CASE BASED 1
MACHINE LEARNING

DISUSUN OLEH :

YUNOLVA ANIS RAMAZIYAH (1301204096)

KELAS : IF-44-09

KODE DOSEN : IZA

Saya mengerjakan tugas ini dengan cara yang tidak melanggar aturan perkuliahan dan kode etik akademisi

BAB I Penggunaan Data

Data yang digunakan merupakan data arrhythmia. Di mana arrhythmia merupakan sebuah penyakit irama jantung yang tidak teratur. Sehingga di dalam data tersebut terdapat kolom yang menunjukkan data – data terkait dengan penyakit arrhythmia. Kemudian arrhythmia terdiri dari 452 instance, dengan 279 attribute.

Daftar attribute yang dimiliki adalah :

1. Age
2. Sex
3. Height
4. Weight
5. QRS duration: Average of QRS duration in msec., linear
6. P-R interval: Average duration between onset of P and Q waves in msec., linear
7. Q-T interval: Average duration between onset of Q and off set of T waves in msec., linear
8. T interval: Average duration of T wave in msec., linear
9. P interval: Average duration of P wave in msec., linear
10. QRS
11. T
12. P
13. QRST
14. J
15. Heart rate: Number of heart beats per minute, linear
16. Q wave
17. R wave
18. S wave
19. R' wave, small peak just after R
20. S' wave
21. Number of intrinsic deflections, linear
22. Existence of ragged R wave, nominal
23. Existence of diphasic derivation of R wave, nominal
24. Existence of ragged P wave, nominal
25. Existence of diphasic derivation of P wave, nominal
26. Existence of ragged T wave, nominal
27. Existence of diphasic derivation of T wave, nominal

28 .. 39 (similar to 16 .. 27 of channel DI)

Of channels DIII:

40 .. 51

Of channel AVR:

52 . .63

Of channel AVL:

64 .. 75

Of channel AVF:

76 . .87

Of channel V1:

88 … 99

Of channel V2:

100 … 111

Of channel V3:

112 … 123

Of channel V4:

124 … 135

Of channel V5:

136 ...147

Of channel V6:

148 ...159

Of channel DI:

160 . JJ wave, linear
161 . Q wave, linear
162 . R wave, linear
163 . S wave, linear
164 . R’ wave, linear
165 . S’ wave, linear
166 . P wave, linear
167 . T wave, linear
168 . QRSA , Sum of areas of all segments divided by 10, (Area= width * height / 2), linear
169 . QRSTA = QRSA + 0.5 * width of T wave * 0.1 * height of T wave. (If T is diphasic then the bigger segment is considered), linear

Of channel DII:

170 .. 179

Of channel DIII:

180 .. 189

Of channel AVR:

190 … 199

Of channel AVL:

200 … 209

Of channel AVF:

210 … 219

Of channel V1:

220 … 229

Of channel V2:

230 … 239

Of channel V3:

240 … 249

Of channel V4:

250 … 259

Of channel V5:

260 … 269

Of channel V6:

270 .. 279

Kemudian dilakukan pengecekan data dan didapatkan kolom yang tidak sesuai karena beirisi data Nan. Sehingga kolom tersebut akan di drop.

```
[155] summary(df)
       X1              X2              X3              X4
 Min.   : 0.00   Min.   :0.0000   Min.   :105.0   Min.   :  6.00
 1st Qu.:36.00   1st Qu.:0.0000   1st Qu.:160.0   1st Qu.: 59.00
 Median :47.00   Median :1.0000   Median :164.0   Median : 68.00
 Mean   :46.47   Mean   :0.5509   Mean   :166.2   Mean   : 68.17
 3rd Qu.:58.00   3rd Qu.:1.0000   3rd Qu.:170.0   3rd Qu.: 79.00
 Max.   :83.00   Max.   :1.0000   Max.   :780.0   Max.   :176.00
       X5              X6              X7              X8              X9
 Min.   : 55.00   Min.   :  0.0   Min.   :232.0   Min.   :108.0   Min.   :  0
 1st Qu.: 80.00   1st Qu.:142.0   1st Qu.:350.0   1st Qu.:148.0   1st Qu.: 79
 Median : 86.00   Median :157.0   Median :367.0   Median :162.0   Median : 91
 Mean   : 88.92   Mean   :155.2   Mean   :367.2   Mean   :169.9   Mean   : 90
 3rd Qu.: 94.00   3rd Qu.:175.0   3rd Qu.:384.0   3rd Qu.:179.0   3rd Qu.:102
 Max.   :188.00   Max.   :524.0   Max.   :509.0   Max.   :381.0   Max.   :205
      X10             X11             X12             X13
 Min.   :-172.00   Min.   :-177.00   Min.   :-170.00   Min.   :-135.00
 1st Qu.:   3.75   1st Qu.:  14.00   1st Qu.:  41.00   1st Qu.:  12.00
 Median :  40.00   Median :  41.00   Median :  54.50   Median :  40.00
 Mean   :  33.68   Mean   :  36.15   Mean   :  48.91   Mean   :  36.72
 3rd Qu.:  66.00   3rd Qu.:  63.00   3rd Qu.:  64.00   3rd Qu.:  62.00
 Max.   : 169.00   Max.   : 179.00   Max.   : 176.00   Max.   : 166.00
      X15             X16             X17             X18
 Min.   : 44.00   Min.   : 0.000   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 65.00   1st Qu.: 0.000   1st Qu.: 40.00   1st Qu.: 0.00
 Median : 72.00   Median : 0.000   Median : 48.00   Median :20.00
 Mean   : 74.46   Mean   : 5.628   Mean   : 51.63   Mean   :20.92
 3rd Qu.: 81.00   3rd Qu.:12.000   3rd Qu.: 60.00   3rd Qu.:36.00
 Max.   :163.00   Max.   :88.000   Max.   :156.00   Max.   :88.00
```

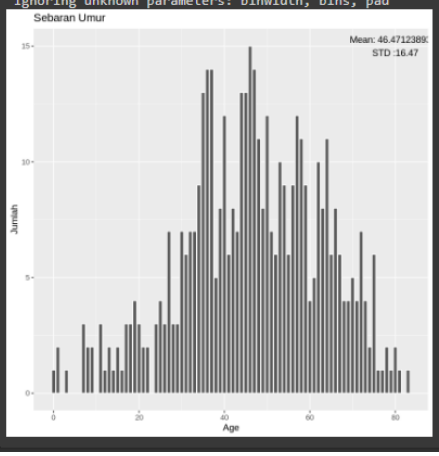
```
[154] df <- data.frame(do.call("rbind", strsplit(as.character(df$V1), ",", fixed = TRUE)))
      df[df == "?"] <- NA
      df <- subset(df, select=-X14)
      for(i in 1:ncol(df)) {
        df[ , i] <- as.numeric(df[ , i])
        df[ , i][is.na(df[ , i])] <- mean(df[ , i], na.rm = TRUE)
      }
```

Lalu berikut merupakan visualisasi dari beberapa kolom yang ada di dalam database arrhythmia.

1. Kolom Umur
   Berdasarkan visualisasi tersebut umur pasien yang menderita penyakit arrhythmia bervariasi, mulai dari 0 tahun sampai dengan lebih dari 80 tahun.

```
(p1<-ggplot(df,aes(`X1`))+
    geom_histogram(bins = 15,binwidth = 2, stat = 'count', color="white")+
    labs(x="Age",y="Jumlah",title = "Sebaran Umur")+
    annotate("text",x= 80,y=15,label = paste0("Mean: ",as.character(round(mean(df$X1),10)),"\nSTD :",as.character(round(sd(df$X1),2) ))))
```

```
Warning message:
"Ignoring unknown parameters: binwidth, bins, pad"
```

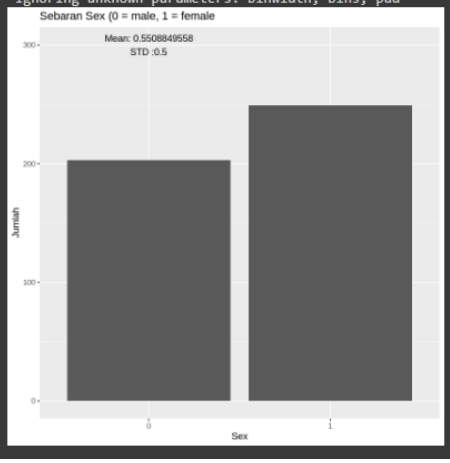


2. Kolom Sex

   Pada kolom ini hanya terbagi menjadi dua label saja yakni 0 dan 1, di mana 0 melambangkan male dan 1 melambangkan female. Berdasarkan visualisasi data tersebut pasien yang paling banyak menderita

```
[167] s2 <- summary(df$X2)
      (p2<-ggplot(df,aes(x=factor(`X2`)))+
          geom_histogram(stat = 'count')+
          labs(x="Sex",y="Jumlah",title = "Sebaran Sex (0 = male, 1 = female)")+
          annotate("text",x= 1,y=300,label = paste0("Mean: ",as.character(round(mean(df$X2),2)),"\nSTD :",as.character(round(sd(df$X2),2) ))))
```

```
Warning message:
"Ignoring unknown parameters: binwidth, bins, pad"
```

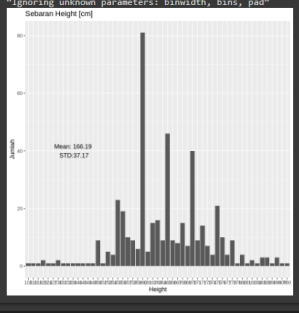


3. Kolom Height

```
library(scales)

s3 <- summary(df$X3)
(p3<-ggplot(df,aes(x=factor(`X3`)))+
    geom_histogram(stat = 'count')+
    labs(x="Height",y="Jumlah",title = "Sebaran Height [cm]")+
    annotate("text",x= 10,y=40,label = paste0("Mean: ",as.character(round(mean(df$X3),2)),"\n STD:",as.character(round(sd(df$X3),2) ))))
```

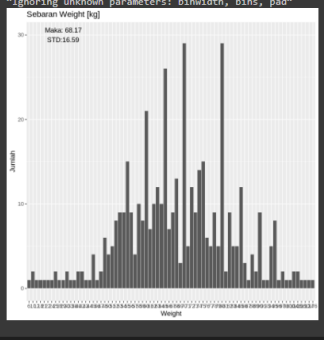
```
Warning message:
"Ignoring unknown parameters: binwidth, bins, pad"
```



4. Kolom Weight

```
[16] s4 <- summary(df$X4)
     (p4<-ggplot(df,aes(x=factor(`X4`)))+
         geom_histogram(stat = 'count')+
         labs(x="Weight",y="Jumlah",title = "Sebaran Weight [kg]")+
         annotate("text",x= 10,y=30,label = paste0("Maka: ",as.character(round(mean(df$X4),2)),"\nSTD:",as.character(round(sd(df$X4),2) ))))
```

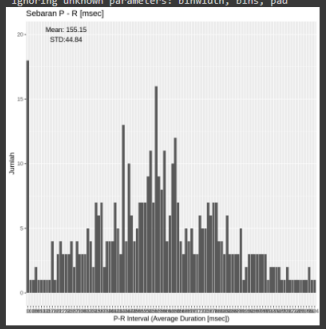
```
Warning message:
"Ignoring unknown parameters: binwidth, bins, pad"
```



5. Kolom P – R Interval

```
s6 <- summary(df$X6)
(p6<-ggplot(df,aes(x=factor(`X6`)))+
    geom_histogram(stat = 'count')+
    labs(x="P-R Interval (Average Duration [msec])",y="Jumlah",title = "Sebaran P - R [msec]")+
    annotate("text",x= 15,y=20,label = paste0("Mean: ",as.character(round(mean(df$X6),2)),"\nSTD:",as.character(round(sd(df$X6),2) ))))
```

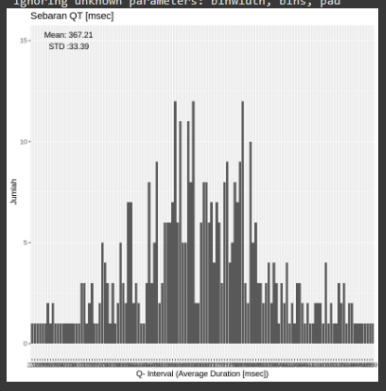
```
Warning message:
"Ignoring unknown parameters: binwidth, bins, pad"
```



6. Kolom Q-T Interval

```
s7 <- summary(df$X7)
(p7<-ggplot(df,aes(x=factor(`X7`)))+
    geom_histogram(stat = 'count')+
    labs(x="Q- Interval (Average Duration [msec])",y="Jumlah",title = "Sebaran QT [msec]")
    + annotate("text",x= 15,y=15,label = paste0("Mean: ",as.character(round(mean(df$X7),2)),"\nSTD :",as.character(round(sd(df$X7),2) ))))
```

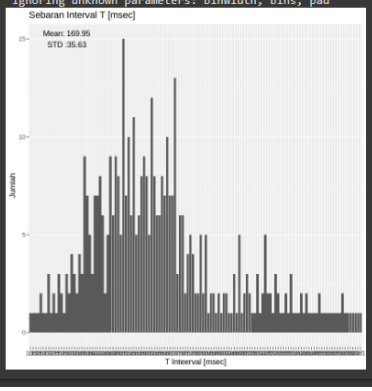
```
Warning message:
“Ignoring unknown parameters: binwidth, bins, pad”
```



7. Kolom T Interval

```
s8 <- summary(df$X8)
(p8<-ggplot(df,aes(x=factor(`X8`)))+
    geom_histogram(stat = 'count')+
    labs(x="T Inteerval [msec]",y="Jumlah",title = "Sebaran Interval T [msec]")+
    annotate("text",x= 15,y=15,label = paste0("Mean: ",as.character(round(mean(df$X8),2)),"\nSTD :",as.character(round(sd(df$X8),2) ))))
```

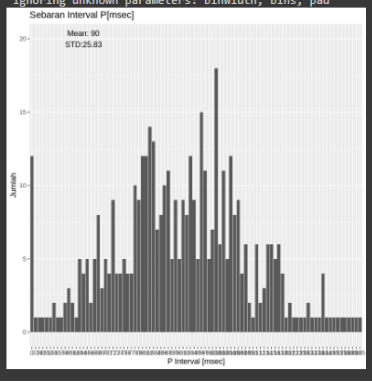
```
Warning message:
“Ignoring unknown parameters: binwidth, bins, pad”
```



8. Kolom P Interval

```
[21] s9 <- summary(df$X9)
     (p9<-ggplot(df,aes(x=factor(`X9`)))+
         geom_histogram(stat = 'count')+
         labs(x="P Interval [msec]",y="Jumlah",title = "Sebaran Interval P[msec]")+
         annotate("text",x= 15,y=20,label = paste0("Mean: ",as.character(round(mean(df$X9),2)),"\nSTD:",as.character(round(sd(df$X9),2) ))))
```

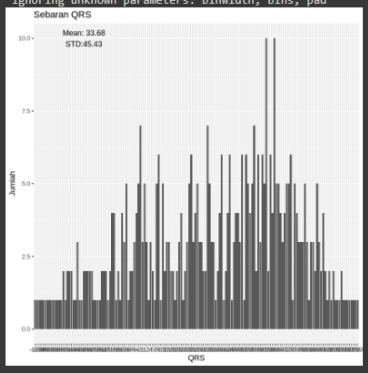
```
Warning message:
“Ignoring unknown parameters: binwidth, bins, pad”
```



9. Kolom QRS

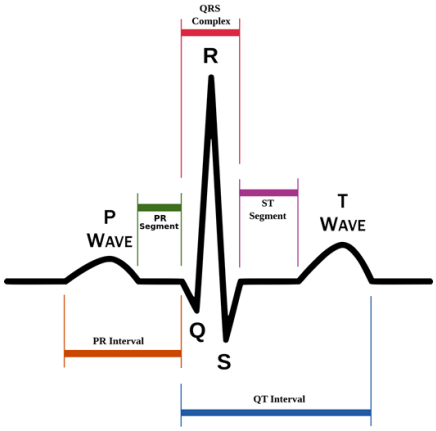
```
[22] s10 <- summary(df$X10)
     (p10<-ggplot(df,aes(x=factor(`X10`)))+
         geom_histogram(stat = 'count')+
         labs(x="QRS",y="Jumlah",title = "Sebaran QRS")+
         annotate("text",x= 25,y=10,label = paste0("Mean: ",as.character(round(mean(df$X10),2)),"\nSTD:",as.character(round(sd(df$X10),2) ))))
```

```
Warning message:
“Ignoring unknown parameters: binwidth, bins, pad”
```



Bab II Pra Pemrosesan

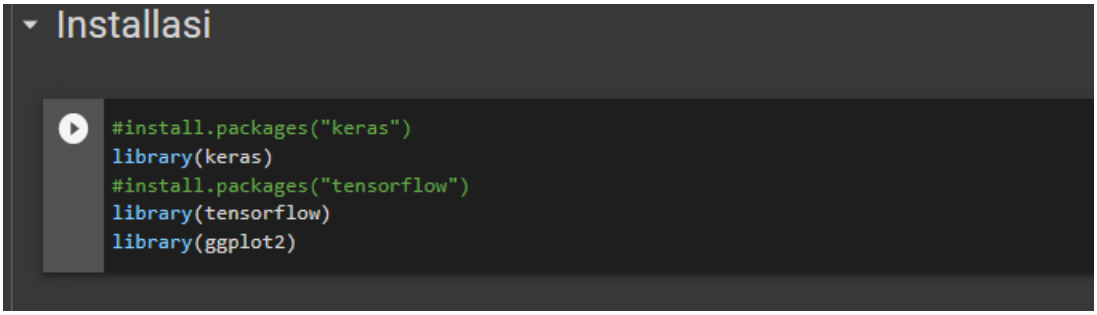
Sebelum menuju ke pra pemrosesan, perlu diketahui konsep dari Arrhythmia itu sendiri. Di dalam Arrhythmia, terdapat beberapa kolom yang sukar, seperti P, Q, R, S, dan T. Kolom – kolom tersebut mengindikasikan rekam dari jatung pasien.



P wave merupakan depolarisasi dari atrium, Q wave merupakan aktivasi dari anterioseptal ventricular myocardium, R wave merupakan depolarisasi dari ventricular myocardium, S wave adalah aktivasi dati posteriobasal ventricles, dan T wave adalah rapid ventricular repolarization. Pada kasus Arrhythmia diputuskan untuk menggunkan MLP yang merupakan bagian dari Supervised Learning Method.

Bab III Penerapan Algoritma

Installasi



Melakukan penginstalan terhadap beberapa library yang akan digunakan kedepannya seperti Tensorflow dan keras.

Proses

Melakukan import dataset dan juga penghapusan kolom yang terdapat nilai NaN, dan kolom tersebut adalah kolom ke X-14.

```
[4] path <- "arrhythmia.data"
    df <- read.table(path)
    df <- data.frame(do.call("rbind", strsplit(as.character(df$V1), ",", fixed = TRUE)))
    df[df == "?"] <- NA
    df <- subset(df, select=-X14)
    for(i in 1:ncol(df)) {
      df[ , i] <- as.numeric(df[ , i])
      df[ , i][is.na(df[ , i])] <- mean(df[ , i], na.rm = TRUE)
    }
```

Menampilkan summary dari dataset, pada summary tersebut dapat terlihat bahawa kolom X-14 sudah tidak ada.

```
summary(df)
```

```
       X1              X2              X3              X4        
 Min.   : 0.00   Min.   :0.0000   Min.   :105.0   Min.   :  6.00  
 1st Qu.:36.00   1st Qu.:0.0000   1st Qu.:160.0   1st Qu.: 59.00  
 Median :47.00   Median :1.0000   Median :164.0   Median : 68.00  
 Mean   :46.47   Mean   :0.5509   Mean   :166.2   Mean   : 68.17  
 3rd Qu.:58.00   3rd Qu.:1.0000   3rd Qu.:170.0   3rd Qu.: 79.00  
 Max.   :83.00   Max.   :1.0000   Max.   :780.0   Max.   :176.00  
       X5              X6              X7              X8              X9        
 Min.   : 55.00   Min.   :  0.0   Min.   :232.0   Min.   :108.0   Min.   :  0  
 1st Qu.: 80.00   1st Qu.:142.0   1st Qu.:350.0   1st Qu.:148.0   1st Qu.: 79  
 Median : 86.00   Median :157.0   Median :367.0   Median :162.0   Median : 91  
 Mean   : 88.92   Mean   :155.2   Mean   :367.2   Mean   :169.9   Mean   : 90  
 3rd Qu.: 94.00   3rd Qu.:175.0   3rd Qu.:384.0   3rd Qu.:179.0   3rd Qu.:102  
 Max.   :188.00   Max.   :524.0   Max.   :509.0   Max.   :381.0   Max.   :205  
      X10             X11             X12             X13        
 Min.   :-172.00   Min.   :-177.00   Min.   :-170.00   Min.   :-135.00  
 1st Qu.:   3.75   1st Qu.:  14.00   1st Qu.:  41.00   1st Qu.:  12.00  
 Median :  40.00   Median :  41.00   Median :  54.50   Median :  40.00  
 Mean   :  33.68   Mean   :  36.15   Mean   :  48.91   Mean   :  36.72  
 3rd Qu.:  66.00   3rd Qu.:  63.00   3rd Qu.:  64.00   3rd Qu.:  62.00  
 Max.   : 169.00   Max.   : 179.00   Max.   : 176.00   Max.   : 166.00  
      X15             X16             X17             X18        
 Min.   : 44.00   Min.   : 0.000   Min.   :  0.00   Min.   : 0.00  
 1st Qu.: 65.00   1st Qu.: 0.000   1st Qu.: 40.00   1st Qu.: 0.00  
 Median : 72.00   Median : 0.000   Median : 48.00   Median :20.00  
 Mean   : 74.46   Mean   : 5.628   Mean   : 51.63   Mean   :20.92  
 3rd Qu.: 81.00   3rd Qu.:12.000   3rd Qu.: 60.00   3rd Qu.:36.00  
 Max.   :163.00   Max.   :88.000   Max.   :156.00   Max.   :88.00  
      X19             X20             X21             X22        
 Min.   : 0.0000   Min.   :0   Min.   :  0.00   Min.   :0.000000  
 1st Qu.: 0.0000   1st Qu.:0   1st Qu.: 24.00   1st Qu.:0.000000  
 Median : 0.0000   Median :0   Median : 28.00   Median :0.000000  
 Mean   : 0.1416   Mean   :0   Mean   : 30.04   Mean   :0.002212  
 3rd Qu.: 0.0000   3rd Qu.:0   3rd Qu.: 36.00   3rd Qu.:0.000000  
 Max.   :24.0000   Max.   :0   Max.   :100.00   Max.   :1.000000  
      X23             X24             X25             X26        
 Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000  
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000  
 Median :0.00000   Median :0.00000   Median :0.000000   Median :0.000000  
 Mean   :0.01106   Mean   :0.01106   Mean   :0.004425   Mean   :0.004425  
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
```

Setelah itu menyimpan df ke dalam variable baru yakni df2 dan mengubah tipe datanya menjadi numeric.

```
df2 <- df
df2[,279] <- as.numeric(df2[,279]) -1
```

Kemudian melakukan split df2 menjadi data training dan test serta data training target dan juga data training test target.

```
df2.training <- df2[ind==1, 1:278]
df2.test <- df2[ind==2, 1:278]

[113] df2.trainingtarget <- df2[ind==1, 279]
      df2.testtarget <- df2[ind==2, 279]
```

Lalu dilanjut dengan mengencode data target

```
df2.trainLabels <- to_categorical(df2.trainingtarget)

df2.testLabels <- to_categorical(df2.testtarget)
```

Kemudian memanggil keras_model_sequencial agar dapat membuat sequential model.

```
model <- keras_model_sequential()
```

Melakukan pengetesan dengan layer yang ada seperti layer dense dan dropout dan dengan di dalamnya menggukan fungsi sigmoid dan softmax. Adanya penggunaan sigmoid ini menjadikan gradien lebih fleksibel dan jangkauannya menjadi lebih luas dibanding dengan penggunaan fungsi lain yang sejenis yakni Relu.

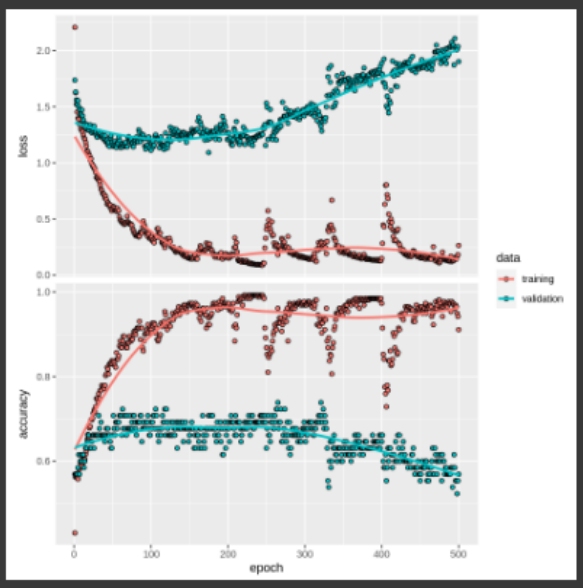
```
model %>% compile(
  loss = 'categorical_crossentropy',
  optimizer = 'adam',
  metrics = 'accuracy'
)
```

Dilanjutkan dengan pembentukan model di mana di dalam model tersebut, loss akan dihitung dengan bantuan fungsi ‘categorical_crossentropy’, dan di tuning dengan menggunakan ‘adam’ optimizer, lalu yang terakhir metric yang digunakan adalah metrics ‘accurancy’.

```
[30] history <- model %>% fit(
       df2.training,
       df2.trainLabels,
       epochs = 500,
       batch_size = 5,
       validation_split = 0.2
     )
```

Setelah itu ditentukan seberapa banyak epochs dan juga batch sizenya, sehingga didapatlah plot sebagai berikut :

```
[31] plot(history)
```



Kemudian, untuk score akhir yang didapat adalah accuracy sebesar 67% dan loss sebesar 1%.

```
[35] score <- model %>% evaluate(df2.test, df2.testLabels, batch_size = 128)
     print(score)

         loss  accuracy
    1.3461293 0.6716418
```

BAB IV Evaluasi

Berdasarkan hasil score yang telah saya dapatkan, menurut saya hasil tersebut belumlah sesuai. Hal tersebut dikarenakan dari banyaknya rujukan yang telah saya baca, MLP memperoleh nilai tertinggi disbanding beberapa kategori lain seperi CNN dan ANN. Dan saya menyadari di dalam laporan saya terdapat banyak kekurangan seperti tidak adanya normalisasi data, yang mungkin menyebabkan nilai accuracy tersebut drop.

Link Dokumentasi :

Video : https://youtu.be/k7nLn4J1_ew

Source Code : https://colab.research.google.com/drive/1xXspt_mz2s_lk6XiB1h_nzacoQpdKafM#scrollTo=RvRvC3AfYSAH

Link Referensi :

https://ceur-ws.org/Vol-2563/aics_33.pdf