Practical - 3

AIM: Demo of Logistic Regression

THEORY:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on

prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or

more existing independent variables. For example, a logistic regression could be used to predict whether a

political candidate will win or lose an election or whether a high school student will be admitted or not to a

particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college

acceptance, the logistic function could consider factors such as the student's grade point average, SAT score

and number of extracurricular activities. Based on historical data about earlier outcomes involving the same

input criteria, it then scores new cases on their probability of falling into one of two outcome categories.

Logistic regression has become an important tool in the discipline of machine learning. It allows algorithms

used in machine learning applications to classify incoming data based on historical data. As additional relevant

data comes in, the algorithms get better at predicting classifications within data sets.

Logistic regression can also play a role in data preparation activities by allowing data sets to be put into

specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the

information for analysis.

CODE AND OUTPUT:

```
> x=read.csv("C:/Users/VAISHAKH/OneDrive/Desktop/DS ASSIGNMENT AND JOURNAL/weather1.csv")
> x
    outlook temp humidity windy play.golf
1     rainy  hot     high FALSE        no
2     rainy  hot     high  TRUE        no
3  overcast  hot     high FALSE       yes
4     sunny mild     high FALSE       yes
5     sunny cool   normal FALSE       yes
6     sunny cool   normal  TRUE        no
7  overcast cool   normal  TRUE       yes
8     rainy mild     high FALSE       yes
9     rainy cool   normal FALSE       yes
10    sunny mild   normal FALSE       yes
11    rainy mild   normal  TRUE       yes
12 overcast mild     high  TRUE       yes
13 overcast  hot   normal FALSE       yes
14    sunny mild     high  TRUE        no
> x$humidity=ifelse(test=x$humidity=="high",yes=1,no=0)
> x
    outlook temp humidity windy play.golf
1     rainy  hot        1 FALSE        no
2     rainy  hot        1  TRUE        no
3  overcast  hot        1 FALSE       yes
4     sunny mild        1 FALSE       yes
5     sunny cool        0 FALSE       yes
6     sunny cool        0  TRUE        no
7  overcast cool        0  TRUE       yes
8     rainy mild        1 FALSE       yes
9     rainy cool        0 FALSE       yes
10    sunny mild        0 FALSE       yes
11    rainy mild        0  TRUE       yes
12 overcast mild        1  TRUE       yes
13 overcast  hot        0 FALSE       yes
14    sunny mild        1  TRUE        no
> x$play=ifelse(test=x$play=="yes",yes=1,no=0)
> x
```

```
    outlook temp humidity windy play.golf play
1     rainy  hot        1 FALSE        no    0
2     rainy  hot        1  TRUE        no    0
3  overcast  hot        1 FALSE       yes    1
4     sunny mild        1 FALSE       yes    1
5     sunny cool        0 FALSE       yes    1
6     sunny cool        0  TRUE        no    0
7  overcast cool        0  TRUE       yes    1
8     rainy mild        1 FALSE       yes    1
9     rainy cool        0 FALSE       yes    1
10    sunny mild        0 FALSE       yes    1
11    rainy mild        0  TRUE       yes    1
12 overcast mild        1  TRUE       yes    1
13 overcast  hot        0 FALSE       yes    1
14    sunny mild        1  TRUE        no    0
> x$windy=ifelse(test=x$windy=="FALSE",yes=0,no=1)
> x
    outlook temp humidity windy play.golf play
1     rainy  hot        1     0        no    0
2     rainy  hot        1     1        no    0
3  overcast  hot        1     0       yes    1
4     sunny mild        1     0       yes    1
5     sunny cool        0     0       yes    1
6     sunny cool        0     1        no    0
7  overcast cool        0     1       yes    1
8     rainy mild        1     0       yes    1
9     rainy cool        0     0       yes    1
10    sunny mild        0     0       yes    1
11    rainy mild        0     1       yes    1
12 overcast mild        1     1       yes    1
13 overcast  hot        0     0       yes    1
14    sunny mild        1     1        no    0
> #partitioning dataset
> s=sample(nrow(x),.7*nrow(x))
> x_tr=x[s,]
> x_test=x[-s]
> nrow(x)
[1] 14
> nrow(x_tr)
[1] 9
> nrow(x_test)
[1] 14
>
```

```
> lmod=glm(play~windy,data=x_tr,family=binomial,control=list(maxit=100)
+ lmod
Error: unexpected symbol in:
"lmod=glm(play~windy,data=x_tr,family=binomial,control=list(maxit=100)
lmod"
> lmod=glm(play~windy,data=x_tr,family=binomial,control=list(maxit=100))
> lmod

Call:  glm(formula = play ~ windy, family = binomial, data = x_tr, control = list(maxit = 100))

Coefficients:
(Intercept)        windy
      20.57       -21.66

Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
Null Deviance:      11.46
Residual Deviance: 4.499        AIC: 8.499
> summary(lmod)

Call:
glm(formula = play ~ windy, family = binomial, data = x_tr, control = list(maxit = 100))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.75853  -0.75853   0.00005   0.00005   1.66511

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    20.57    7929.26   0.003    0.998
windy         -21.66    7929.26  -0.003    0.998

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11.4573  on 8  degrees of freedom
Residual deviance:  4.4987  on 7  degrees of freedom
AIC: 8.4987

Number of Fisher Scoring iterations: 19

> lmod=glm(play~humidity,data=x_tr,family=binomial,control=list(maxit=100))
> summary(lmod)
```

```
Call:
glm(formula = play ~ humidity, family = binomial, data = x_tr,
    control = list(maxit = 100))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651  -1.3537   0.7585   1.0108   1.0108

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0986     1.1547   0.951    0.341
humidity     -0.6931     1.4720  -0.471    0.638

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11.457  on 8  degrees of freedom
Residual deviance: 11.229  on 7  degrees of freedom
AIC: 15.229

Number of Fisher Scoring iterations: 4

> lmod=glm(play~temp,data=x_tr,family=binomial,control=list(maxit=100))
> summary(lmod)

Call:
glm(formula = play ~ temp, family = binomial, data = x_tr, control = list(maxit = 100))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651  -1.1774   0.7585   0.9005   1.1774

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 8.019e-17  1.414e+00   0.000    1.000
temphot     6.931e-01  1.871e+00   0.371    0.711
tempmild    1.099e+00  1.826e+00   0.602    0.547

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11.457  on 8  degrees of freedom
Residual deviance: 11.090  on 6  degrees of freedom
AIC: 17.09

Number of Fisher Scoring iterations: 4
```

Conclusion: Hence we successfully implemented Demo Of Logistic Regression