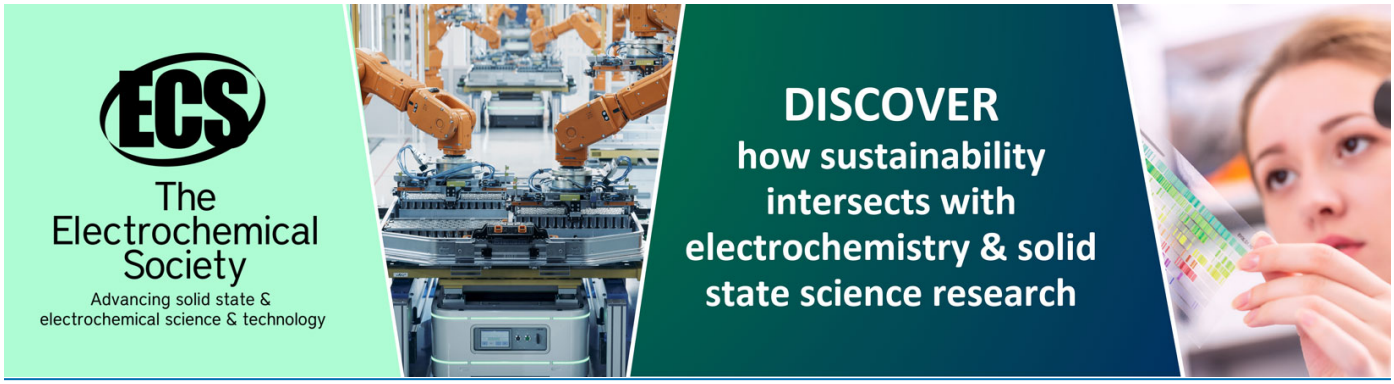
Identifying of factor associated with parkinson's disease subtypes using random forest



IOP Conf. Series: Journal of Physics: Conf. Series **1108** (2018) 012064 doi:10.1088/1742-6596/1108/1/012064

Identifying of factor associated with parkinson's disease subtypes using random forest

E Latifah[1,*], S Abdullah[1] and S M Soemartojo[1]

[1] Departement of Mathematics, Universitas Indonesia, Depok, 16424, Jawa Barat, Indonesia
*Corresponding author: esti.latifah@sci.ui.ac.id

Abstract. Parkinson's disease (PD) is a neurodegenerative disorder that caused by the result of lack of dopamine in a specific area of the brain called bangsal ganglia. It is a long term degenerative disorder of the central nervous system that mainly effect the motor system that has some impacts such as difficulty in speech, problem in swallowing, and dressing, trouble with handwriting or even doing some activities, and tremor. Based on this problem, researchers use the Parkinson's Progression Markers Initiative (PPMI) database to classify subtypes: Tremor Dominant (TD) and Postural Instability Gait Difficulty (PIGD). Identifying the factors of Parkinson's disease subtypes is crucial in understanding the appropriate therapy for Parkinson's disease patient. Furthermore, it gives some characteristics of patient that is classified into TD or PIGD. Classification method is used to identify the factors of parkinson's disease subtypes on 207 patient with PD and 47 variables obtained from Movement Disorder Society-Unified Parkinson Disease Rating Scale (MDS-UPDRS) part II and part III in event V12. The result is PD patient who is classified to PIGD class have the lower value in constancy of rest tremor, rest tremore amplitude (RUE), tremor, rest tremor amplitude (LUE), and postural tremor of right hand than PD patient with TD and the higher value in postural stability, walking and balance, and freezing than PD patient with TD.

1. Introduction

Parkinson's disease (PD) is a disorder of brain function that results in the degeneration of nerve cells in a specific area of the brain called basal ganglia. Degeneration of these nerve cells is due to dopamine degradation in substansia nigra pars compacta (SNC) and striatumcorpus. Dopamine decrease causes the brain activity cannot function normally that has impact to some symptoms on motor and nonmotor aspects of the body. Notice that Parkinson's disease is the most common neurodegenerative disease after Alzheimer's [11].

Parkinson's disease affects millions of the world's population or about 1% of the world's population [8]. Based on WHO data, the incidence of Parkinson's disease in Asia shows that 1.5 to 8.7 cases per year in China and Taiwan, whereas in Singapore, Wakayama and Japan there are 6.7 to 8.3 cases per year. It happens in the range of age 60 to 69. The prevalence of Parkinson's disease in Indonesia is 876,665 inhabitants [8]. The deaths of Parkinson's patients are usually not caused by Parkinson's disease itself but rather because of secondary infection [5].

There are two subtypes of Parkinson's disease: a subtype that has tremor dominant (TD) and postural instability gait disorder (PIGD) [7]. Both of these Parkinson's subtypes differ in clinical, imaging, genetic, and pathological features [6]. Parkinson with tremor dominant subtype (TD) exhibits slower progression and has a better prognosis than the postural subtype of instability gait disorder (PIGD) [15]. Parkinson's disease with TD and PIGD have different forms of motoric and non-motoric symptoms. Therefore, researchers want to examine the motor factors of Parkinson's disease that influence the

subtype of Parkinson's disease using classification method that is random forest so that proper and efficient therapy can be done.

2. Data and Variables

Parkinson's Progression Markers Initiative (PPMI) data base is a study to verify the development of Parkinson's disease. The study was conducted on 145 patients with Parkinson's disease obtained from the PPMI database. The data were taken under the Movement Disorder Society-Unified Parkinson's Disease Rating Scales (MDSUPDRS) Part 2 and Part 3. Respondents selected for this study were Parkinson's patients with the following criteria:

- Patients with Parkinson's disease who have been undergoing treatment for 36 months (V12).
- Patients observed are only sufferers of motor disorders.

The following variables are used in this study:

1. Response Variable (Y)
   Y is the ratio of certain variables in the MDS-UPDRS part 2 and part 3. The mean of MDS-UPDRS items 2.10, 3.15a, 3.15b, 3.16a, 3.16b, 3.17a, 3.17b, 3.17c, 3.17d, 3.17e, and 3.18 is divided by the mean of MDS-UPDRS items 2.12, 2.13, 3.10, 3.11, and 3.12. If the ratio is greater than or equal to 1.15, then the patient is classified with TD. If the ratio is less than or equal to 0.90, then the patient is classified with PIGD [14].
2. Predictor Variable (X)
   Variable predictors consist of :
   - Data MDS-UPDRS Part 2 (13 variables). The data is a test of the motor aspects of daily life of people with Parkinson's such as speaking, eating, writing, and so on.
   - Data MDS-UPDRS Part 3 (34 variables). The data is a test performed physically by medical personnel such as response when spoken to, hand and foot movements, and so on.

3. Classification Methode

In this section the statistical methods developed and used in this study will be described.

3.1 Decision Tree

Decision tree is one of the methods used to classify a sample of unknown class data into existing classes. The use of the decision tree method depends on the response variable. If the response variable is a continuous variable, then the method used is the regression tree. If the response variable is a categorical variable, then the method used classification tree. However, in this study will only be focused on the classification problem. There are many algorithms for building a decision tree, such as ID3 (Iterative Dichotomiser 3) [16] and its further development C4.5 [17], and CART (classification and regression tree) [1].

In growing decision tree, it begins at the root node. The node that contains all the data to build a tree. From the root node will be formed branches and other nodes called parent node. Parent node itself will divide and generate new node is called child node that consist of right node and left node. Child node that can not splitting anymore is called terminal node. To understand the concept, see the ilustratin in Figure 1.

Moreover, each parent node will split into exactly two child nodes (right node and left node) based on a particular predictor variable. This process is called binary partitioning. Next, the node separation process is repeated on every child node (The child node do splitting successfully will be a parent node and so on) until the splitting process cannot be done anymore. The process of formation is known as binary recursive partitioning.

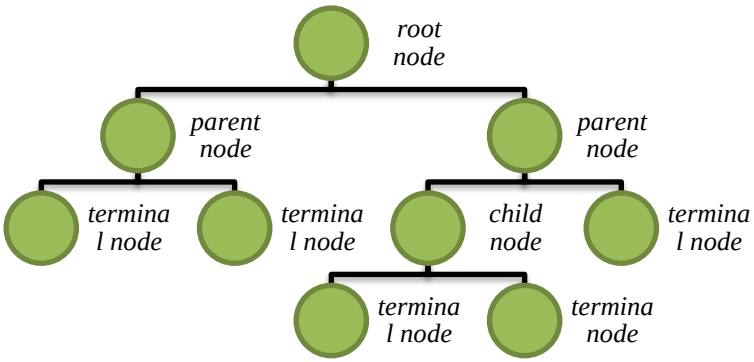


Figure 1. Structure of decision tree.

Splitting is the process of dividing the parent node into two child nodes through certain splitting rules. Parent node is divided into two child nodes based on the goodness of split criterion. The splitting criterion must fulfill the characteristic that the grouping results in the child node are more homogeneous than the preceding nodes.

In the classification and regression tree (CART), separation criteria are based on Gini index [1]. Suppose that there are n observations. These observations will be grouped into two classes (i.e. $j = A, B$). The classification rule is based on a number of M predictor variables (notation x_m, $m = 1, \dots, M$). At each node (notation node t), the Gini index is expressed as follows,

$$Gini(t) = 1 - \sum_j {p_j}^2$$

with,

- $p_j = \frac{N_j(t)}{N_t}$ is the proportion of the number of objects that enter into class j on a node t.
- $N_j(t)$ is the number of objects that enter into class j on a node$t$.
- N_t is the total number of objects on a node t.

The Gini index describes the non-homogenous child nodes of the node t due to the splitting based on x_m variabel. The Gini index has a range of values between 0 and 1. The lower value of Gini index, the greater the homogeneity of a node so that the better process of separating objects into existing classes.

It takes a comparison of the non-homogeneous levels between the parent nodes (before splitting) with the non-homogeneous levels of the child node (after splitting) to determine how well an predictor variable is in splittig a node. A predictor that produces the maximum difference of Gini index between the parent node and child node to be selected as the best splitter. The difference of Gini index can be calculated as follows,

$$\Delta G(t, m) = G(t_P) - p_R G(t_R) - p_L G(t_L)$$

with,

- $\Delta G(t, m)$ is the difference of Gini index value.
- $G(t_P)$ is the Gini index valuein parent node.
- $G(t_R)$ is the Gini index value on the right node.
- $G(t_L)$ is the Gini index value on the left node.
- p_R dan p_L is the proportion of the number of objects on right and left node.

The splitting process will continue until there is no increase in homogeneity in the child node is generated.

Unfortunately, decision trees are prone to overfitting [12]. Overfitting occurs when the algorithm continues to grow the tree to reduce the error for the training data, yet this results in an increased error for the test data. The training data is a set of data that is used to build the tree. The test data is a set of data that is used to evaluate the predictive accuracy of the tree.

3.2 Random Forest

The random forest is an ensemble approach that offers a robust alternative to the decision trees [2]. In this method, the forming structure is similar to the decision tree method but the way of working of the two methods are quite different. There are two important steps in this method. First, the sample is obtained by taking a random sample with the replacement of data. It is called bootstrapping [2]. Second, the certain size of candidate variable is randomly taken of all predictors that used for splitting a node. It is called random feature selection [2]. As a result, the two steps above will produce many trees to form a forest with the differences of size and shape of the trees.

Suppose that there are n observations and consist of M predictor variables so that random forest is built using the following step [3]:

1. Do n random sampling of a random sample with replacement in training data. The sample generated from this step is called the bootstrap sample.
2. By using a bootstrap sample, the tree is built until it reaches its maximum size. At each node, the split selection is performed by taking $p \ll M$ of the predictors randomly. This step is called random feature selection. The best splitter is chosen from that candidat variable p.
3. Repeat steps 1 and 2 until$K$ trees are built so that the forest is formed.

The method based on bootstrapping and random feature selection is to reduce uncertainty in model predictions. In the random forest method, multiple decision trees are fitted, where each tree is fitted to a random subset of the data, and bootstrap sampled with replacement. However, to obtain a good bootstrapping result is not only through with replacement, but also be done sampling without replacement if the sample size is large enough [4].

Within a single tree, splits are determined from a random subset of predictors sampled at each node, as a means of reducing correlation among trees. The number of predictors randomly selected for each node is constant at $\sqrt{M}$ [2], where M is the total number of predictor variables. The decision on the best split is based on the CART algorithm [1] as previously described in the decision tree section. There is no pruning is performed in the growing process of each tree.

Once a tree is grown, a prediction is made for each of the cases in the out-of-bag (OOB) data. The OOB data are a subset of the bootstrap sample that is not included in the model building. For each tree K where case i is in the OOB data, class j is assigned for case i if i is predicted to be in that class. The class with the majority of assignments will be the assigned class for case i. The random forest accuracy is then calculated as the proportion of OOB cases that are correctly classified. Hence, the random forest method is appealing for its internal cross-validation embedded in the procedure.

To find the most significant variable, random forest is used Mean Decrease Gini (MDG). MDG is a measurement to find variable importance in the model. The larger MDG value of a variable predictor, the variable is more importance. MDG can be calculated as follows [3]

$$MDG_m = \frac{1}{k}\sum_t [\Delta G(t,m) I(t,m)]$$

with,

- $\Delta G(t,m)$ is the decreasing Gini index value for x_m.
- $I(t,m)$ is indicator function. $I(t,m) = \begin{cases} 1, & x_m \text{ splitting at node } t \\ 0, & \text{others} \end{cases}$
- $k = 1,2,\dots,K$ is the number of trees formed in random forest.

4. The results for modelling data by Random Forest

The random forest method is run by using software R-version 3.4.3 (free download). Then, the following output is obtained from running the data on software R.

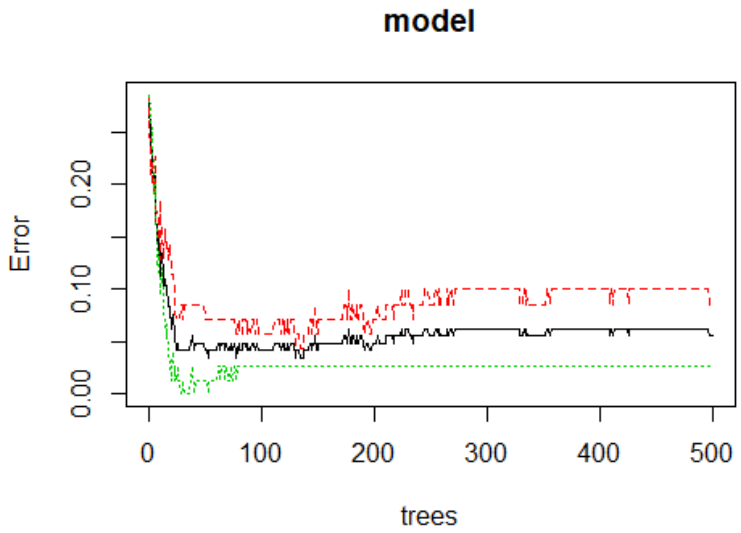


Figure 2. Error model plot with 500 trees. The red is misclassification of PIGD class plotting, the green is misclassification of TD class plotting, and the black is OOB error plotting.

Figure 2 shows the misclassifcation (error) for different classes. The red plot describes the error of the sample classified into the PIGD class, but it is wrongly predicted. In other words, the predicted PIGD sample error goes to the TD class. The error value in this sample is more constant when the tree is formed by 300 trees. In the picture, if we built a number of 300 to 500 trees, then the error is about 8%. The green plot describes the error of the predicted TD sample which goes wrongly to the PIGD class. This plot is different from PIGD sample error plot. When the tree is formed by 100 trees, the error value will be constant. If more trees are built (i.e. 100 to 500 trees), then the TD sample error is about 2%. Furthermore, the black plot is OOB error. From the picture shows if the more trees are formed, the OOB error will be more constant. When a tree is formed around 300 to 500 trees, the OOB error is about 5%. It means that random forest able to predict the patient with PD in the amount of 95%.

Afterward, Table 1 is given to find out the accuracy of the model.

Table 1. The results of confusion matrix on Parkinson's disease from software R.

		Predicted class	
Actual Class		PIGD	TD
	PIGD	64	6
	TD	2	73

Table 1 shows the confusion matrix on Parkinson's disease. The accuration of the model can easily be obtained from the confusion matrix. The accuration model is about 94,5%. It means that the model able to classify Parkinson's disease correctly.

```
Call:
 randomForest(formula = VAR.RESPON ~ ., data = trainParkins
on,      ntree = 500, mtry = 6, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 5.52%
```

Figure 3. Output R.

The random forest model depends on $ntree$ and $mtry$. $ntree$ is how many trees will be built, while $mtry$ is the number of candidate variables used to split each node. Figure 3 shows that in this model, the tree to be built is 500 trees and the number of variables used for splitting is 6 variables for each node.

The random forest will produce an OOB error. From the Figure 2, the value of OOB error approach its OOB error plotting. The OOB error is about 5.52%. It can be interpreted that the resulting accuracy for the prediction sample is 94.48%.

Thereafer, Figure 4 shows that the rank of all predictors. The constancy of rest variable has the higest MDG value overall predictors, whereasin the second place is rest tremor amplitude (RUE) variable. Tremor variable is in the third place after two variables above. In the next place is rest tremor mplitude (LUE) and postural stability variable. Further inforrmations can be seen in the picture below.

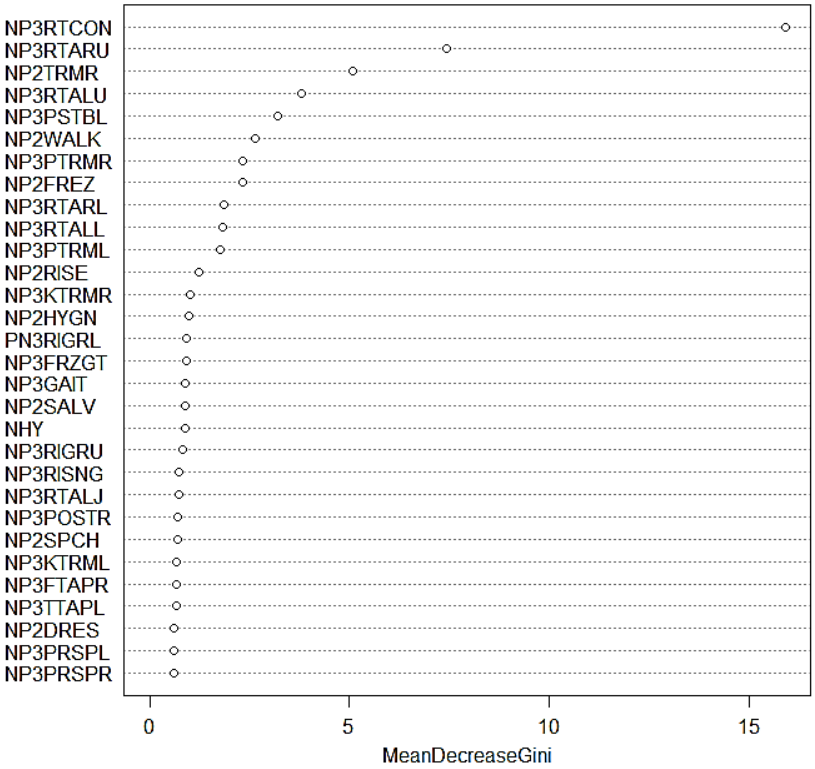


Figure 4. Variable importance plotting using MDG.

5. Conclusion

Data of 207 patients with PD were analysed to determine factors that associated with Parkinson's disease subtypes. The factors of PD subtypes have been identified. There are some importance variables based on MDG with the value > 5 such as constancy of rest tremor, rest tremore amplitude (RUE), tremor, rest tremor amplitude (LUE), postural stability, walking and balance, postural tremor of right hand, freezing. Those variabeles are the factors that influence Parkinson's disease subtypes. PD patient who is classified to PIGD class have the lower value in constancy of rest tremor, rest tremore amplitude (RUE), tremor, rest tremor amplitude (LUE), and postural tremor of right hand than PD patient with TD and the higher value in postural stability, walking and balance, and freezing than PD patient with TD.

6. Acknowledgments

This research is funded by Universitas Indonesia via PITTA 2018.

7. Reference

[1] Breiman L., et al 1984 *Classification and Regression trees* (Wadsworth: Wadsworth International Group)
[2] Breiman L 2001 Random Forests *J. of Machine Learning*, **45** pp 5–32.
[3] Breiman L and Cutler A https://info.salford-systems.com/an-introduction-to-random-forests-for-beginners (accessed: Jan 20th 2018).
[4] Buja A and Stuetzle W 2006 Observations on Bagging *J. of Statistica Sinica* **16** pp 323-351

[5] Joesoef A A 2007 *Parkinson's Disease : Basic Science* in *Parkinson's Disease & Other Movement Dissorder* (Jakarta: Pustaka Cendikia Press)
[6] Marras C and Lang A 2013 Parkinson's Disease Subtypes: Lost in Translation? *J. Neurol Psychiatry*, **84**(4)
[7] McDermott M P, et al 1995 Factors Predictive of The Need for Levodopa Therapy in Early, Untreated Parkinson's Disease *Archives of neurology* **52**(6) pp 565-570.
[8] Noviani, et al 2010 Hubungan antara Merokok Dengan Penyakit Parkinson Di RSUD Prof. Dr. Margono Soekarjo Purwokerto *Journal of health*, **4**.
[9] Parkinson's Foundation http:/www.parkinson.org (accessed: June 3rd 2018).
[10] Parkinson's Progressive Markers Initiative https://www.ppmi.org (accessed: April 5th 2018).
[11] PERDOSSI 2008 *Modul Gangguan Gerak Penyakit Parkinson*.
[12] Rokach L and Maimon O 2014 *Proactive data mining with decision trees* (New York: Springer Science & Business Media)
[13] Abdullah S 2017 *Statistical Methods for Modelling Fall and Symptoms Progression in Patients with Early Stage of Parkinson's disease* (Brisbane: Queensland University of Technology)
[14] Stebbins G T, et al 2013 How to Identify Tremor Dominant and Postural Instability/Gait Difficulty Groups with The Movement Disorder Society Unified Parkinson's Disease Rating Scale: Comparison with The Unified Parkinson's Disease Rating Scale *Movement Disorders* **28**(5) pp 668-670.
[15] Thenganatt M A and Jankovic J 2014 Parkinson Disease Subtypes *JAMA neurology*, **71**(4) pp 499-504.
[16] Quinlan J R 1986 Induction of Decision Trees *J. of Machine Learning*, **1** pp 81-106.
[17] Quinlan, J, R 1993 *C4.5: Program for Machone Learning* (New York: Spriger)