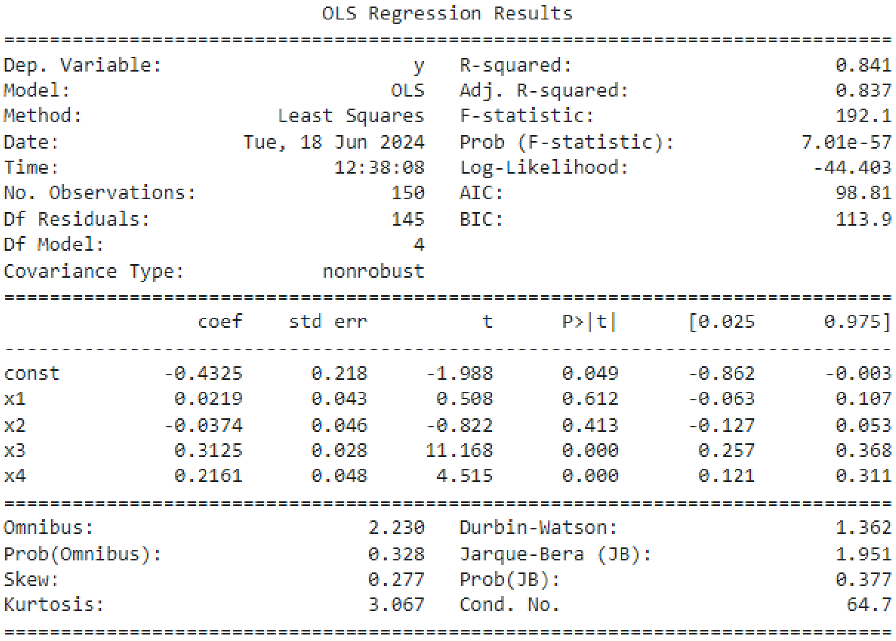
Machine Learning **Sep 04, 2024**

QUESTION SHEET

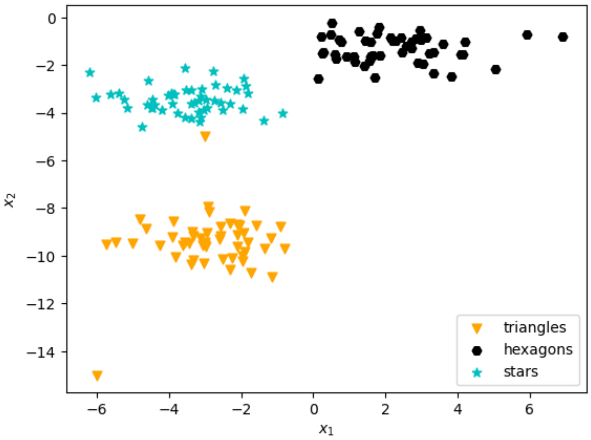
Question 1 ♣ Considering the OLS table shown below, mark the True(correct) statements.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.841
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                     192.1
Date:                Tue, 18 Jun 2024   Prob (F-statistic):           7.01e-57
Time:                        12:38:08   Log-Likelihood:                -44.403
No. Observations:                 150   AIC:                             98.81
Df Residuals:                     145   BIC:                             113.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.4325      0.218     -1.988      0.049      -0.862      -0.003
x1             0.0219      0.043      0.508      0.612      -0.063       0.107
x2            -0.0374      0.046     -0.822      0.413      -0.127       0.053
x3             0.3125      0.028     11.168      0.000       0.257       0.368
x4             0.2161      0.048      4.515      0.000       0.121       0.311
==============================================================================
Omnibus:                        2.230   Durbin-Watson:                   1.362
Prob(Omnibus):                  0.328   Jarque-Bera (JB):                1.951
Skew:                           0.277   Prob(JB):                        0.377
Kurtosis:                       3.067   Cond. No.                         64.7
==============================================================================
```

- A The features ‘x3’ and ‘x4’ should be removed because their p values are close to zero.
- ■ Removing features ‘x1’, and ‘x2’ from the regression model does not guarantee that the model will become valid.
- C Upsampling the dataset can make the regression model valid.
- ■ If we provide a higher number of observations, it may help the model detect significant coefficients.
- ■ Based on the OLS report, we cannot confirm that the ‘x3’ has a positive relationship with the dependent variable.
- F *None of these answers are correct.*

Question 2 ♣ Considering the following figure that shows the result of a clustering task, mark the True(correct) statements.



■ K-medoids method is less sensitive to outliers compared with K-means.

B Using **Single Linkage** distance, the stars will merge with the Hexagons in the next step of Agglomerative clustering.

C The silhouette coefficients for all Triangles are below zero.

■ Using Ward's distance, merging Stars and Hexagons is preferred over merging Hexagons and Triangles.

■ Complete Linkage distance tends to create globular clusters

F *None of these answers are correct.*

Question 3 ♣ Referring to regression, mark the True(correct) statements.

A In a Support Vector Regressor, feature selection is performed automatically as part of the training process.

■ In a linear regression model, the scale of the variables impacts the corresponding regression coefficients.

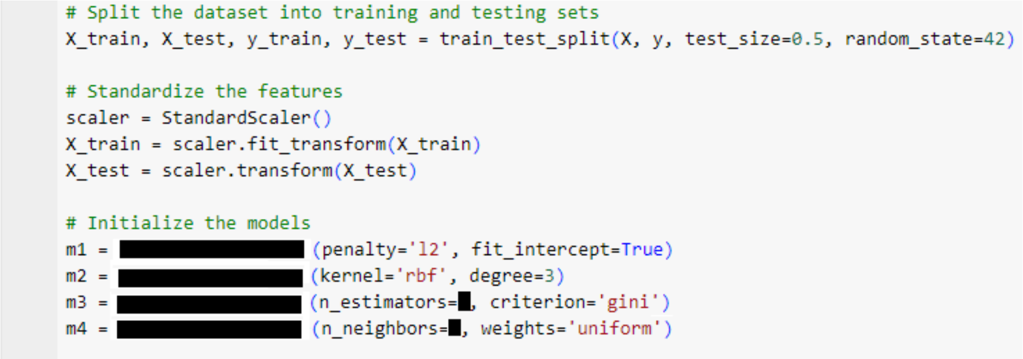
C A Naive Bayesian Regressor assumes that the explanatory features are independent of the target variable.

■ The adjusted R-squared coefficient can take negative values.

■ Boosting models such as AdaBoost are among supervised learning algorithms.

F *None of these answers are correct.*

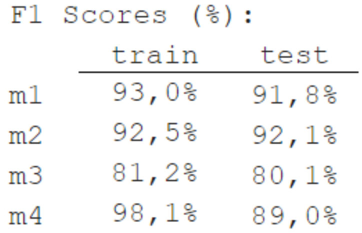
Question 4 ♣ Regarding a piece of code for a classification task with 3 different models, mark the True(correct) statements.



```
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=42)

# Standardize the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Initialize the models
m1 = [redacted](penalty='l2', fit_intercept=True)
m2 = [redacted](kernel='rbf', degree=3)
m3 = [redacted](n_estimators=[redacted], criterion='gini')
m4 = [redacted](n_neighbors=[redacted], weights='uniform')
```



```
F1 Scores (%):
         train    test
m1       93,0%    91,8%
m2       92,5%    92,1%
m3       81,2%    80,1%
m4       98,1%    89,0%
```

- ■ We should increase ‘n-neighbors’ for m4.
- B By increasing only the ‘degree’ value, m2 will eventually overfit.
- ■ The ‘l2’ term in m1 discourages large coefficients, which can help prevent overfitting.
- ■ Adding more estimators to m3 could improve its performance.
- ■ m3 is an ensemble learning model.
- F *None of these answers are correct.*

Question 5 ♣ Referring to classification, mark the True(correct) statements.

- ■ Perfect classification can still be achieved even if the classes of the target variable are not linearly separable, using non-linear classifiers.
- ■ If our binary classifier yields an AUC of 0.01 on the test set, it suggests that by reversing the labels we will have an AUC of 0.99.
- ■ In a multivariate binary tree, each split can be based on multiple explanatory features.
- ■ Support vectors are the training data points closest to the decision boundary and within or on the margin.
- E The generalization error tends to decrease as the hypothesis space becomes more and more complex.
- F *None of these answers are correct.*

Question 6 ♣ Mark the True(correct) statements.

- A The Pearson correlation coefficient is designed to measure the strength of a nonlinear relationship between two variables.
- B Filter methods for feature selection depend on the learning algorithm used for the task.
- ■ Observations outside the $[\mu - 4\sigma, \mu + 4\sigma]$ interval should be considered outliers.
- ■ After finding the best model using an n-fold cross-validation, we can use all the samples in the training set to re-train the best model.
- E For imbalanced datasets, it is often more advisable to use upsampling on the majority class.
- F *None of these answers are correct.*

Question 7 ♣ Mark the True(correct) statements.

A Increasing data dimensions can help prevent overfitting in supervised learning models.

B In the case of data incompleteness, it is always preferable to consider elimination techniques to prevent the injection of noise into the dataset.

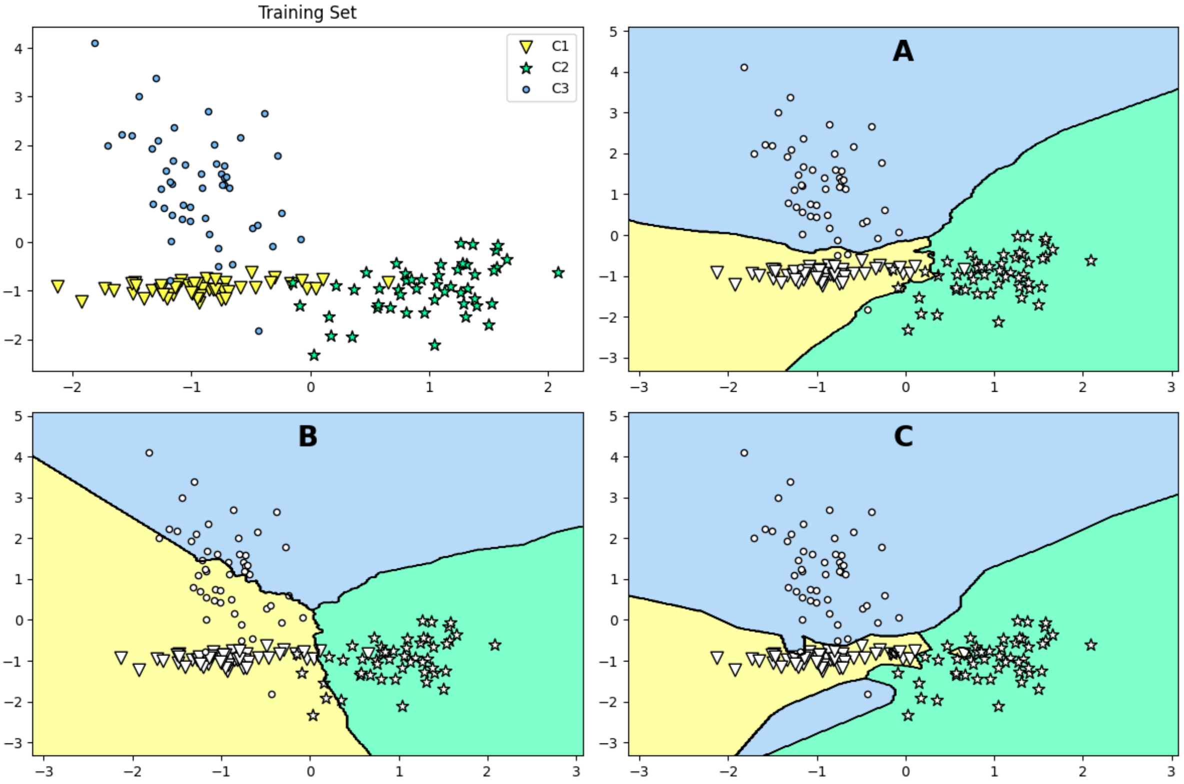
■ A zero correlation between an explanatory feature and the target should **not** be interpreted as the absence of a relation between the variables.

■ The Chebyshev theorem provides an interval that can be used to identify potential outliers in a dataset, regardless of the underlying distribution.

■ GridSearch systematically explores a range of hyperparameters to find the best combination for optimizing model performance.

F *None of these answers are correct.*

Question 8 ♣
The figure shows a training set for a multi-class classification task and the results of three k-NN models with different parameters. The training set (upper-left) has two numerical features and a categorical target with three classes: C1 (triangle), C2 (star), and C3 (dot). The coloured regions in the outputs represent the predicted values (C1, C2, and C3) for new observations based on the three different models.
Regarding the figure, mark the True(correct) statements.



A In model B, better results can be obtained by increasing the regularization parameter to give more importance to the regularization term.

■ Model B performs slower than model C, in terms of runtime.

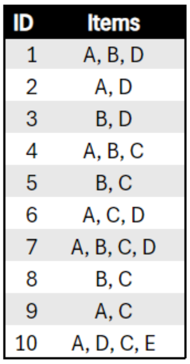
■ In model C, the parameter k is likely to be equal to one.

■ Among the three models, we expect that model A will have a higher accuracy on the test set.

■ The training precision of class C3 is greater in model C than that of model B.

F *None of these answers are correct.*

Question 9 ♣ Considering the list of transactions in the table below. Referring to the association rules, mark the True(correct) statements.



ID	Items
1	A, B, D
2	A, D
3	B, D
4	A, B, C
5	B, C
6	A, C, D
7	A, B, C, D
8	B, C
9	A, C
10	A, D, C, E

■ The apriori algorithm helps improve the efficiency of computing strong rules.

■ There is a negative association between $\{A, D\}$ and $\{C\}$.

■ The support of $\{A \Rightarrow D\}$ is 0.5.

■ The confidence of rule $\{\{A, D\} \Rightarrow C\}$ is 0.6.

■ If an itemset is not frequent, all its supersets are also not frequent.

F *None of these answers are correct.*

Question 10 In the text box of this question, you can add any comment related to the test (withdraw, clarification, etc).

Question 11 ♣ Black out the answer to this question (**in the answer sheet!**) to withdraw from the exam.

■

ANSWER SHEET

- **Make sure you read the instructions carefully.**
- Total time: 50 minutes.
- Usage of texts, notes, or digital devices is forbidden.
- Questions can have multiple correct answers.
- Wrong answers are attributed a negative score (half the score of a correct answer).
- Completely blacken the box beside a correct answer **in the answer sheet ONLY**. Answers indicated in other sheets will be ignored. Use a white-out marker to correct and unmark previously marked boxes.

  DOs: □ ■ **DON'Ts**: ☒ ☒ ☑ **Absolute DON'Ts**: ◉ NO
- No questions are allowed during the exam. Comments can be included in the last box. If your comment refers to a question, make sure you reference the question (for example, "In question 1, answer B, ...").
- It is your responsibility to write in a clear and understandable way.
- **DO NOT** black out the boxes in the top right corner of open text questions.
- **To withdraw** from the exam, completely blacken out the box associated to Question 11 in the answer sheet.
- Please encode your **codice persona** (8 digits) below, blacking out one digit per column. For instance, if your codice persona is 12345678, the "1" digit must be blacked out in the leftmost column, and the "8" digit must be blacked out in the rightmost column.

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

- Write your first and last names in the box below.

Last name: ....

First name: ....

Corrected

Answers:

Question 1: A ■ C ■ ■ F

Question 2: ■ B C ■ ■ F

Question 3: A ■ C ■ ■ F

Question 4: ■ B ■ ■ ■ F

Question 5: ■ ■ ■ ■ E F

Question 6: A B ■ ■ E F

Question 7: A B ■ ■ ■ F

Question 8: A ■ ■ ■ ■ F

Question 9: ■ ■ ■ ■ ■ F

Question 10: ■

....

....

....

....

....

....

....

....

....

....

Question 11: ■