DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT230 – DATA MINING

Mini Project

DUE: On/before 29/05/2022 @ 5:00pm

This assignment can be completed in groups of five (5).

Context

In this dataset, we will take a look at vaccination, a key public health measure used to fight infectious diseases. Vaccines provide immunization for individuals, and enough immunization in a community can further reduce the spread of diseases through "herd immunity."

Beginning in spring 2009, a pandemic caused by the H1N1 influenza virus, colloquially named "swine flu," swept across the world. Researchers estimate that in the first year, it was responsible for between 151,000 to 575,000 deaths globally.

A vaccine for the H1N1 flu virus became publicly available in October 2009. In late 2009 and early 2010, the United States conducted the National 2009 H1N1 Flu Survey. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves. These additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviours towards mitigating transmission. A better understanding of how these characteristics are associated with personal vaccination patterns can provide guidance for future public health efforts.

Content

The goal is to classify individuals into whether they receive their H1N1/ seasonal flu vaccines or not.

Each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey and there are two target variables and you have to one target variable for this exercise.

- **h1n1_vaccine**: Whether respondent received H1N1 flu vaccine.
- **seasonal_vaccine**: Whether respondent received seasonal flu vaccine.

Both are binary variables: 0 = No; 1 = Yes.

You are provided a dataset with 37 columns with the last two columns are the target variables. The remaining 35 features are described below (For all binary variables: 0 = No; 1 = Yes) :

- **age_group** - Age group of respondent.
- **education** - Self-reported education level.
- **race** - Race of respondent.
- **sex** - Gender of respondent.

- **income_poverty** - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- **marital_status** - Marital status of respondent.
- **rent*or*own** - Housing situation of respondent.
- **employment_status** - Employment status of respondent.
- **h1n1_concern** - Level of concern about the H1N1 flu. 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
- h1n1_knowledge - Level of knowledge about H1N1 flu. 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.
- **behavioral*wash*hands** - Has frequently washed hands or used hand sanitizer. (binary)
- **behavioral*large*gatherings** - Has reduced time at large gatherings. (binary)
- **behavioral*antiviral*meds** - Has taken antiviral medications. (binary)
- **behavioral_avoidance** - Has avoided close contact with others with flu-like symptoms. (binary)
- **behavioral*face*mask** - Has bought a face mask. (binary)
- **behavioral*outside*home** - Has reduced contact with people outside of own household. (binary)
- **behavioral*touch*face** - Has avoided touching eyes, nose, or mouth. (binary)
- doctor*recc*h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- **doctor*recc*seasonal** - Seasonal flu vaccine was recommended by doctor. (binary)
- **chronic*med*condition** - Has any of the following chronic medical conditions: asthma or an other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)
- **child*under*6_months -** Has regular close contact with a child under the age of six months. (binary)
- **health_worker** - Is a healthcare worker. (binary)
- **health_insurance** - Has health insurance. (binary)
- **opinion*h1n1*vacc_effective** - Respondent's opinion about H1N1 vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- **opinion*h1n1*risk** - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- **opinion*h1n1*sick*from*vacc** - Respondent's worry of getting sick from taking H1N1 vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- **opinion*seas*vacc_effective** - Respondent's opinion about seasonal flu vaccine effectiveness. 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- **opinion*seas*risk** - Respondent's opinion about risk of getting sick with seasonal flu without vaccine. 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- **opinion*seas*sick*from*vacc** - Respondent's worry of getting sick from taking seasonal flu vaccine. 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- **hhs*geo*region** - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
- **census_msa** - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
- **household_adults** - Number of other adults in household, top-coded to 3.
- **household_children** - Number of children in household, top-coded to 3.
- **employment_industry** - Type of industry respondent is employed in. Values are represented as short random character strings.

- **employment_occupation** - Type of occupation of respondent. Values are represented as short random character strings.

Using the above data, perform the following tasks:

a. Produce descriptive statistics for the following variables: **age_group, h1n1_concern,** and **sex.**
b. Construct a contingency table for **h1n1_concern (row)** and **age_group (column).**
c. Perform a chi-square test of independence to test the association between **h1n1_concern (response variable)** and age**_group (explanatory variable)** at 5% level of significance.
d. Use Cramer's V contingency coefficient to measure the strength of the association between **h1n1_concern** and **age_group.**
e. Build a decision tree algorithm for the Vaccine dataset using either **h1n1_vaccine** or **seasonal_vaccine** as the target variables. Determine the overall accuracy of the algorithm[**Hint**: Use 70% (18692) of the dataset for the training set and 30% (8,012) for the testing set]

Below is a sample code for building a decision tree:

```
#In this exercise we classify the iris data into four species
#using Species as the target variable and the remaining
#variables as the explanatory variables

#Load the tree package
library(tree)



set.seed(2)

#Dividing the Data set into training and test sets
train<-sample(1: nrow(iris), 100) # sample 100 records for the training
iris_test<- iris[-train, ] # Using the remaining records for the testing

#Extract the target variable in the iris data
#excluding those in the training set
Species_test<- iris$Species[- train ]
```

```
#Training the algorithm
tree_iris<-tree(Species~.,iris,subset = train )
#Plot the resulting tree
plot(tree_iris, type = "uniform")
#Add text to the tree
text(tree_iris, all = TRUE)

#Making prediction using the built algorithm
tree.pred<-predict (tree_iris ,iris_test,type = "class")

#Construct a confusion table for the prediction
Tab<-table(tree.pred ,Species_test)

#Compute the accuracy of the algorithm
Accuracy<-sum(diag(Tab))/sum(Tab)*100

#Print out the resulat
print(paste("Accuracy for Test",Accuracy))



set.seed(7)
#Perform a cross-validation to choose tree complexity
cv.iris<-cv.tree(tree_iris, FUN =prune.misclass)
cv.iris

par(mfrow=c(1,2))
#Plot size of the tree against deviance
plot(cv.iris$size, cv.iris$dev, type = "b")
#plot k against deviance
plot(cv.iris$k,cv.iris$dev, type = "b")
par(mfrow=c(1,1))
```

```
#Prunning the tree to obtain 4-node tree
prune.iris<-prune.misclass( tree_iris, best = 4)

#Plot the resulting tree after prunning
plot(prune.iris,type = "uniform")

#Add text to the tree
text(prune.iris, all = TRUE)

#Make prediction using the prunned tree
tree.pred<-predict(prune.iris,iris_test,type = "class")

#Construct a confusion table
TT<-table(tree.pred,Species_test)

#Compute accuracy
Accuracy<-sum(diag(TT))/sum(TT)*100

#print output
print
```

NOTE 1: Write a very nice report for your analysis, explaining your results in context. Add the R codes for the analysis as an appendix.

NOTE 2: You can choose not to use the tree package and it associated functions. There is another R package called ***rpart*** *which can also be used to build a decision tree. You can tutorials on it on YouTube.*