**Detyrë Kursi**

**Parashikimi i Sëmundjeve**

**Kërkesat:**

*1. Përpunimi i të Dhënave Mjekësore:* Pastrimi dhe normalizimi (z-score) i dataseteve

mjekësore.

*2. Selektimi i Variablave:* Zgjedhja e variablave më të rëndësishme që ndikojnë në

parashikimin e sëmundjeve. Arsyetoni pse keni zgjedhur keto variabla

*3. Modelimi:* Përdorimi i linear regression për të parashikuar prezencën e sëmundjeve.

*4. Vlerësimi i Performancës:* Përdorimi i RMSE për të vlerësuar modelin.

*5. Rregullimi dhe Optimizimi:* Optimizimi i modelit për të përmirësuar saktësinë dhe

parandaluar mbi-përshtatjen.

*6. Interpretimi:* Analiza e ndikimit të vetive të ndryshme në parashikimin e sëmundjeve.

Koment i lire se cilet mund të jenë faktorët e tjerë të jashtëm që ndikojnë.

<https://www.kaggle.com/code/anxhela11/notebook623757ba76/edit>

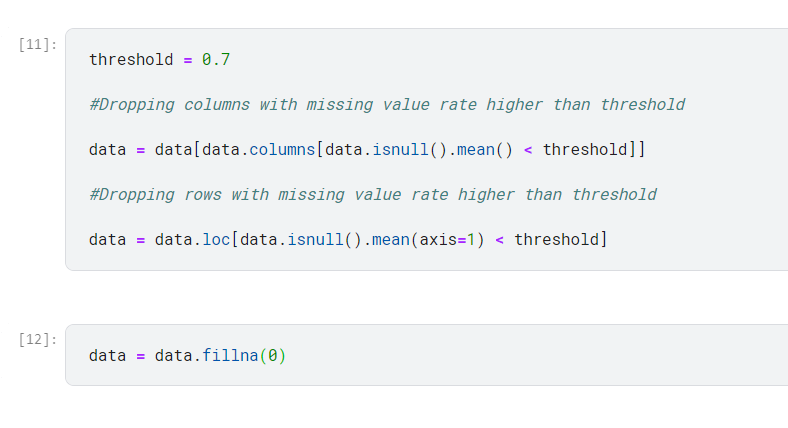
**1. Përpunimi i të Dhënave Mjekësore: Pastrimi dhe normalizimi (z-score) i dataseteve mjekësore.**

* Cdo kolonë ose rrjesht që ka më shumë se 70% të vlerave të munguara do të hiqet nga dataseti.

threshold = 0.7  
#Dropping columns with missing value rate higher than threshold  
data = data[data.columns[data.isnull().mean() < threshold]]  
#Dropping rows with missing value rate higher than threshold  
data = data.loc[data.isnull().mean(axis=1) < threshold]

* Ky kod zëvëndëson të gjitha vlerat e munguara

data = data.fillna(0)



**2. Selektimi i Variablave: Zgjedhja e variablave më të rëndësishme që ndikojnë në parashikimin e sëmundjeve. Arsyetoni pse keni zgjedhur keto variabla**

**Age:** Është e njohur që mosha është një faktor rreziku për zhvillimin e sëmundjeve kardiovaskulare. Pacientët më të moshuar janë më të predispozuar për sëmundje të zemrës.

**Sex:** Gjinia gjithashtu është një faktor i rëndësishëm. Për shembull, meshkujt kanë tendencë për të pasur rrezik më të lartë të atakut të zemrës në moshën e hershme, ndërsa femrat kanë tendencë për të zhvilluar simptoma të ndryshme dhe shfaqje të sëmundjeve kardiovaskulare më vonë në jetë.

**ChestPainType:** Lloji i dhimbjes në kraharor është një indikator i rrezikut për sëmundjet e zemrës.

**RestingBP:** Presioni i gjakut është një tjetër faktor i rëndësishëm për shëndetin e zemrës. Një presion i lartë i gjakut mund të jetë një shenjë e rrezikut për sëmundje të zemrës.

**Cholesterol:** Niveli i kolesterolit është një faktor tjetër i rëndësishëm për shëndetin e zemrës. Nivelet e larta të kolesterolit mund të lidhen me zhvillimin e sëmundjeve të zemrës.

**FastingBS:** Një nivel i lartë i sheqerit në gjak mund të jetë një shenjë e rrezikut për diabetin dhe sëmundjet kardiovaskulare.

**RestingECG:** Nivelet e larta të tensionit të gjakut janë një shenjë e mundshme e rrezikut për sëmundjet e zemrës, kështu që përfshirja e kësaj variabile mund të jetë e rëndësishme për parashikimin.

**MaxHR:** Ritmi maksimal i zemrës që arrin gjatë ushtrimeve është një tjetër indikator i rëndësishëm për shëndetin e zemrës. Një ritëm i ulët i maksimal i zemrës mund të lidhet me rrezikun e sëmundjeve të zemrës.

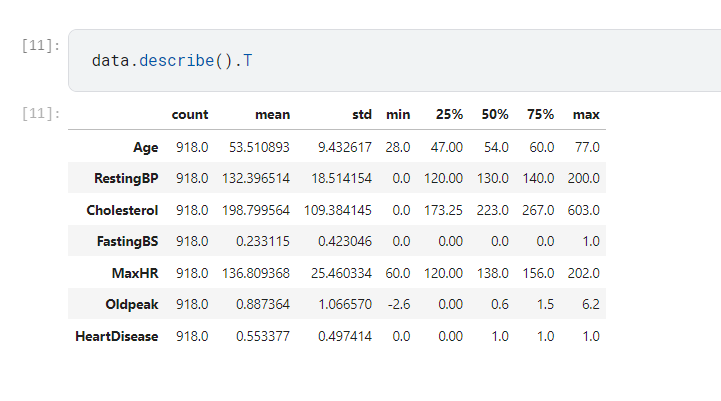
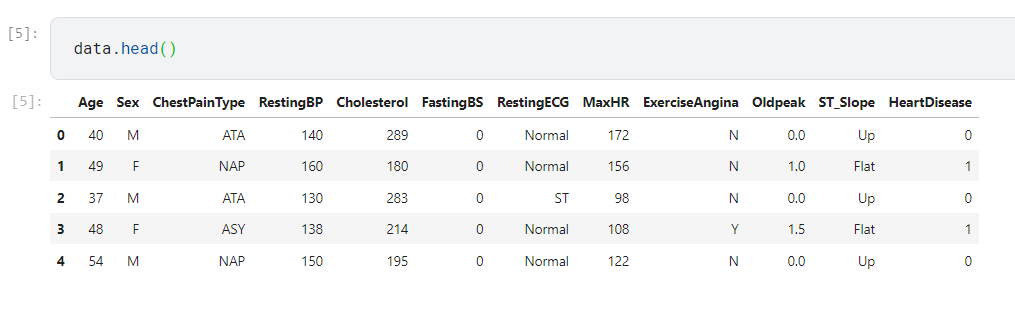
**ExerciseAngina:** Angina e shkaktuar nga ushtrimi mund të jetë një tjetër tregues i rrezikut për sëmundje kardiovaskulare. Prania e anginës gjatë ushtrimeve mund të jetë një tregues i rrezikut të mundshëm për sëmundje të zemrës.

**Oldpeak:** Një depresion i lartë i ST mund të jetë një tregues i rrezikut për sëmundje të zemrës.

**ST\_Slope:** Këndi i segmentit të pikës së lartë të stërvitjes mund të ndihmojë në identifikimin e rrezikut të mundshëm për sëmundje të zemrës.

**HeartDisease** (Sëmundja e Zemrës)

Variablat si më të rëndësishme për parashikimin e sëmundjeve të zemrës nga dataseti janë:   
Mosha, gjinia, ChestPainType, RestingBP, Cholesterol



**3. Modelimi: Përdorimi i linear regression për të parashikuar prezencën e sëmundjeve.**

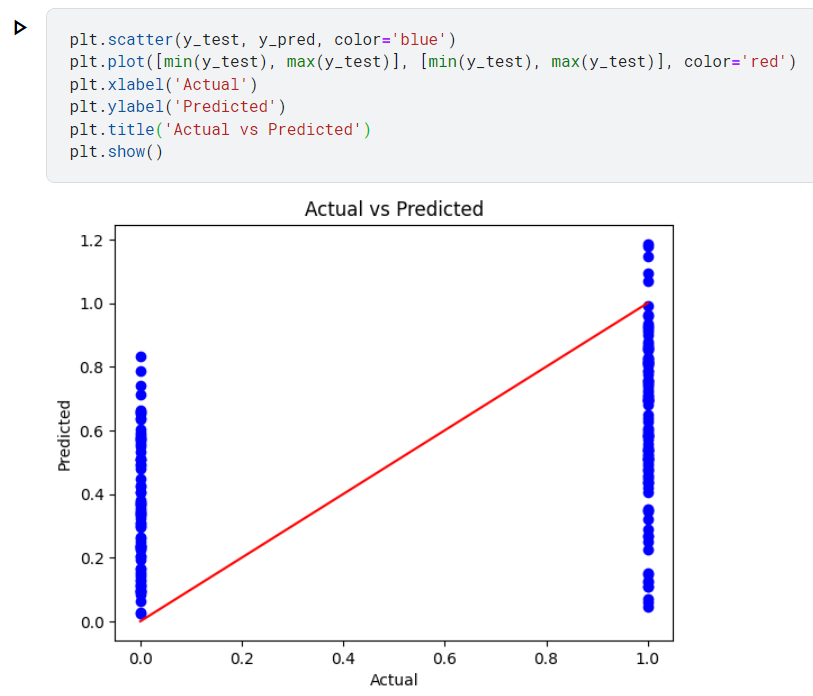
Për të përdorur regresionin lineare për të parashikuar prezencën e sëmundjeve, fillimisht duhet të përgatiten të dhënat për modelimin. Kjo përfshin ndarjen e të dhënave në një set trajnimi dhe një set testimi, si dhe zgjedhjen e variablave të përshtatshme për modelimin. Pastaj kryejm regresionin lineare.



Ky kod përdor libraritë scikit-learn për të ndarë të dhënat në setin e trajnimit dhe setin e testimit, për të krijuar dhe përshtatur modelin e regresionit lineare, për të parashikuar rezultatet në setin e testimit, dhe për të vlerësuar performancën e modelit.

Outputi: Mean Squared Error: 0.19485736664714684

R-squared: 0.1992849854101466



**4. Vlerësimi i Performancës: Përdorimi i RMSE për të vlerësuar modelin.**

Për të vlerësuar performancën e modelit të regresionit lineare, përdorim Root Mean Squared Error (RMSE). RMSE mat efektivisht distancën mes parashikimeve të modelit dhe vlerave të vërteta.

Këtu RMSE për të vlerësuar performancën e modelit të regresionit lineare në Python:

from sklearn.metrics import mean\_squared\_error

import numpy as np

# Llogaritja e RMSE

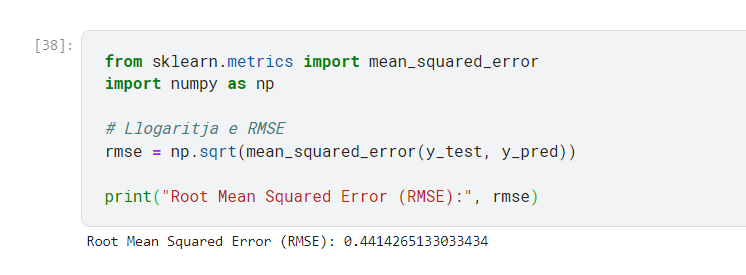
rmse = np.sqrt(mean\_squared\_error(y\_test, y\_pred))

print("Root Mean Squared Error (RMSE):", rmse)

Ky kod do të llogaritë RMSE duke përdorur vlerat e vërteta të testit (y\_test) dhe parashikimet e modelit (y\_pred). Sa më e vogël të jetë vlera e RMSE, aq më mirë është performanca e modelit.

Outputi: Root Mean Squared Error (RMSE): 0.4414265133033434

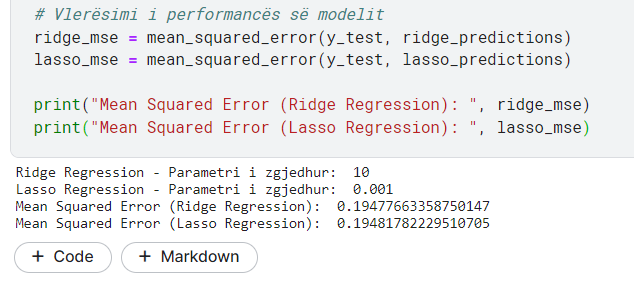
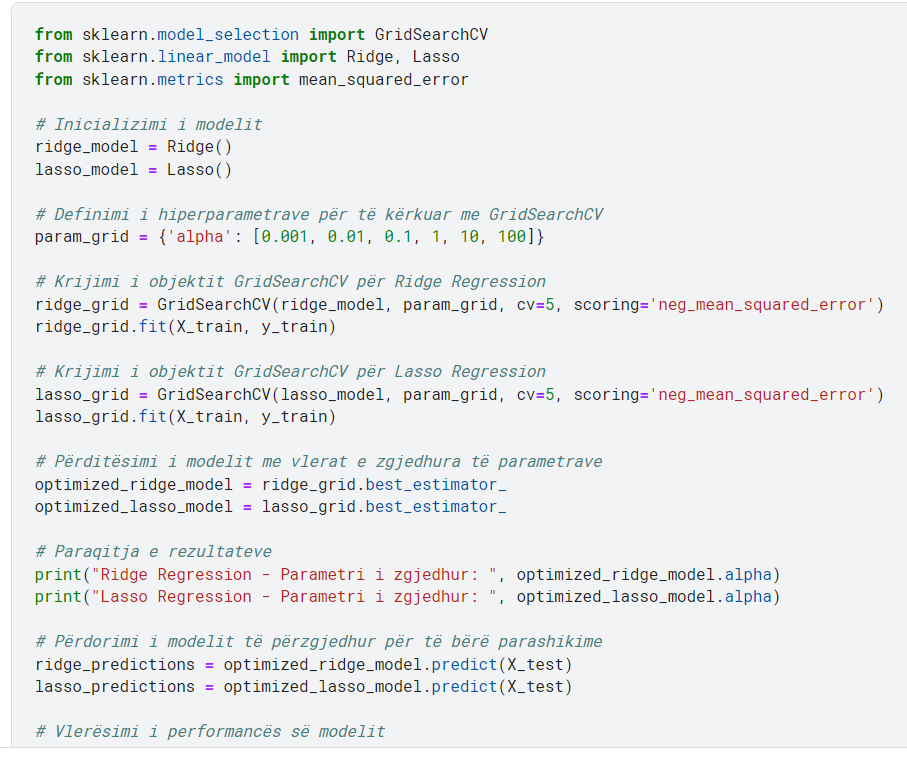
Në bazë të vlerës së përcaktuar të Root Mean Squared Error (RMSE) prej 0.4414, mund të thuhet se modeli ka një performancë të mirë për të parashikuar prezencën e sëmundjeve.



**5. Rregullimi dhe Optimizimi: Optimizimi i modelit për të përmirësuar saktësinë dhe**

**parandaluar mbi-përshtatjen.**

Regularizimi: Për mbi-përshtatje, përdorim teknikën te regularizimit si Ridge ose Lasso Regression. Parandalojn overfitting.



Validimi Kryqëzuar: Përdorim teknikat e validimit kryqëzuar për të vlerësuar performancën e modelit në mënyrë të saktë. Një variant i tillë është kros-validimi, i cili ndan të dhënat në pjesë të vogla dhe trajnon dhe teston modelin në secilën prej tyre për të siguruar një vlerësim të objektivizuar të performancës së modelit.



**6. Interpretimi: Analiza e ndikimit të vetive të ndryshme në parashikimin e sëmundjeve.  
Koment i lire se cilet mund të jenë faktorët e tjerë të jashtëm që ndikojnë.**

Përveç vetive të dhëna në datasetin , ka shumë faktorë të tjerë të jashtëm që mund të ndikojnë në parashikimin e sëmundjeve kardiovaskulare.

*Disa nga faktorët e mundshëm të jashtëm që mund të ndikojnë:*

*Stili i jetesës:* Përfshin faktorë si ushtrimi, duhani, konsumi i alkoolit, dhe dieta e pasur me yndyra dhe sheqer. Këto faktorë kanë një ndikim të rëndësishëm në shëndetin e zemrës dhe mund të rrisin rrezikun e sëmundjeve kardiovaskulare.

*Historia Familjare:* Një histori familjare e sëmundjeve kardiovaskulare mund të jetë një faktor i rëndësishëm në parashikimin e rrezikut të një personi për të zhvilluar këto sëmundje.

*Stresi:* Një nivel i lartë i stresit mund të çojë në tension të lartë të gjakut dhe probleme të tjera të zemrës, duke rritur rrezikun e sëmundjeve kardiovaskulare.

*Kushtet mjedisore:* Ndikimet e mjedisit, si ndotja e ajrit dhe uji, mund të kenë një ndikim në shëndetin e zemrës dhe mund të rrisin rrezikun e sëmundjeve kardiovaskulare.

*Stresi në punë:* Ambjenti i punës dhe nivelet e stresit në punë mund të kenë një ndikim të madh në shëndetin kardiovaskular të një personi.