SPICED Data Science Quiz

Question 1

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

description	**function**
Shows the first n rows	df.head(n)
Writes a CSV file	df.to_csv()
Replaces index by a new one	df.reset_index(), df.set_index(), df.reindex()
Converts long to wide format	df.pivot()
Removes rows with missing values	df.dropna()
Swaps rows and columns in a DataFrame	df.transpose()
Calculates minimum, median, mean, maximum etc.	df.describe()
Defines moving window over a time series	df.rolling()
Converts wide to long format	df.melt()
Reads data from an Excel spreadsheet	df.read_excel()

Question 2

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

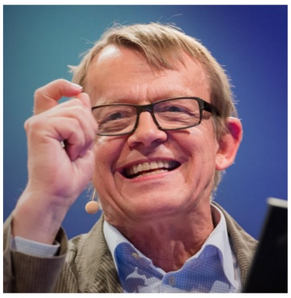
y_true	1.2	3.4	5.6	7.8	9.0	10.11
y_pred	1.1	2.2	3.3	4.4	5.5	6.66

Result: 7.075

sum((y_true - y_pred)^2) / len(y_true)

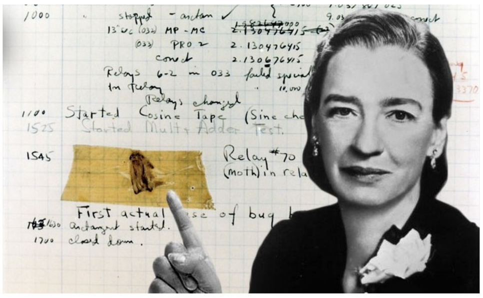
Question 3

Identify these persons? *(6 points)*



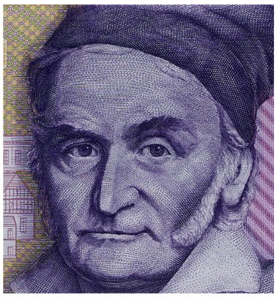
a)

Hans Rosling



b)

Grace Hopper



c)

Karl-Friedrich Gauss

Question 4

Find 5 bugs: *(5 points)*

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

       load_iris
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)))
```

Question 5

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

`git pull`	Fetches updates from the remote reposiory and merges them into the local repository
`git log`	Displays the history of git commands of your repo
`git checkout orange`	Switches to branch orange
`git remote add origin <url>`	It links your local repository to a remote repository
`git add .gitignore`	Adds the file .gitignore to the staging area

Question 6

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

Question 7

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

Classification:

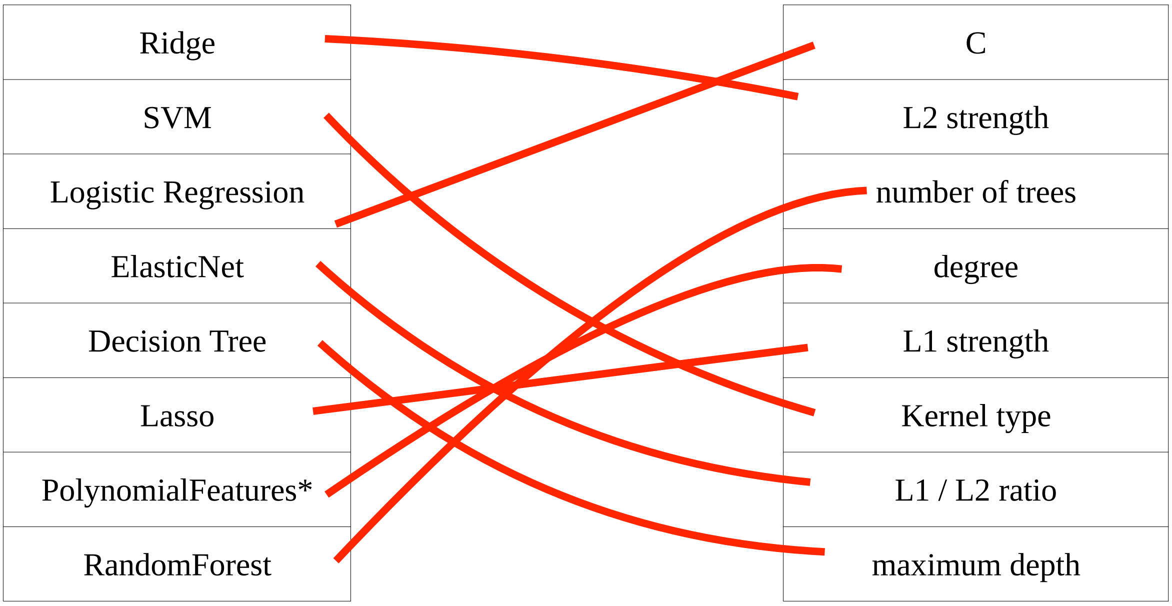
- Random Forrest
- Logistic Regression
- Decision Tree

Regression:

- (Multivariate) Linear Regression
- SVM
- Forecasting (AR, ARIMA)

Question 8

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



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