COMP 309
Assignment three

Name: Siwen Feng
ID: 300363512
Dataset: Tramping

Core

Business understanding:

The business objective of this assignment is to predict the time consumption of walking track by using given attributes(length, difficulty, and location). Application of this model can help the development of tourism. The local government of the walking track can use this technique to categorize tracks and give walking suggestion if the tourist wants to visit.

Data understanding:

I collect the initial data from Kaggle's competition, the datasets have been split into the training set and test set. we got 5 attributes, they are the difficulty, shape length, latitude, and longitude respectively. The task is to classify instances into 3 different classes, they are 0, 1, 2. The distinct classes determined that classification is the best way to make the prediction. In the following part of the report, I will address how was the data being cleaned and optimized.

Data preparation:

Examining data quality:

This dataset is well organized and has a clear structure. However, from the observation of the dataset, we still can find problems for the dataset which limit us to get better performance.

- Irrelated attributes.
- Outliers.
- Imbalanced class.

Attributes selection:

The original training dataset has low complexity and easy to understand. However, the relation between the targeted time and attributes is uncertain. In this step, I will use the attributes selection method to filter out irrelated attributes. Correlation attributes eval is on the right. Based on the diagram, we can find that location and Id only have little relation with our targeted classes. I cleared those two attributes as noises. As anticipated, the performance increase dramatically.

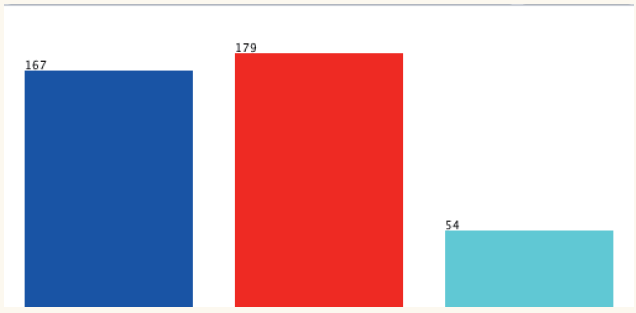
```
average merit      average rank  attribute
 0.33  +- 0.006    1   +- 0      3 Shape__Length
 0.246 +- 0.005    2   +- 0      2 difficulty
 0.091 +- 0.013    3.3 +- 0.46   5 Y
 0.076 +- 0.012    3.7 +- 0.46   4 X
 0.016 +- 0.01     5   +- 0      1 Id
```

Outliers or extreme values:

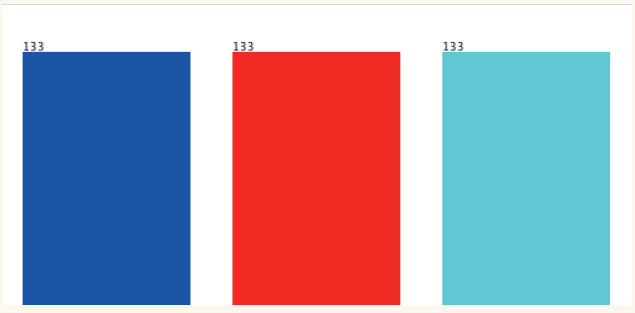
In our dataset, it excludes many other attributes regarding the difficulty, Thus, some track even very short and still can be advanced level track. Some long length track also defined as easy level. I consider those values as extreme values which may lead the model to get a less accurate prediction. I preferred to replace extreme value to missing value. Then, the model based on the training set can be more accurate.

Balance dataset:

Initial instances from distinct classes are imbalanced, class 2, with a small number of instances. To give a fair evaluation, I use the resample function to balance dataset into an equal distribution. Thus, the characteristics of the data can be more obvious.

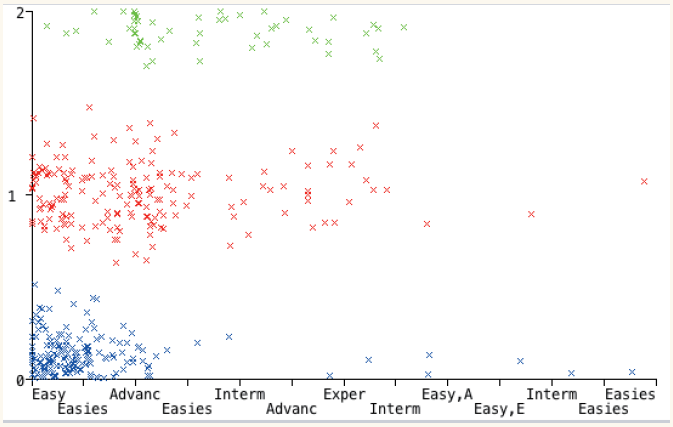


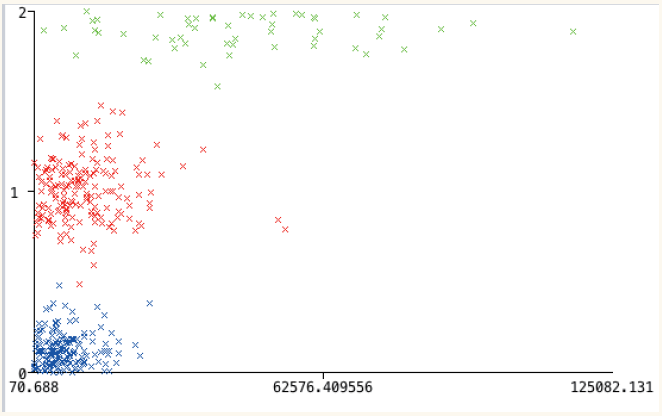
Before



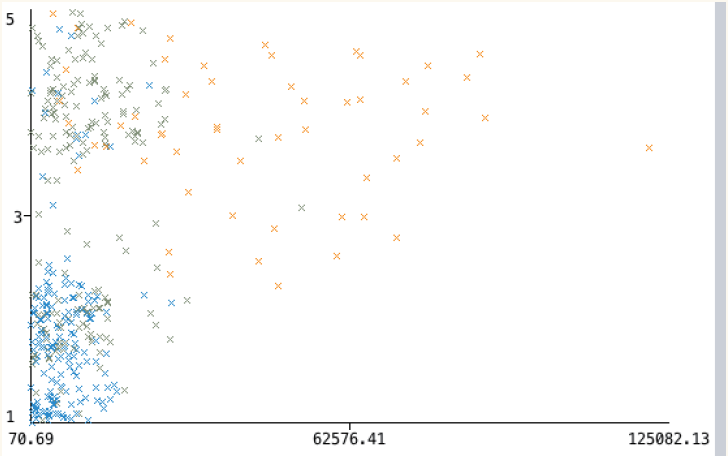
After

Feature interactions:





These diagrams show that both difficulty and shape length has a positive correlation with the targeted class. Majority of "class 0" instances have a short length and easy difficulty. Class 2 is more spreading out and got more instances in the advanced level. Meantime, the length is larger as well. Furtherly, the relations between difficulty and shape length also interact with each other. In the following part, we can explore the relationships between these two attributes. Thus, two attributes can be used for classification.



For simplifying the attributes, I replace difficulty to numbers. 1-5 representing easiest, easy, intermediate, advanced, and expert. As we can see left, almost every instance with short length are defined as easy or easiest. Longer paths have higher difficult level. Hence, we can deduce that shape length and difficulty have a correlation. At this stage, we still cannot identify which one is dependent value and which one is independent value. I will test the influence brought by both of them and address best result.

Modeling:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         321               80.25   %
Incorrectly Classified Instances        79               19.75   %
Kappa statistic                          0.6719
Mean absolute error                      0.1834
Root mean squared error                  0.3154
Relative absolute error                 45.2554 %
Root relative squared error             70.0917 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.862     0.159     0.796       0.862   0.828      0.697    0.911     0.835     0
               0.765     0.163     0.792       0.765   0.778      0.605    0.866     0.844     1
               0.741     0.017     0.870       0.741   0.800      0.775    0.965     0.839     2
Weighted Avg.  0.803     0.142     0.804       0.803   0.802      0.666    0.898     0.839

=== Confusion Matrix ===

   a   b   c   <-- classified as
 144  22   1 |   a = 0
  37 137   5 |   b = 1
   0  14  40 |   c = 2
```

Name	Submitted	Wait time	Execution time	Score
AnswerKeyDummy.csv	4 days ago	0 seconds	0 seconds	0.94444

After tried few techniques, I found that Multiple layer perceptron got highest accuracy for the training set. The result for testing set, I get 91% eventually, It is a good result overall.

Important findings:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         268               67      %
Incorrectly Classified Instances       132               33      %
Kappa statistic                          0.4422
Mean absolute error                      0.2905
Root mean squared error                  0.3853
Relative absolute error                 71.7002 %
Root relative squared error             85.6398 %
Total Number of Instances              400
```

Only use difficulty to predict.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         323               80.75   %
Incorrectly Classified Instances        77               19.25   %
Kappa statistic                          0.6795
Mean absolute error                      0.1814
Root mean squared error                  0.3165
Relative absolute error                 44.7683 %
Root relative squared error             70.3346 %
Total Number of Instances              400
```

Only use shape length to predict.

As we can see, on the left, if I only use difficulty to predict the training dataset, It will give us a poor performance. only 67%. I apply the model to predict testing set. The result is very bad. We can only get about 80% accuracy. However, if we use length to predict. The results for both testing and training are improved. Training set increased from 80.25 to 80.75. the testing set prediction rise to 94.4%. Based on the performance, we have reason to infer that difficulty could be considered as less related attributes for our predicting process.

Completion(Old dataset)

Data preparation:

Examining the quality of data:

In the completion part, we are given two training datasets with slightly different attributes. The second dataset contains all the information we need. However, training dataset 1 miss shape length column. At the same time, these two datasets have many irrelated noises which should be clear. This dataset consumption time is inconsistent numbers, we also should consistent time to minutes. There are two steps which I can do initially.

- Clear noises.
- Time conversion

Clear dataset:

X	Y	OBJECTID	name	introductior	difficulty	completion	hasAlerts	introductior	walkingAnd	dateLoadedToGIS
176.296638	-39.93176	1909	Aeanes Busl	'Take a walk	'Easiest'	'30 min'	'Refer to DC	https://www	https://www	2018-06-22T04:19:12.000Z
172.401476	-42.379524	1910	'Alpine Natu	'The Alpine	'Easiest'	'20 min'	'Refer to DC	https://www	https://www	2018-06-22T04:19:14.000Z
178.30843	-38.231307	1911	'Anaura Bay	'This track p	'Easy'	'2 hr'	'Refer to DC	https://www	https://www	2018-06-22T04:19:14.000Z
172.622902	-40.752098	1912	'Aorere Golc	'Aorere Golc	'Easy'	'3 hr circuit'	'Refer to DC	https://www	https://www	2018-06-22T04:19:15.000Z
174.093759	-39.272137	1913	'Around the	'The challen	'Advanced,E	'4 - 5 days'	'Refer to DC	https://www	https://www	2018-06-22T04:19:15.000Z
175.138814	-39.727109	1914	'Atene Skyli	'The Atene S	'Advanced'	'6 - 8 hr'	'Refer to DC	https://www	https://www	2018-06-22T04:19:15.000Z

As we can see above, this dataset contains many attributes. Locations, introductions, etc. In the beginning, I failed to load the data in, because Weka is not support loading “space”. The introduction comes with a lot of spaces. For the rest, We have discussed in core part. The only valuable attribute is the shape length. Thus, we filter out rest attributes and leave shape length for predicting.

Time Conversion:

Original time	Calculations	Modified time
1hr	1*60	60
3-4hr	3.5*60	210
1 day	1*12*60	720
2hr one way	2*2*60	240

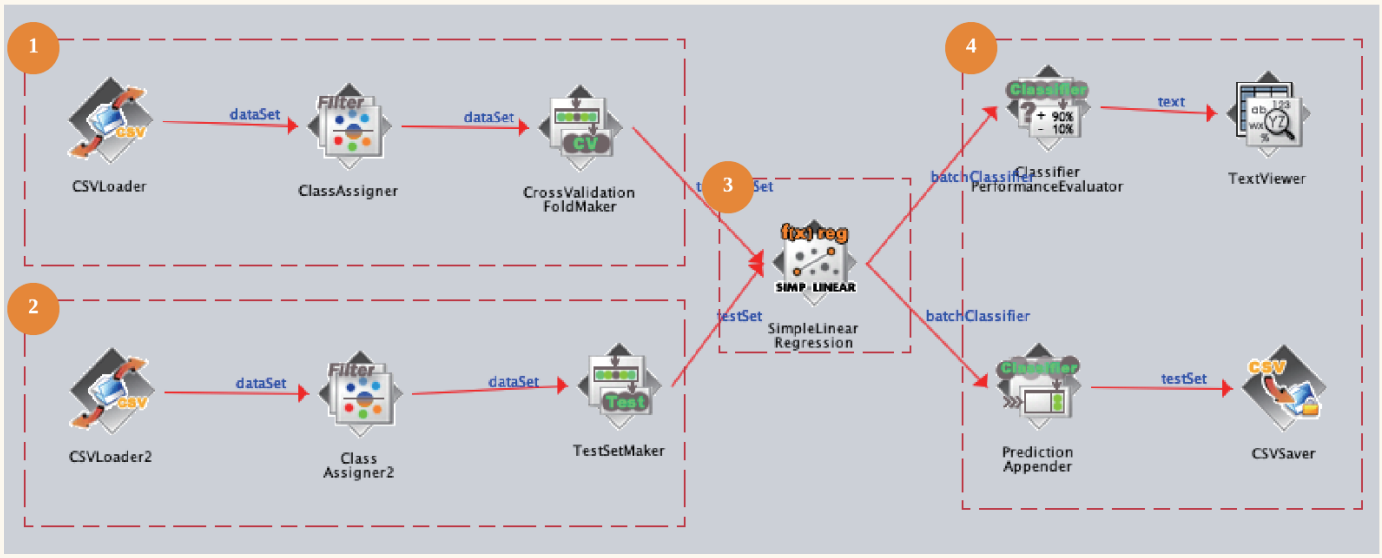
Even though I cleared useless attributes, I still failed to load my data into Weka. The reason is that the completion time contains space. The next step is to convert all inconsistent time to mins. However, it is not an easy job. There are few standards on the left that I followed for altering time to mins.

It doesn’t make sense to consider 1 day as 24 hours, because people need rest and sleep. Thus, I consider 1 day as 12 hours walking is more reasonable. I also chose average value as the representative if the completion time is a range of time. I considered every tramp is returning and doubled time if they are just one way.

Initial design:

For our modified dataset target is time and time is the numeric type. We are not allowed to use classification for this scenario. Thus, I design my initial system with simple regression. I simply trained my data by using regression and apply the model to testing data. The pipeline is down below:

Pipeline by Weka:



1. I load training data and use 10 cross-validations to prevent from overfitting. Time is the target and I set time as a class.

2. In this part I set testing set into the system, then, I can use the model to predict all instance in this testing set.

3. Initially, two inputs are both numeric. Linear regression is feasible to predict numbers. We verified the correlation between time and length in the core part. In the completion part, regression model could give us a predicted number which we expected.

4. After generating the result, I stored predicted numbers in another CSV file and it is my first try.

Challenge-AB-Data12-AnswerKey-Dummy.csv 307.42368
3 days ago by siwen feng
add submission details

The initial try comes with a bad result, the root mean squared errors reaches to 307.4. Thus, there are plenty of places to get improvements. Intermediary systems will give more analysis of regrading performances.

Intermediary systems:

The initial design gives us a very clear indication that we can get improvements from two aspects.

1. Select a better algorithm or technique.
2. Keep increasing the quality of the dataset.

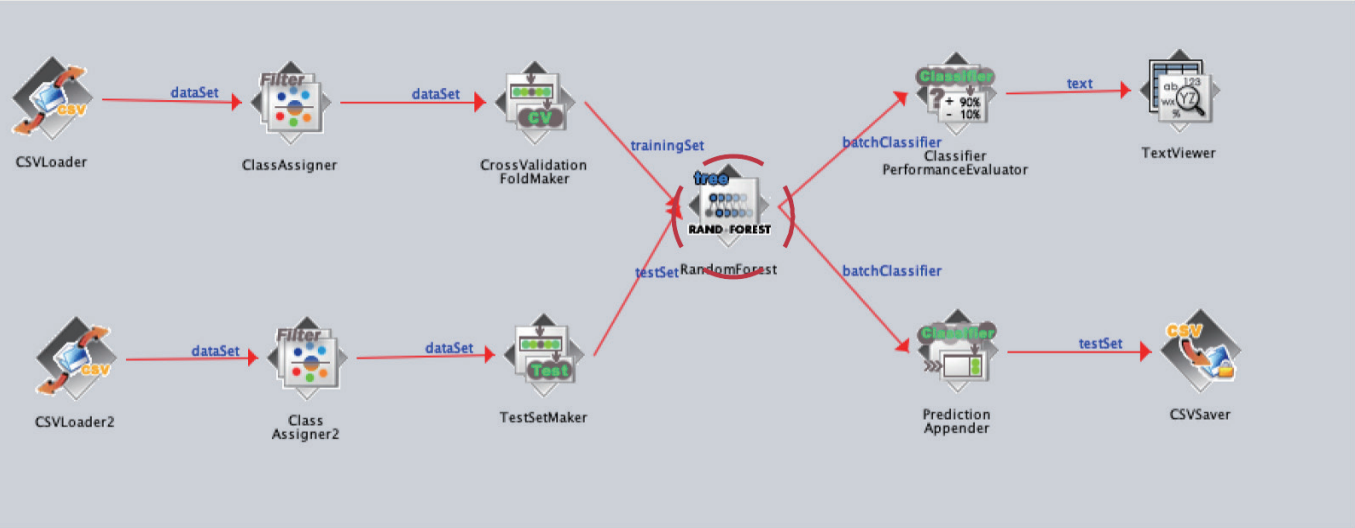
Second system:

In the second system, I swapped linear regression to the random forest. The reason why I changed the algorithm from regression to random forest is that the only attributes we use shape length. However, the shape length has a very big range. Regression is a better option when the variance is low. The shape length has a very high variance which random forest can deal with.

Name: Shape__Length Type: Numeric
Missing: 0 (0%) Distinct: 952 Unique: 942 (98%)

Statistic	Value
Minimum	59.1
Maximum	125082.131
Mean	10178.769
StdDev	14939.326

Pipeline by Weka:



Challenge-AB-Data12-AnswerKey-Dummy.csv 108.25180

3 days ago by siwen feng

random forest

The result is impressive, the error drop from 307.4 to 108.25. Which indicated random forest is a more feasible way to fit this dataset.

Third system:

Next step is to enhance the dataset quality, there is a remaining problem I haven't solved yet. For training set 1. The shape length is missing. In the previous steps. I use "?" replaced instances' shape length. To furtherly decrease the error. I should find an appropriate method to replace missing values.

Because the time is various in training set 1. It is not a good way to replace the missing value with average value or median value. However, training data 2 have matching shape length. Thus, I used data 2 as a training set and data 1 as the testing set to predict shape length. Thus, We give all training set 1 a unique shape length which is better and median number.

Shape__Length	Time
?	30
?	20
?	120
?	360
?	7204
?	486
?	90
?	90
?	90
?	60
?	180
?	40
?	80
?	60
?	75
?	2880
?	30
?	90
?	210
?	330
?	240
?	210
?	60
?	212
?	30

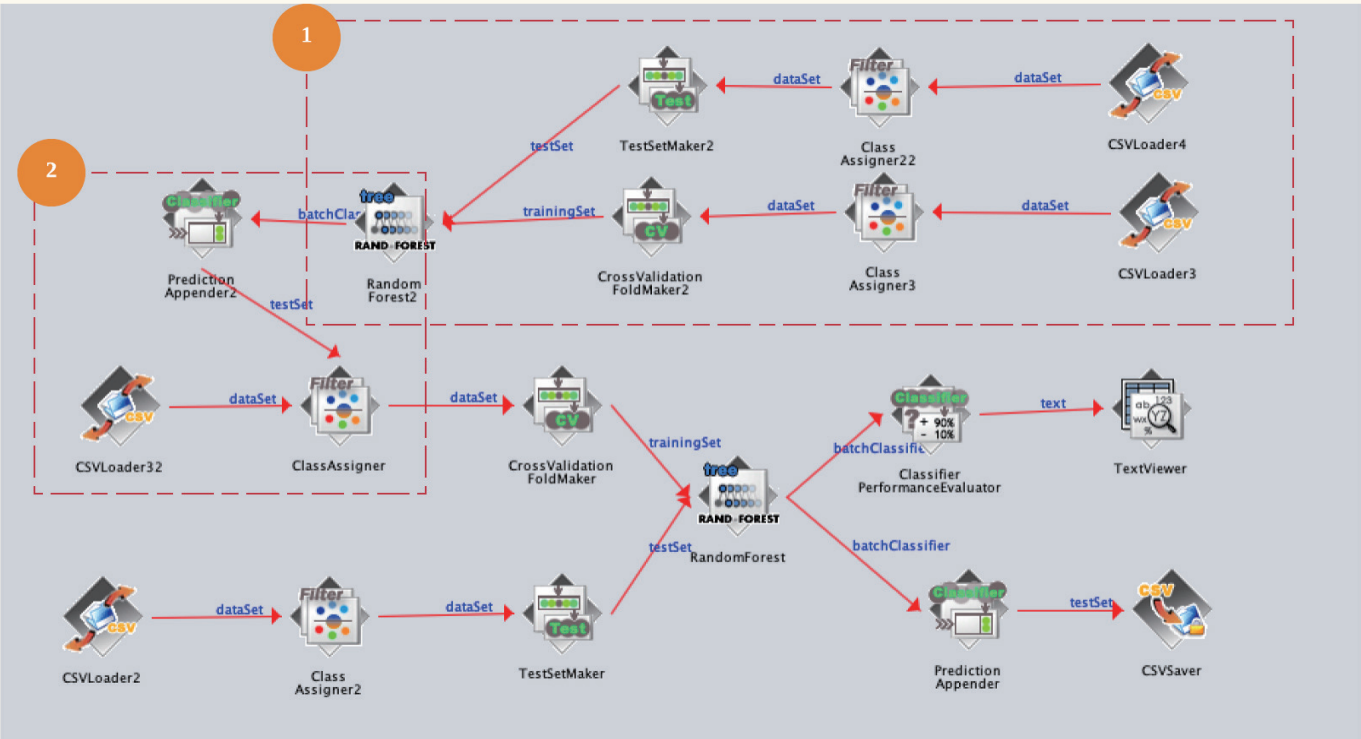
Before

Shape__Length	Time
2359.544	30
2003.295	20
5659.039	120
15186.647	360
70230.086	7204
10436.367	486
4510.014	90
4510.014	90
4510.014	90
3881.216	60
7445.032	180
2606.891	40
3627.371	80
3881.216	60
4252.253	75
45142.551	2880
2359.544	30
4510.014	90
9733.148	210
9217.573	330
9787.153	240
9733.148	210
3881.216	60
9733.148	212
2359.544	30

After

Once data 1 had all shape length, I merged two datasets to increase the amounts of the training set and make the model more accurate. After implementing this step, I have fixed both potential risks already. Thus, this system will be my final system.

Pipeline by Weka:



1. This part is the process of replacing the missing value in dataset 1. I used dataset 2 predicted all possible shape length and applied to dataset one. I also used the Random forest as my regression method. All the predict model aligned with my existing model.
2. we can get a new dataset 1 from the previous step, In this step, I merged two datasets to make sure we get sufficient instances.

Challenge-AB-Data12-AnswerKey-Dummy.csv a day ago by siwen feng random forest with modified dataset	79.78387

The performance increased from 108.25 to 79.78. Thus,It is safe to say that clean data can help us to predict accurately. Modified from both sides contributed to a better and competitive result. The reason why I deployed this system as my final system is that this system provided the best performance and I have used multiple ways to modify and clean dataset. Overall, I identified important features and optimzed them. The result as expected increased gradually. Hence, the final system is reliable to implement.

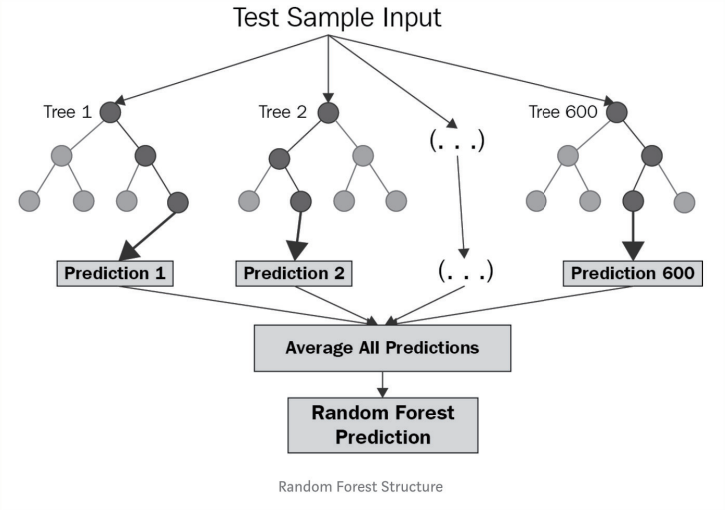
Interesting findings:

- The performance can be increased dramatically with a suitable algorithm. . When I made the second system, I never anticipated I can get 200 increase. Thus, I can extend this experience to many other cases. making a detailed analysis of the dataset and understand all features of datasets could save a lot of time when we asked to do data mining, prediction or other tasks.
- The result is unpredictable unless you did it. At the beginning of this assignment. I firmly believe the difficulty of the track will play an important role and I guess the result definitely will drop if I eliminate this feature. Moreover, the result is the opposite with my personal guess. it tells me making all conclusion based on data instead of personal guess or investigation.

Challenge

Is it easy to interpret:

Random forest is the model I chose for my prediction. The algorithms generated random trees and performing regression. I understand the reason why it comes with the best result. The characteristics of merged dataset indicated that dataset simple and containing a wide range of numbers. In other words. The dataset is low bias and high variance. RF could train each decision tree on a different data sample. RF can give a prediction with higher accuracy because it deals with all instances separately and decreases the influence of variance. ***Hence, RF makes the whole system hard to understand and interpret.***



how random forest work

The initial system is more easy to interpret. The linear regression is an easy model to understand. Cross-validation made sure the data set is not overfitting. training data set help us to build the linear model. in the testing part, we input shape length as an independent value, after calculation with the given model. we can get dependent value completion time. However, its limitation is when high variance in the data the performance is fluctuating.

Ethical consquences:

Go back to business understanding, this technique is helpful to solve the problems I pointed out before. The government can use this technique to develop local tourism, especially like New Zealand which is famous by natural scenes. However, when it comes with ethical consequences we should analyze both pros and cons.

Pros:

- Tourism site can estimate the accessibility of the track and recommend the best track to different trampers. For example, we give some safe and responsible recommendation to tramper with the wheelchair and make sure they can enjoy nature and with low safety risks or worries.
- There are no two identical tramps in this world, the diversity of tracks also representing the dangerous level. This technique could help tramper planning and preparing their trip. Thereby minimizing the risk of injury and maximizing the joy.

Cons:

- Misestimatio has a chance to happen, Thus, it will mislead trampers and make some extra troubles. For the local government, they may give up developing a potential tramp. Thus, we are better to find a balance point not too rely on these techniques.
- Financially, People can use the predicted results as an excuse to destroy the natural environment and gain economic benefits.