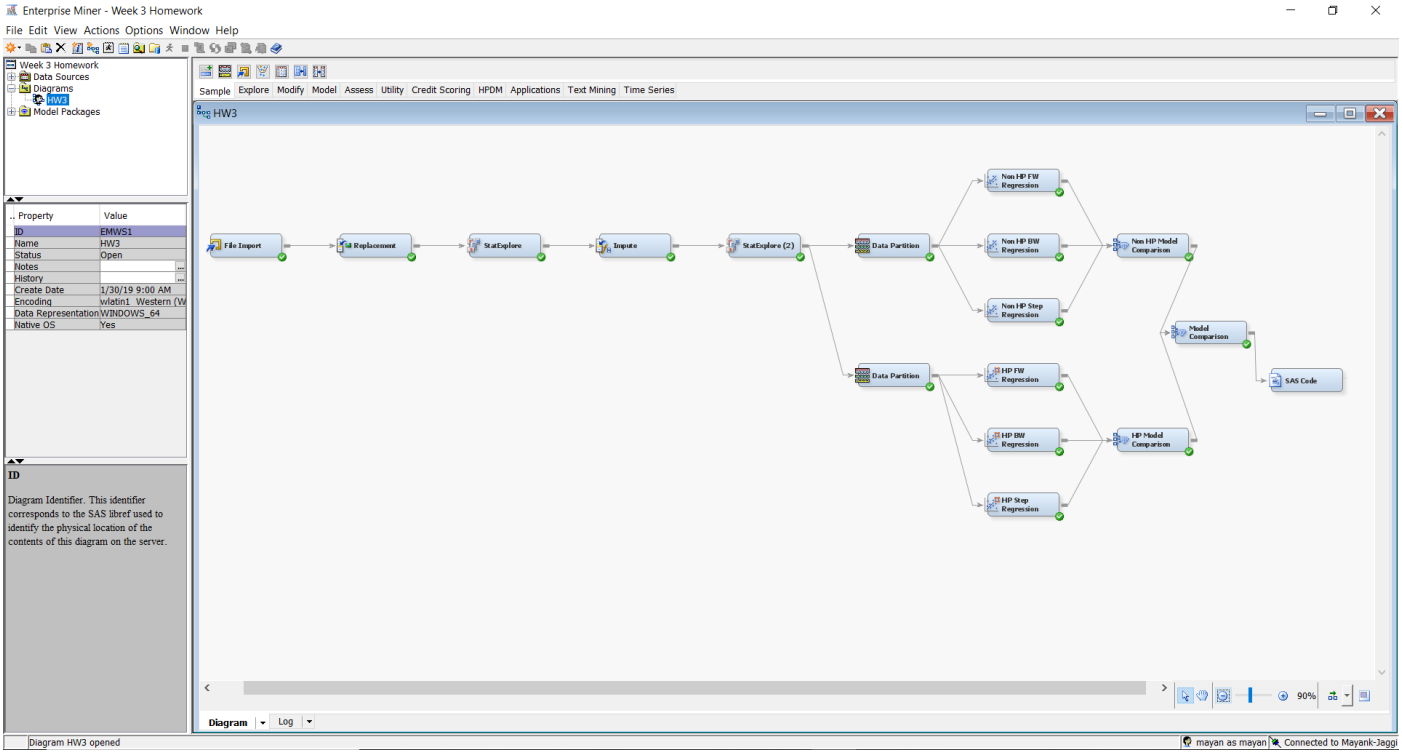
STAT 656
Homework 3

Name-Mayank Jaggi

UIN-526005299

PART 1: SAS ENTERPRISE MINER

Screenshot of Project Window



Note: Couldn’t complete second and third part of Part 1

PART 2
PYTHON PROGRAM

```
# -*- coding: utf-8 -*-
"""
Created on Wed Jan 30 12:29:26 2019

@author: mayank
"""

import pandas as pd
import numpy as np
from AdvancedAnalytics import ReplaceImputeEncode
from sklearn import preprocessing
from sklearn.model_selection import train_test_split

df2 = pd.read_excel("sonar3by5.xlsx")     #data file name
#df2 = df1.dropna(subset = ['object'])  # removing the column object as its the
target


print(df2)

#Missing values and outliers

data_map = {
            'R1': [0,(0,1)]
            ,'R2': [0,(0,1)]
            ,'R3': [0,(0,1)]
            ,'R4': [0,(0,1)]
            ,'R5': [0,(0,1)]
            ,'R6': [0,(0,1)]
            ,'R7': [0,(0,1)]
            ,'R8': [0,(0,1)]
            ,'R9': [0,(0,1)]
            ,'R10': [0,(0,1)]
            ,'R11': [0,(0,1)]
            ,'R12': [0,(0,1)]
            ,'R13': [0,(0,1)]
            ,'R14': [0,(0,1)]
            ,'R15': [0,(0,1)]
            ,'R16': [0,(0,1)]
            ,'R17': [0,(0,1)]
            ,'R18': [0,(0,1)]
            ,'R19': [0,(0,1)]
            ,'R20': [0,(0,1)]
            ,'R21': [0,(0,1)]
            ,'R22': [0,(0,1)]
            ,'R23': [0,(0,1)]
```

```
            ,'R24': [0,(0,1)]
            ,'R25': [0,(0,1)]
            ,'R26': [0,(0,1)]
            ,'R27': [0,(0,1)]
            ,'R28': [0,(0,1)]
            ,'R29': [0,(0,1)]
            ,'R30': [0,(0,1)]
            ,'R31': [0,(0,1)]
            ,'R32': [0,(0,1)]
            ,'R33': [0,(0,1)]
            ,'R34': [0,(0,1)]
            ,'R35': [0,(0,1)]
            ,'R36': [0,(0,1)]
            ,'R37': [0,(0,1)]
            ,'R38': [0,(0,1)]
            ,'R39': [0,(0,1)]
            ,'R40': [0,(0,1)]
            ,'R41': [0,(0,1)]
            ,'R42': [0,(0,1)]
            ,'R43': [0,(0,1)]
            ,'R44': [0,(0,1)]
            ,'R45': [0,(0,1)]
            ,'R46': [0,(0,1)]
            ,'R47': [0,(0,1)]
            ,'R48': [0,(0,1)]
            ,'R49': [0,(0,1)]
            ,'R50': [0,(0,1)]
            ,'R51': [0,(0,1)]
            ,'R52': [0,(0,1)]
            ,'R53': [0,(0,1)]
            ,'R54': [0,(0,1)]
            ,'R55': [0,(0,1)]
            ,'R56': [0,(0,1)]
            ,'R57': [0,(0,1)]
            ,'R58': [0,(0,1)]
            ,'R59': [0,(0,1)]
            ,'R60': [0,(0,1)]
            ,'object': [1,('R','M')]

        }

rie = ReplaceImputeEncode(data_map=data_map, display=True)
df_rie = rie.fit_transform(df2)


#Imputing Missing Values

interval_att=['R1','R2','R3','R4','R5','R6','R7','R8','R9','R10','R11','R12','R13','
R14',
```

```
'R15','R16','R17','R18','R19','R20','R21','R22','R23','R24','R25','R26','R27',

'R28','R29','R30','R31','R32','R33','R34','R35','R36','R37','R38','R39','R40',

'R41','R42','R43','R44','R45','R46','R47','R48','R49','R50','R51','R52','R53',
            'R54','R55','R56','R57','R58','R59','R60']      # list of attributes
with interval data type
interval_data=df2.as_matrix(columns=interval_att)
interval_impute=preprocessing.Imputer(strategy='mean')
interval_data_imputed = interval_impute.fit_transform(interval_data)
print("Imputed Interval Data:\n", interval_data_imputed)


df2[['R1','R2','R3','R4','R5','R6','R7','R8','R9','R10','R11','R12','R13','R14',

'R15','R16','R17','R18','R19','R20','R21','R22','R23','R24','R25','R26','R27',

'R28','R29','R30','R31','R32','R33','R34','R35','R36','R37','R38','R39','R40',

'R41','R42','R43','R44','R45','R46','R47','R48','R49','R50','R51','R52','R53',
            'R54','R55','R56','R57','R58','R59','R60']] = interval_data_imputed
df2.head()


from pandas import ExcelWriter
writer_file = ExcelWriter('Python_Export.xlsx')
df2.to_excel(writer_file)
writer_file.save()


from AdvancedAnalytics import logreg
#from sklearn.datasets import make_regression
from sklearn.linear_model import  LogisticRegression

y = df2['object']
x = df2.drop('object',axis=1)

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3,
random_state=1)

lr= LogisticRegression()



lr.fit(x_train,y_train)
print("\n*** LOGISTIC REGRESSION ***")


y_hat= lr.predict(x_test)
xtest1 = np.asanyarray(x_test)
```

```
ytest1 = np.asanyarray(y_test)

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(ytest1, y_hat)
print("\n*** Confusion Matrix ***\n",cm)

mcr = (cm[0,1]+cm[1,0])/(sum(sum(cm)))
print('Misclassification Rate : ', mcr )

sensitivity1 = cm[0,0]/(cm[0,0]+cm[0,1])
print('Sensitivity : ', sensitivity1 )

specificity1 = cm[1,1]/(cm[1,0]+cm[1,1])
print('Specificity : ', specificity1)


print("\nFirst 15 predicted values\n",y_hat[0:14])
```

OUTPUT

```
*** Confusion Matrix ***
 [[28  4]
 [12 19]]
Misclassification Rate :  0.25396825396825395
Sensitivity :  0.875
Specificity :  0.6129032258064516

First 15 predicted values
 ['M' 'R' 'R' 'M' 'R' 'R' 'M' 'R' 'M' 'R' 'M' 'M' 'M' 'M']
```