4 | 21.0 | 0.262 | | | | |
| | | Random Forest | 319.0 | 13.6 | 0.038 | 0.987 | 0.964 | 0.045 | 0.487 |
| | | | 1.8 | 26.6 | 0.064 | | | | |
| | RUS | Lasso-Logistic | 305.8 | 26.8 | 0.080 | 0.955 | 0.960 | 0.099 | 0.383 |
| | | | 3.8 | 24.6 | 0.140 | | | | |
| | | Random Forest | 318.0 | 14.6 | 0.042 | 0.991 | 0.970 | 0.026 | 0.610 |
| | | | 1.4 | 27.0 | 0.048 | | | | |
| | SMOTE | Lasso-Logistic | 314.2 | 14.4 | 0.044 | 0.958 | 0.908 | 0.070 | 0.365 |
| | | | 5.6 | 22.8 | 0.196 | | | | |
| | | Random Forest | 327.0 | 5.6 | 0.018 | 0.994 | 0.970 | 0.021 | 0.516 |
| | | | 1.8 | 26.6 | 0.063 | | | | |
| Removed | None | Lasso-Logistic | 251.2 | 3.0 | 0.012 | 0.960 | 0.966 | 0.122 | 0.064 |
| | | | 13.2 | 10.4 | 0.552 | | | | |
| | | Random Forest | 244.0 | 2.2 | 0.010 | 0.992 | 0.9981 | 0.035 | 0.306 |
| | | | 3.6 | 20.0 | 0.156 | | | | |
| | ROS | Lasso-Logistic | 222.2 | 24.0 | 0.096 | 0.920 | 0.905 | 0.129 | 0.452 |
| | | | 5.2 | 18.4 | 0.220 | | | | |
| | | Random Forest | 235.4 | 10.8 | 0.044 | 0.976 | 0.964 | 0.044 | 0.532 |
| | | | 1.8 | 21.8 | 0.084 | | | | |
| | RUS | Lasso-Logistic | 229.6 | 16.6 | 0.066 | 0.968 | 0.984 | 0.110 | 0.348 |
| | | | 3.2 | 20.4 | 0.140 | | | | |
| | | Random Forest | 234.8 | 11.4 | 0.046 | 0.991 | 0.990 | 0.044 | 0.541 |
| | | | 0.6 | 23.0 | 0.028 | | | | |
| | SMOTE | Lasso-Logistic | 235.6 | 10.6 | 0.046 | 0.968 | 0.946 | 0.048 | 0.492 |
| | | | 3.2 | 20.4 | 0.146 | | | | |
| | | Random Forest | 242.8 | 3.4 | 0.014 | 0.991 | 0.981 | 0.031 | 0.509 |
| | | | 1.8 | 21.8 | 0.080 | | | | |

### Comparison of Classification Methods

Compare with Lasso-Logistic and Random Forest (RF) , the result in Table 3 shows that Rf, as a nonparametric method, always better than Lasso-logistic. It is because the dataset included lots of factor variables, parametric method gives a linear or nonlinear relationship of the hypothyroid. Sometimes factor variable do not have significant impact of the hypothyroid. In resent years, lots of research found that nonparametric classification methods like CART, Neural Network, SVM, RF always have higher accuracy than Logistic Regression and Discrimination Analysis [13,20].

### New Imputation Random Forest Model

Compared with those models, the results were shown that two variables (query_on_thyroxine and lithium) didn't have much impact in classify the person which suffering from hypothyroidism, even the negative impact of the model. For this reason, these two variables were excluded in the final model. From the models results in Table 3, we weighted the rest of variables importance and ranked was them which was shown in Table 4. As the table shown that, TBG_measured located almost the middle part of the importance, that means although TBG was missing 91.78% of the observations, that should be include in our model. As the multiple imputation didn't preference very well than omit them, thus, a new imputation method was considered. Fortunately, RF was a very useful algorithm, not only in classification and variable selection, but also in imputation[36].

Table 4: Weighted Variables Importance

| Variable | Weight |
|---|---|
| TSH | 96 |
| FTI | 94 |
| $TT_4$ | 80 |
| $T_4U$ | 65 |
| $T_3$ | 60 |
| age | 36 |
| on_thyroxine | 26 |
| thyroid_surgery | 23 |
| query_TRUE | 19 |
| sick | 19 |
| goitre | 7 |
| TBG_measured | 7 |
| TSH_measured | 5 |
| on_antithyroid_medication | 3 |
| sex | 3 |
| pregnant | 2 |
| $T_3$_measured | 2 |
| tumor | 2 |
| query_hyperthyroid | 1 |

After imputation, a RF model was conducted using 90% of observations as training dataset and the rest 10% observations as testing dataset. Compared with lots of parameters value, the parameters in final model are ntree=800 trees to grow, mtry=17 variables randomly sampled as candidates at each split. nPerm=10 times the OOB data are permuted per tree for assessing variable importance.

**Confusion Matrix and AUC**

Since this hypothyroidism dataset is an benchmark data of UCI, lots of machine learning methods performed very well, after imputation and parameters adjustment. Even the imbalanced data, RF classified accuracy is 100% which was shown on Table 5. ROC Curve and AUC was shown on Figure 5. Because no miss-classified in this model, the AUC as equal to 1 and the best cut-off was 0.50125, which almost the middle of the interval [0,1] . Although the data is imbalance, the best cut-off still located near 0.5, it means that the robust of RF, which is ensemble learning method, is much better than other single model.

Table 5: Confusion Matrix of RF Model

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Training | | Testing | |
| | | FALSE | TRUE | FALSE | TRUE |
| Actual | FALSE | 2170 | 0 | 302 | 0 |
| | TRUE | 0 | 135 | 0 | 16 |

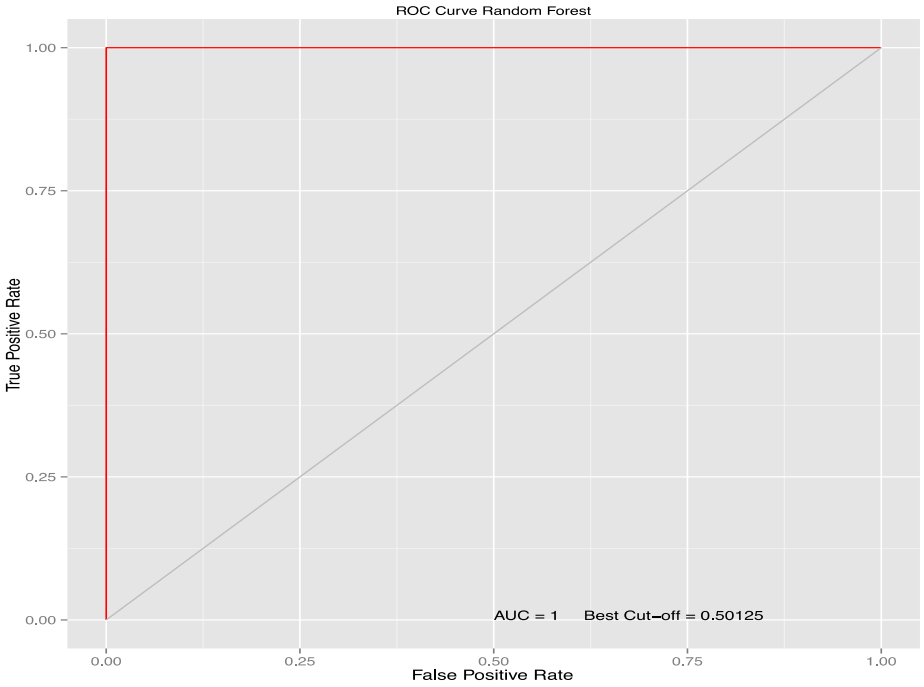


Figure 5: ROC Curve and AUC

### OOB Error and Variable Importance

A Out of Bag (OOB) Error with number of trees is shown on Figure 6. Because the dataset is imbalanced, the error of small proportion side (Hypothyroid=TRUE) goes very high at the beginning when trees grow higher than 200, the error of Hypothyroid=TRUE becomes stabilized. The final OOB estimate of error rate: 1.02% with the Confusion Matrix shown in Table 6. From OOB Error, the model also voted the importance of variables. As Figure 7 shown, TSH, FTI, TBG and TT4 are the most important variables in predicting Hypothyroidism. Age also gives an important impact of Hypothyroidism, even more important than T3 in both Accuracy and Gini criterions. But with Accuracy criteria, thyroid even have a negative impact of the RF.

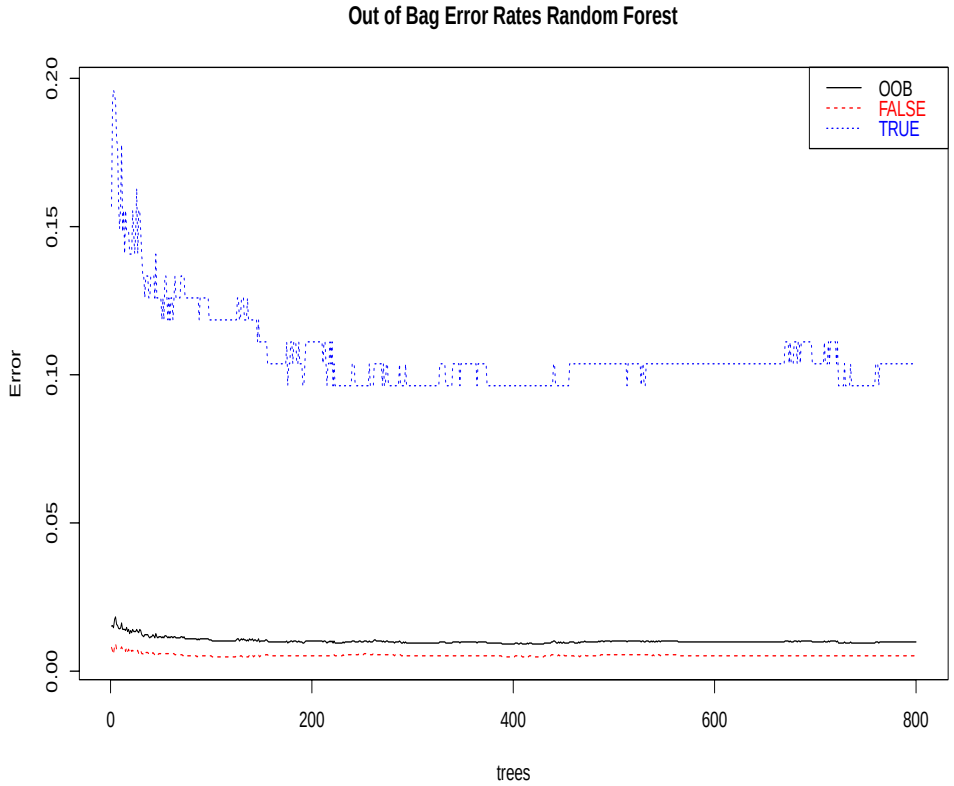


Figure 6: OOB Error of Trees

Table 6: Confusion Matrix of OOB

| | | Predicted | | |
|---|---|---|---|---|
| | | FALSE | TRUE | Error |
| Actual | FALSE | 2695 | 15 | 0.0055 |
| | TRUE | 14 | 121 | 0.1037 |

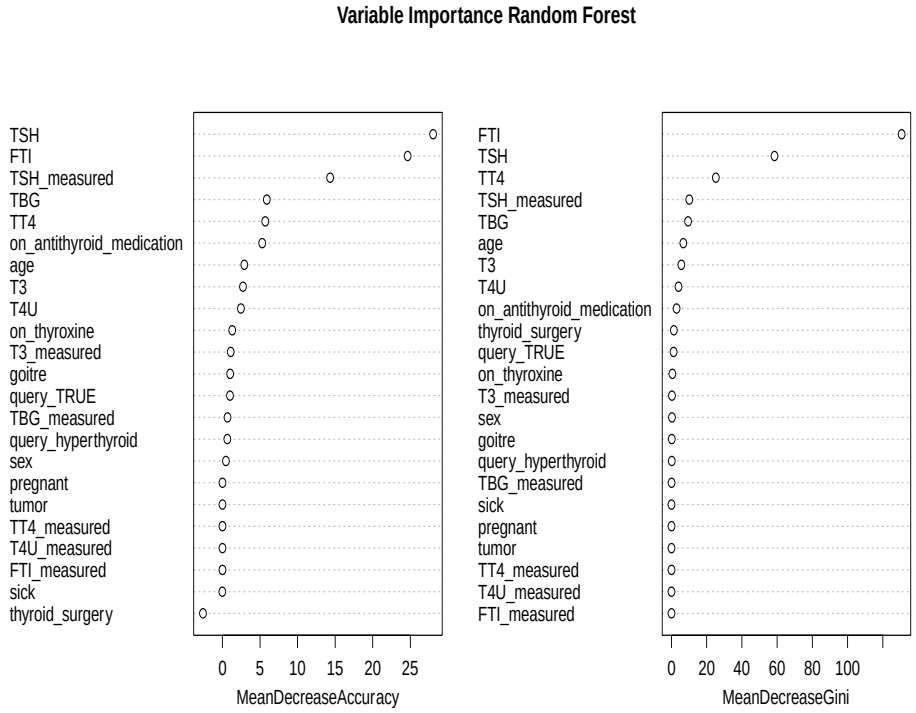


Figure 7: Variable Importance of Final Model

## Final Model

Using all variable model to predict Hypothyroidism probably can make almost 100% accuracy in the real case, But getting these variable always costs a lot of money and time, not really efficient in the real life. Compare with the variable importance in Figure 7, TSH, FTI, TBG, TT4 and age give the majority information of the model. In our final model, only these five variables were used. Confusion Matrix of this new model in Table 7 shows that only one observation was miss-classification with error 0.06. This error will be accepted in the real case. The OOB Error rate is 1.15% against the all-variable model with 1.02% and the AUC is 0.9988 only 0.0012 smaller than the all-variable model.

Table 7: Confusion Matrix of Final Model

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Training | | Testing | |
| | | FALSE | TRUE | FALSE | TRUE |
| Actual | FALSE | 2742 | 0 | 270 | 0 |
| | TRUE | 0 | 135 | 1 | 15 |

## 4. Discussion

The Etiology or causes of hypothyroidism is mainly divided in 2 categories. Congenital agencies of thyroid gland and pituitary gland dysfunction, acquired to surgery, medication of lithium, amiodarone, interleukin and interferon alpha, viral or autoimmune thyroiditis, overdose iodine and radiation therapy. The diagnosis of hypothyroidisms is mainly based on blood test of TSH, T3 and TT4. But in our study, FTI and TBG are much more important than T3, Schulman et al.(1993), Weiss et al.(1990) and Chen et al.(2008) using the similarly dataset also found that TSH and FTI are very important but T3 and TBG are the opposite[33, 39, 8]. The reason is because for the large scale of missing data in TBG, they just omit it in the model, this makes T3 become more important. But after imputation using RF, the result shows that TBG has more importance than T3. Schulman et al.(2003) using Cascade correlation, Local adapt. rates, BP+genetic opt. etc. and Weiss et al.(1990) using CART, PVM, Bayesian etc. also built the model with almost 99.8% accuracy with training data and 98.5% to 99.3% accuracy with testing data. In the final model, only TSH, FTI, TBG, TT4 and age will provide a very high accuracy. Traditional method like Mbah et al.(2011) and Meher et al.(2013) provided a Logistic Regression to predict hypothyroidism like score card and attributes confidence interval[26, 27]. But in our previous result shown that a parametric method sometimes cannot fit very well than nonparametric method, even ensemble nonparametric machine learning method. Also, nowadays, with the development of computer science, we are no need to predict using our brain with score card or something else, just input the data, it will give the probability of a person who is suffering a disease.

## 5. Conclusions

This study found a ensemble machine learning model to predict a patient who will suffer the hypothyroidism. Following with the dataset from UCI and doctor's guide, we tried to deal the data with different missing data method, data balancing method and classification method. After the comparison with these model, RF with RUS will be better to fit the model. But some new idea found from RF also can do imputation, and with TBG was imputed by RF, The model can predict with 100% accuracy in both training data and testing data. Finally, only using TSH, FTI, TBG, TT4 and age to built a RF model is much more suitable in the real case.

## 6. Limitations and Future Study

This Although this study has more than 3000 datasets, it would be better if more detail data were collected especially thyroxine binding globulin. More researches of congenital or acquired hypothyroidism should be studied in different parts of the world as well as different races[1].

Hashimoto's thyroiditis is an autoimmune disease that attacks thyroid gland and causes of hypothyroidism and risk of thyroid cancer, but this dataset does not classify different type of thyroid. Medication of lithium, carbamazepine and valproate increase risk of hypothyroidism.

With the promotion of thyroid function test, early screening and detection of hypothyroidism get better early treatment [9, 10].

Lasso-Logistic didn't perform well in this data with lots of binary variables which makes nonparametric model better than parametric model. Also, Lasso-Logistic model has no robust with imbalanced data. A new adjustment method should be built in the future.

## 7. Author Contributions

Conceived and designed the research process: YJ. Performed the lecture review: KK TW ML. Analyzed the data: YJ KK TW ML. Contributed reagents/materials/analysis tools: YJ KK. Wrote the paper: YJ KK TW ML.

## References


[1] Fuster V, et al. Medical Underwriting for Life Insurance. McGraw-Hill's AccessMedicine. 2008;.

[2] Pandey S, Miri R, Tandan S. Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques. In: International Journal of Engineering Research and Technology. vol. 2.ESRSA Publications; 2013. .

[3] National Health Insurance Administration MoH, Welfare ROC Taiwan. National Health Insurance Annual Report 2014-2015. NHIRD; 2014.

[4] Gaikwad S, Pise N. AN EXPERIMENTAL STUDY ON HYPOTHYROID USING ROTATION FOREST. International Journal of Data Mining & Knowledge Management Process. 2014;4(6):31.

[5] Margret JJ, Lakshmipathi B, Kumar SA. Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules. molecular biology. 2012;3:4.

[6] Akbaş A, Turhal U, Babur S, Avci C. Performance improvement with combining multiple approaches to diagnosis of thyroid cancer. Engineering. 2013;5(10):264.

[7] Kousarrizi MN, Seiti F, Teshnehlab M. An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification. International Journal of Electrical & Computer Sciences IJECS-IJENS. 2012;12(01):13–20.

[8] Radwan E, Assiri AM. Thyroid diagnosis based technique on rough sets with modified similarity relation. Thyroid. 2013;4(10).

[9] Asuncion A, Newman D. UCI machine learning repository; 2007.

[10] Dayan CM. Interpretation of thyroid function tests. The Lancet. 2001;357(9256):619–624.

[11] Soley-Bori M. Dealing with missing data: Key assumptions and methods for applied analysis. Technical Report; 2013.

[12] Graham JW, Cumsille PE, Elek-Fisk E. Methods for handling missing data. Handbook of psychology. 2003;.

[13] Schafer JL. Imputation of missing covariates under a multivariate linear mixed model. Tech; 1997.

[14] Elkan C. The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17. Citeseer; 2001. p. 973–978.

[15] Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE; 2003. p. 435–442.

[16] Chawla NV. Data mining for imbalanced datasets: An overview. In: Data Mining and Knowledge Discovery Handbook. Springer; 2010. p. 875–886.

[17] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;p. 321–357.

[18] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;p. 267–288.

[19] Breiman L. Better subset regression using the nonnegative garrote. Technometrics. 1995;37(4):373–384.

[20] Tibshirani R, et al. The lasso method for variable selection in the Cox model. Statistics in medicine.1997;16(4):385–395.

[21] Lokhorst J. The lasso and generalised linear models. Honors Project, The University of Adelaide,Australia. 1999;.
[22] Roth V. The generalized LASSO. Neural Networks, IEEE Transactions on. 2004;15(1):16–28.
[23] Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics. 2003;19(17):2246–2253.
[24] Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics. 2007;49(3):291–304.
[25] Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2005;27(6):957–968.
[26] Breiman L. Random forests. Machine learning. 2001;45(1):5–32.
[27] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer. 2005;27(2):83–85.
[28] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making. 2011;11(1):51.
[29] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
[30] Fang K, Wu J, Zhu J, Xie B. Forecasting of Credit Card Credit Risk Under Asymmetric Information Based on Nonparametric Random Forests. Economic Research S. 2010;1:97–107.
[31] Huang CL, Chen MC, Wang CJ. Credit scoring with a data mining approach based on support vector machines. Expert systems with applications. 2007;33(4):847–856.
[32] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–118.
[33] Schiffmann W, Joost M, Werner R. Comparison of optimized backpropagation algorithms. In: ESANN. vol. 93. Citeseer; 1993. p. 97–104.
[34] Weiss SM, Kapouleas I. An empirical comparison of pattern recognition, neural nets and machine learning classification methods. Readings in machine learning. 1990;p. 177–183.
[35] Chen SY, Xu WB. Rule Induction on Mining Large Database. In: 2008 Information Communication Technology Management and Applications Conference. Shu-Te University; 2008. p. 65–76.
[36] Mbah AU, Ejim EC, Onodugo OD, Ezugwu FO, Eze MI, Nkwo PO, et al. Two logistic models for the prediction of hypothyroidism in pregnancy. BMC research notes. 2011;4(1):205.
[37] Meher L, Raveendranathan S, Kota S, Sarangi J, Jali S, et al. Prevalence of hypothyroidism in patients with metabolic syndrome. Thyroid Research and Practice. 2013;10(2):60.
[38] Ahn D, Heo SJ, Park JH, Kim JH, Sohn JH, Park JY, et al. Clinical relationship between Hashimoto's thyroiditis and papillary thyroid cancer. Acta Oncologica. 2011;50(8):1228–1234.
[39] Chen YK, Lin C, Cheng FT, Sung F, Kao C. Cancer risk in patients with Hashimoto's thyroiditis: a nationwide cohort study. British journal of cancer. 2013;109(9):2496–2501.

[40] Chen CY, Lee KT, Lee CTC, Lai WT, Huang YB. Epidemiology and clinical characteristics of congenital hypothyroidism in an Asian population: a nationwide population-based study. Journal of Epidemiology. 2013;23(2):85.

Received Aril 10, 2015; accepted November 24, 2015.