**PRAGYA VERMA – (101276204)**

**SEIS 763: 01 Machine Learning**

**Assignment #2**

**1 - Load the patient data from “ML\_HW\_Data\_Patients.csv” file.**

#import libraries  
  
import pandas as pd  
from sklearn.preprocessing import StandardScaler  
from sklearn.linear\_model import LinearRegression  
from sklearn.model\_selection import train\_test\_split  
import statsmodels.formula.api as smf  
import statsmodels.api  
import statsmodels.api as sm  
import scipy.stats as stats  
import pylab  
import matplotlib.pyplot as plot  
  
  
  
#Load dataset  
Pt\_data=pd.read\_csv("ML\_HW\_Data\_Patients.csv")  
  
#print top 5 rows of dataset  
print(Pt\_data.head())

**2 - Use variables Age, Gender, Height, Weight, Smoker, Location, SelfAssessedHealthStatus to build a linear regression model to predict the systolic blood pressure.**

#Extracting variables to be used in linear regression  
  
Extvar\_Model= Pt\_data[["Age", "Gender", "Height","Weight","Smoker","Location","SelfAssessedHealthStatus","Systolic"]]  
print(Extvar\_Model.head())  
  
#scale the numerical columns uniformly  
#Using StandardScaler/Zscaler to uniformly scale the numerical columns.  
BP\_scaler = StandardScaler()  
Extvar\_Model[['Age','Height', 'Weight']] = BP\_scaler.fit\_transform(Extvar\_Model[['Age','Height', 'Weight']])  
print(Extvar\_Model.head())  
  
#create dummy variables for Categorical columns and drop first category  
Extvar\_Model = pd.get\_dummies(Extvar\_Model,drop\_first=True)  
print(Extvar\_Model.head())  
  
#Rename column names which are long  
Extvar\_Model.rename(columns = {"Gender\_'Male'":"Male","Location\_'St. Mary's Medical Center'": "Loc\_StMaryMedCtr", "Location\_'VA Hospital'": "Loc\_VAHosp", "SelfAssessedHealthStatus\_'Fair'": "HealthStat\_Fair","SelfAssessedHealthStatus\_'Good'": "HealthStat\_Good","SelfAssessedHealthStatus\_'Poor'": "HealthStat\_Poor" }, inplace = True)  
print(Extvar\_Model.head())  
  
# creating formula for regression model .  
Formula = 'Systolic ~ Age + Height + Weight + Smoker + Male + Loc\_StMaryMedCtr + Loc\_VAHosp + HealthStat\_Fair + HealthStat\_Good + HealthStat\_Poor'  
  
#Fitting Linear Regression model and displaying the coefficients.  
model = smf.ols(formula= Formula, data = Extvar\_Model).fit()  
model.params.sort\_values()  
print(model.params.sort\_values())  
  
  
print(model.resid.sort\_values())  
print(model.summary())

**3 – What are the regression coefficients (thetas).**

**Variables θ (Theta)**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

HealthStat\_Fair -2.750968

Loc\_VAHosp -1.734841

Male -1.479391

Loc\_StMaryMedCtr -0.856501

Weight -0.354757

HealthStat\_Poor 0.459343

Age 0.576204

HealthStat\_Good 0.586379

Height 1.325387

Smoker 9.673087

Intercept 121.161481

**4 - How do you interpret those numbers in thetas?**

**For Numeric Variable (Age, Weight, height) –** Presence of numerica variables means how much blood pressure is going to change when we add or subtract 1 unit in from these variables. Since we have normalized the numeric columns, we use the standard deviation as a unit.

**Example:** If age goes by 1 unit, or one standard deviation, their systolic blood pressure will go up by **+0. 576204**.

**For Categorical Variable (Health Status, Location, Gender, Smoker) –** Categorical variable represents whether something is true or not. When categorical variables are presents, theta factor will affect the blood pressure.

**Example:** if a patient smokes, his/her blood pressure will go up by **+9. 673087** if we compare it to a person who does not smoke**.**

**5 - If you need to identify one or few useless features (independent variables or predictors), which one(s) will you choose? Why do you reach this conclusion?**

**Ans:** To check which variable or feature is useless, we can use theta and p value of variables.

print(model.summary())

**High P-value (> 0.05):** The feature is not significant and we can remove.

**Low P-value (≤ 0.05):** The feature is significant.

* We can see that Weight **coefficient** has the lowest value **-0.354757**
* Weight has the highest **p-value 0.819.** It has minimal effect and can be removed from systolic blood pressure prediction.

