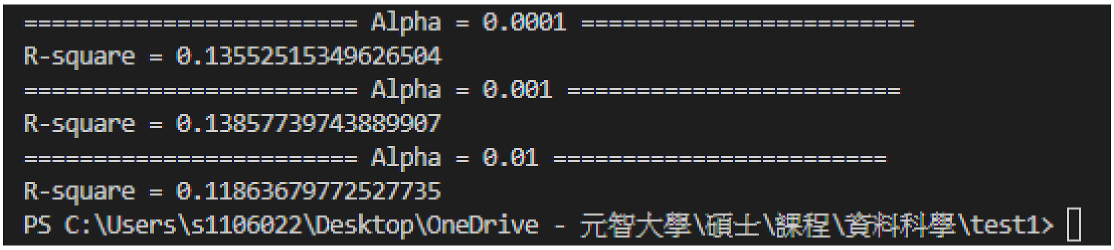
資料科學 Report of Test 1

1106022 陳柏嘉

Code 與註解：

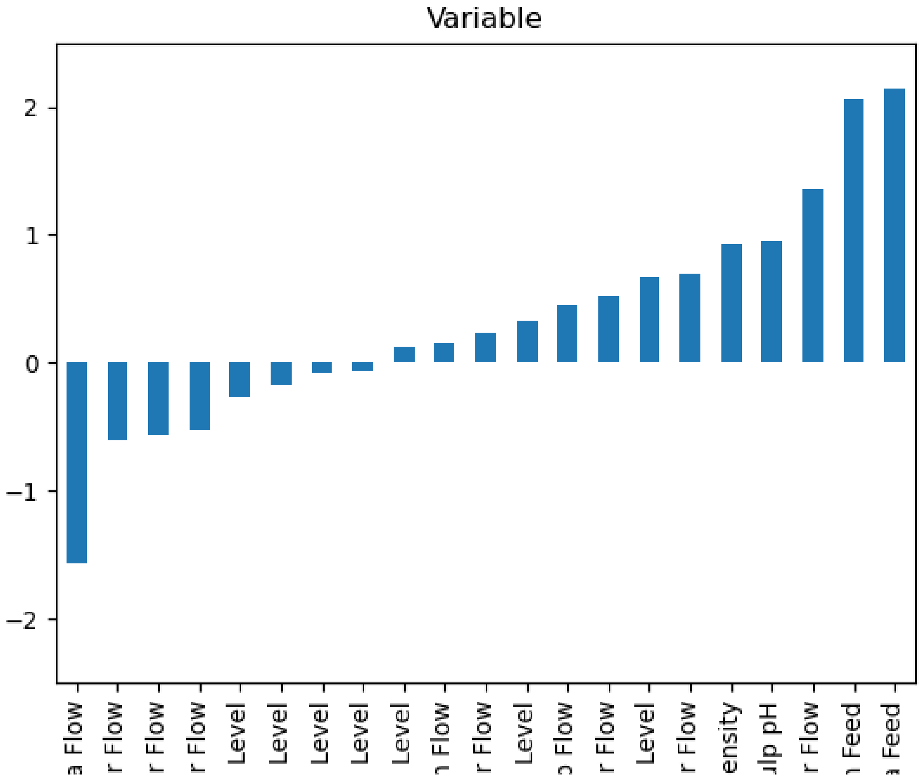
```
from sklearn.linear_model import SGDRegressor
from sklearn import preprocessing
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

df = pd.read_csv("MiningProcess_Flotation_Plant_Database.csv")  # 讀取 CSV 檔
df.drop('date', inplace=True, axis=1)   # 把 date 這個 column 丟掉，在這邊屬於無用資訊
list_of_column_names = list(df.columns) # 取得 CSV 檔的 column list

print(df.shape) # 查看資料筆數與欄位數

for col in list_of_column_names:
    df[col] = df[col].str.replace(',', '.') # 此資料集小數點是","，替換成"."
    df[col] = df[col].astype('float')       # 轉成 float 型別

Y=df['% Iron Concentrate']  # 要預測的欄位
X=df.drop(['% Iron Concentrate', '% Silica Concentrate'],axis=1)    # 用來訓練的欄位

min_max_scaler = preprocessing.MinMaxScaler()
X_scaled = pd.DataFrame(min_max_scaler.fit_transform(X), columns=X.columns) # 將資料縮放到 0~1 之間

X_train, X_test, Y_train, Y_test = train_test_split(X_scaled, Y, test_size = 0.2, random_state = 3) # 切割出訓練用資料與測試用資料

alphas = [0.0001, 0.001, 0.01]

for a in alphas:
    print("======================== Alpha = %s ========================" %(a))
    sgdr = SGDRegressor(loss='squared_loss', penalty='l2', alpha = a)   # 建立預測模型
    sgdr.fit(X_train, Y_train)  # 導入訓練資料
    Y_prdict = sgdr.predict(X_test) # 預測

    print("R-square =", sgdr.score(X_scaled, Y))    # 計算 R-square 值

    predictors = X_train.columns

    coef = pd.Series(sgdr.coef_,predictors).sort_values()   # sort predictors
    coef.plot(kind='bar', title='Variable')   # 建立直方圖
    plt.ylim(-2.5,2.5)
    plt.show()
```

輸出結果：

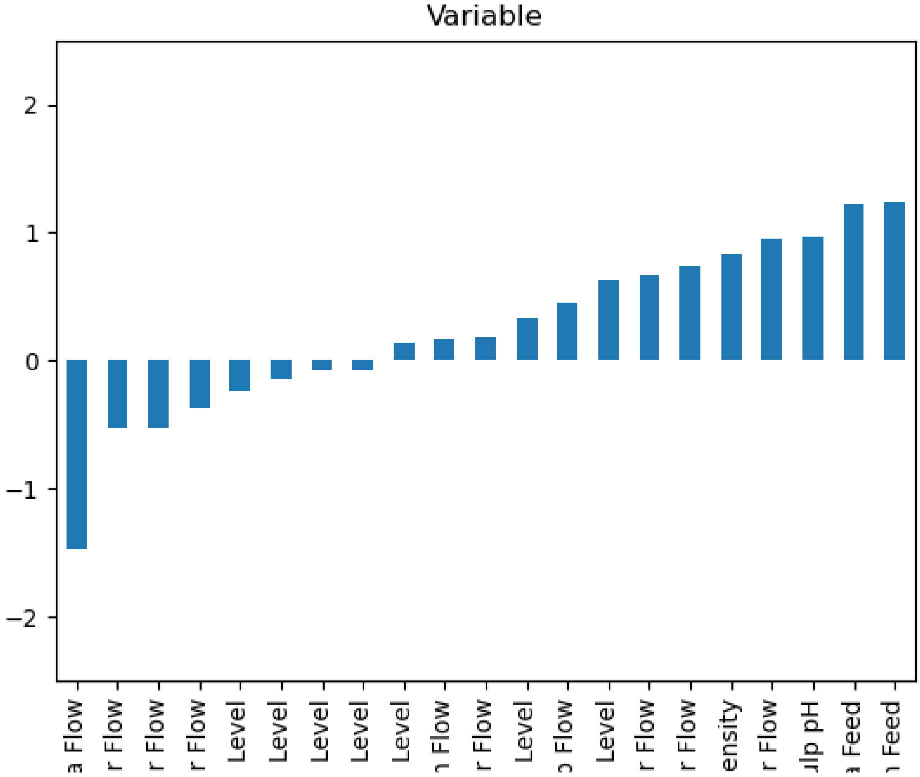


```
======================== Alpha = 0.0001 ========================
R-square = 0.13552515349626504
======================== Alpha = 0.001 ========================
R-square = 0.13857739743889907
======================== Alpha = 0.01 ========================
R-square = 0.11863679772527735
PS C:\Users\s1106022\Desktop\OneDrive - 元智大學\碩士\課程\資料科學\test1> 
```

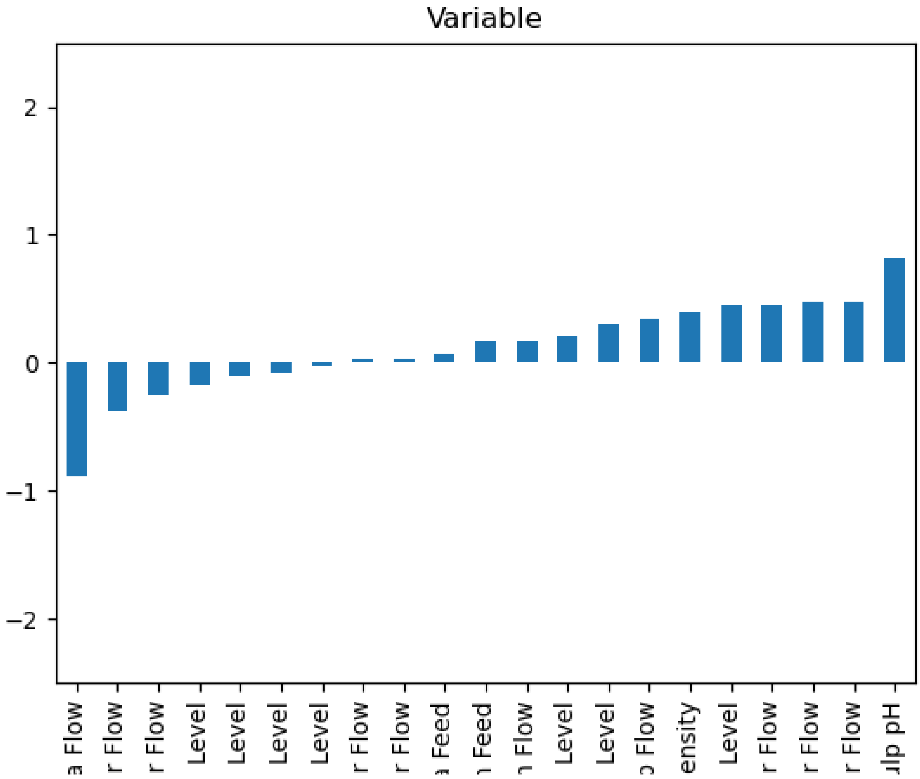
Alpha = 0.0001



Alpha = 0.001



Alpha = 0.01



Variable
2
1
0
−1
−2
a Flow
r Flow
r Flow
Level
Level
Level
Level
r Flow
r Flow
a Feed
ı Feed
ı Flow
Level
Level
ɔ Flow
ensity
Level
r Flow
r Flow
r Flow
ılp pH