Q9:

Text

Description automatically generatedNB is the model that can be done by hand.

Assumed variables are independent, so some variables are omitted.

As shown in figure !!, there are four predictor variables for this model based on the important variable selected in Q!!, namely Sunshine, Windy, HighPressure and Cold. The values of “Yes” and “No” are given. To judge “Yes” or “No”, there are threshold values. When the value is beyond them, it is considered “Yes” else “No.” For example, when the Pressure level is beyond 1013, which is standard average level, the HighPressure value is “Yes”. Moreover, it is known that the probability of being CloudyTomorrow is 210 / ( 326 + 210 ) = .39 and being not “CloudyTomorrow” is 1- .39 = .61.

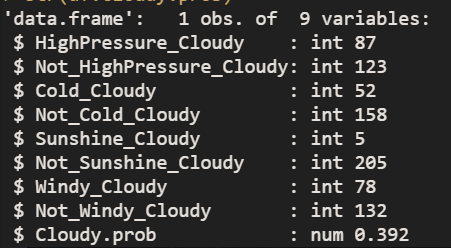


Figure !! is the count of all variables group by “yes” and “no” in which tomorrow is being Cloudy. With this graph, we can calculate P( Variables | CloudyTomorrow ).

For example, P(High Pressure | CloudyTomorrow ) = number of high pressure / number of being CloudyTomorrow = 87 / 210 = .414

Text

Description automatically generatedP(Not high Pressure | CloudyTomorrow ) = number of high pressure / number of being CloudyTomorrow = 123 / 210 = .586



With knowing the steps of calculating the probability of each variable given that CloudyTomorrow is Yes, the results are shown in figure !!.

Likewise, Figure !! is the count of all variables group by “yes” and “no” in which tomorrow is being NOT Cloudy. With this graph, we can calculate P( Variables | Not CloudyTomorrow ).

For example, P(High Pressure | Not CloudyTomorrow ) = number of high pressure / number of being CloudyTomorrow = 180 / 326 = .552

P(Not high Pressure | Not CloudyTomorrow ) = number of high pressure / number of being Not CloudyTomorrow = 146 / 326 = .448

With knowing the steps of calculating the probability of each variable given that CloudyTomorrow is Yes, the results are shown in figure !!.

To test the model, the testing dataset is shown in figure !!







Text

Description automatically generated



Accuracy = 5+3 / (5+3+2) = .8

What it can tell :

The reason of high accuracy is the number of testing row is not sufficient

Randomise and lucky training set