Data Science Quiz

Question 1

Identify functions from the pandas library *(10 points)*

description	function
Shows the first n rows	head(n)
Writes a CSV file	df.to_csv("final_lyrics_dataf.csv")
Replaces index by a new one	set_index[“new index”]
Converts long to wide format	pd.dataframe.transpose()
Removes rows with missing values	pd.dropna(axis=0)
Swaps rows and columns in a DataFrame	df.pivot_table
Calculates minimum, median, mean, maximum etc.	pd.dataframe.describe()
Defines moving window over a time series	pd.dataframe.rolling(“time_window”)
Converts wide to long format	pandas.wide_to_long
Reads data from an Excel spreadsheet	pd.read_excel(“xyz.xlml”)

Question 2

Calculate the MSE from the values below *(5 points)*

y_true	1.2	3.4	5.6	7.8	9.0	10.11
y_pred	1.1	2.2	3.3	4.4	5.5	6.66

mean_squared_error(y_true,y_pred)

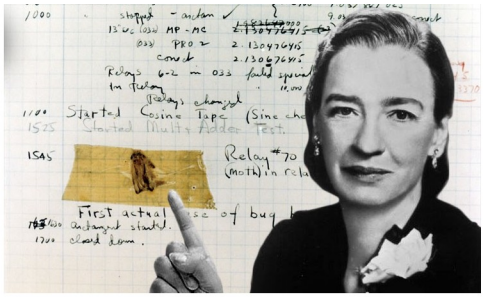
result:7.075

Question 3

Identify these persons? *(6 points)*

a)

Hans Rosling



b)

Grace Hopper

c)

Gauss

Question 4

Find 5 bugs: *(5 points)*

```
from sklearn.datasets import iris
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X, y = iris(return_X_y=True)

m = LogisticRegression(max_depth=3)
Xtrain, ytrain, Xtest, ytest = train_test_split(X, y,
random_state=42)
m.fit_transform(Xtrain, ytrain)
print('test    :', m.score(ytest, Xtest)
```

load_iris

max_depth

Xtrain,Xtest,ytrain,ytest

m.fit(Xtrain,ytrain)

Xtest,ytest

Question 5

What do the following git commands do? *(5 points)*

`git pull`	update the local version of a repository from a remote
`git log`	The git log command shows a list of all the commits made to a repository.
`git checkout orange`	navigate between the branches created by git branch
`git remote add origin <url>`	Add origin to a new remote repository
`git add .gitignore`	We are using to add something in our local respository

Question 6

Describe three assumptions of a linear regression model. *(9 points)*

1)Linear dependence at X and mean y
2)X features are independent,
3)y_true— y_linear follow normal

Question 7

Name 3 different classification and 3 regression models. *(6 points)*

Classification—>Random Forest, Decision Tree,Regression
Regression—>Logistic,Linear, Nearest Neighbors Regression

Question 8

Match each model with exactly one hyperparameter. *(8 points)*

Ridge	L2	C
SVM	Kernel type	L2 strength
Logistic Regression	C	number of trees
ElasticNet	L1/L2 ratio	degree
Decision Tree	max_depth	L1 strength
Lasso	L1	Kernel type
PolynomialFeatures*	Degree	L1 / L2 ratio
RandomForest	number of trees	maximum depth

*PolynomialFeatures is not a statistical model but a Feature Engineering Technique that transforms your input data.

Question 9

Check the correct answers. *(4 points)*

9.1 Which does **not** help against overfitting?

a) More training data
b) More test data b
c) Regularization
d) Simpler model

9.2 To reduce the regularization strength, should you increase or decrease the regularization hyperparameter 'alpha'?

a) increase
b) decrease decrease
c) neither

9.3 What is a linear Ridge regression model with an 'alpha' of zero equivalent to?

a) Lasso
b) ElasticNet
c) simple linear regression
d) Logistic Regression

9.4 Why would you want to use Lasso instead of Ridge Regression?

a) To discard unnecessary features
b) To apply stronger regularization
c) L1 is better as a first attempt than L2