REPORT
BIG TASK OF MACHINE LEARNING

BY :
NAME : HANIF FADHLURRAHMAN
ID : 1301174609

INFORMATICS MAJOR
SCHOOL OF COMPUTING
TELKOM UNIVERSITY
BANDUNG
2020

A. Problem Formulation

In this assignment, we given a task where we must do a classification and clustering for the chosen data. The data that I get is *Air_Bnb.csv*, this data is a collection of an online services for place to stay.

B. Data Exploration and Preparation

In this program, there are 4 model of data preparations. 2 models for classification and 2 models for clustering. The models are :

1. For the first model for classification, I use Neighborhood group, Neighbourhood, room type, minimum nights, number of reviews, last review, reviews per month, calculated host listings count, and availability_365 as the columns. And I replace the value in column 'availability_365' by available (if the value inside is not 0) and full (if the value inside is 0). And for the preparation, to split the data I use train_test_split which a method from sklearn library and encoded it using label encoder.

2. For the second model for classification, I use room type, minimum nights, and number of review as the columns. And has the same preparation as number 1

3. For the third model for clustering, I only use latitude and longitude as the columns. The data have been grouped by the neighbors, so that each neighbors has it's own latitude and longitude.

4. for the forth model for clustering, I only use price and number of review as the columns. The data have been grouped by the neighbors, so that each neighbors has it's own price and number of review.

C. Modeling

In modeling I use 2 models of classification and 1 model of clustering. For the classification I used Naïve Bayes, and K-Nearest Neighbors. And for the clustering I used Hierarchical Agglomerative clustering with Complete Link Maximum. For the classification I use Naïve Bayes because I already use it in the previous assignment. And for the K-Nearest Neighbors is because it's a recommendation from the internet. And for the clustering, I use Hierarchical Agglomerative with Complete Link Maximum because I learn it and I know the algorithm very well.

D. Experiment

For the experiment, I 6 experiments in total, 2 for clustering and 4 for classification. There are :

1. The first experiment is for Classification using Naïve Bayes. In this experiment, I use data number 1 (in B section) and using availability as my label, and others as feature.

2. The second experiment is for Classification using K-Nearest Neighbors. In this experiment, for the label and feature is the same as experiment number 1.

3. The third experiment is for classification using Naïve Bayes. In this experiment, I use data number 2 (in B section) and using room type as my label, and others as feature.

4. The forth experiment is for classification using K-Nearest Neighbors. In this experiment, for the label and feature is the same as experiment number 3.

5. the fifth experiment is for Clustering using Agglomerative Clustering with Complete Link Maximum. In this experiment, for the feature I used data number 3 (in B section) and by grouping using the average value of latitude and longitude of each neighbors, for searching which neighbors that has the same neighbors group. And because the neighbors group has 8 value (based on neighbourhood group column), so the clustering is divided into 8 clusters.

6. The sixth experiment is for Clustering using Agglomerative Clustering with Complete Link Maximum, in this experiment for the feature I used data number 4 (in section B) and by grouping using the average of price and number of reviews in each neighbor. For searching which neighbors is worth to stay, neutral, and not worth to stay. This clustering is divided into 3 clusters.

E. Evaluation

Based on the experiment that I have done, the result can be seen as shown :

```
Experiment 1

Naive Bayes Classification Accuracy : 0.6948568969227459
KNearestN Classification Accuracy : 0.8050355067785668
```

(fig. E.1)

1. The result of the first experiment and second experiment can be seen at fig E.1.

```
Experiment 2

Naive Bayes Classification Accuracy : 0.6998063266623629
KNearestN Classification Accuracy : 0.789541639767592
```

(fig E.2)

2. the result of the third and forth experiment can be seen at fig E.2

```
Clustering latitude and longitude to find which neighbourhood stays in the same neighbourhood_Group

cluster  1  =  [2, 20, 13, 17, 29, 49, 14, 101, 96, 86, 115, 52, 21, 97, 92, 55, 112, 100, 30, 122, 121, 90, 91, 134, 87, 85, 98, 18, 95, 32, 118, 135, 53, 94]
cluster  2  =  [0, 12, 3, 46, 67, 36, 48, 40, 8, 66, 9, 15, 56, 60, 71, 70, 47, 58, 61, 25]
cluster  3  =  [1, 22, 10, 116, 73, 68, 42, 69, 26, 110, 39, 62, 63, 51, 107, 117, 72, 132, 88, 119, 108, 27, 74, 106, 33, 120, 111, 130, 89, 79, 57]
cluster  4  =  [4, 5, 11, 7, 23, 16, 34, 6, 103, 50, 54, 123, 80, 59, 81, 104, 93, 77, 35, 102, 38, 31, 105, 84, 76, 78]
cluster  5  =  [19, 24, 125, 41, 44, 114, 45, 129, 43, 131, 113, 28]
cluster  6  =  [75, 109, 99]
cluster  7  =  [64, 82, 124, 127, 65, 126, 128, 83]
cluster  8  =  [37, 133]
```

(fig E.3.1)

```
data Cluster 1
                      latitude  longitude
neighbourhood
Adlershof             52.436217  13.543875
Albrechtstr.          52.456513  13.336666
Alexanderplatz        52.522516  13.404078
Allende-Viertel       52.447435  13.597893
Alt  Treptow          52.490455  13.449758
...                         ...        ...
Wilhelmstadt          52.527162  13.190948
Zehlendorf  Nord      52.445922  13.256900
Zehlendorf  Südwest   52.425791  13.185577
nördliche Luisenstadt 52.501865  13.427402
südliche Luisenstadt  52.496617  13.435156

[136 rows x 2 columns]
```

(fig E.3.2)

3. The result of the fifth experiment can be seen at fig E.3.1. in this result, each cluster has an array of numbers. The numbers are representing the name of the neighbors (see fig E.3.2) for example, number 0 is representing Adlershof, number 1 is representing Albrechtstr, etc. And in each cluster the neighbors are in the same group.

```
Clustering price and number of review to find which neighbourhood is worth to stay

cluster  1  =  [2, 24, 30, 5, 49, 6, 99, 51, 41, 100, 10, 57, 65, 56, 119, 39, 79, 98, 114, 27, 9, 66, 53, 13, 106, 18, 71, 133, 97
, 73, 95, 110, 104, 96, 67, 44, 42, 115, 108, 74, 109, 118, 135, 130, 134, 132, 111]
cluster  2  =  [0, 8, 3, 38, 22, 1, 23, 28, 103, 15, 12, 45, 48, 11, 61, 55, 14, 78, 16, 35, 127, 20, 47, 86, 124, 4, 69, 7, 82, 32
, 58, 62, 77, 17, 46, 64, 34, 37, 68, 52, 113, 87, 54, 60, 21, 75, 29, 94, 90, 84, 59, 80, 112, 19, 63, 125, 89, 117, 72, 76, 123,
129, 88, 91, 81, 33, 122, 107, 40, 50, 128, 36, 131, 116, 102, 26, 92, 25, 101, 121, 126, 85, 83, 93, 105, 70]
cluster  3  =  [31, 43, 120]
```

(fig 4.1)

```
data Cluster 2
                          price  number_of_reviews
neighbourhood
Adlershof             54.142857           8.666667
Albrechtstr.          44.117647          15.482353
Alexanderplatz        85.372017          35.742950
Allende-Viertel       25.000000          14.000000
Alt  Treptow          54.136691          15.093525
...                         ...                ...
Wilhelmstadt          45.875000           7.083333
Zehlendorf  Nord      64.150943          19.735849
Zehlendorf  Südwest   67.807692          18.000000
nördliche Luisenstadt 63.179245          21.955189
südliche Luisenstadt  58.432836          19.535448
```

(fig 4.2)

4. the result of the sixth experiment can be seen at fig E.4.1. in this result each cluster has an array of numbers. The numbers are representing the name of the neighbors (see fig 4.2) for example, number 0 is representing Adlershof, number 1 is representing Albrechtstr, etc. And in clusters, each cluster represent the value of worth, not worth it and neutral (if you have money, then go, if not don't). And cluster 1 is representing as value Worth it, cluster 2 is representing as value Neutral, and cluster 3 is representing as value Not Worth it. For example, Adlershof is neutral to visit. Alexanderplatz is worth to visit, etc

F. Conclusions

Using Naïve Bayes and K-Nearest Neighbors classification, it seems that the K-Nearest Neighbors classification is better than Naïve Bayes classification. As we can see on chapter E, the value of accuracy is bigger if compared it to Naïve Bayes accuracy. And for the Clustering, we can know which neighbors that are in the same group with other neighbors. And we can also know which neighbors is more worth to visit during holiday.