Mock Exam

Exercise 1

A manufacturer records ten correlated sensor readings from 400 wearable devices. The devices occasionally enter one of three hidden operating modes. You are given the data frame `wearable_sensors.csv`.

1. Perform PCA. Plot the scree curve and the cumulative variance curve. How many components explain at least 95% of the variance? [Upload the scree curve to MS forms]
2. Plot the loadings of the first 2 PCs and interpret them. [Upload the plot to MS forms]
3. Using the retained PCs, compare hierarchical clustering and K–means.
   (a) For the K-Means, use the elbow method to select the optimal number of clusters.
   (b) For the hierarchical clustering, use the average linkage and cut the dendrogram to obtain the same number of clusters as K-means.

   For K-means: Report the within cluster sum of square associated with the optimal number of cluster.

   Report the silhouette scores for both K-means (with optimal number of clusters) and hierarchical clustering, and comment on which algorithm performs better.

   Fit Gaussian Mixture Models with 1–5 components, select the best model via BIC, and compare its cluster assignments with K–means using the adjusted Rand index.

Exercise 2

The file `credit_default.csv` contains 15 applicant features and the response `default` (1 = defaulted).

Use the following command to split the dataset into training a test set:

```
 X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.3, random_state=42, stratify=y)
```

Scale the features to have zero mean and unit variance.

1. Fit a standard logistic regression. Estimate test accuracy via 10–fold cross–validation. Compare the estimated test accuracy with the actual test accuracy.

2. Repeat with ridge and lasso penalties. Use 10-fold cross-validation to estimate the parameters. Report the best C values and compare test accuracies.

3. Train a random forest (300 trees, `oob_score=True`). Report OOB accuracy, test accuracy, and plot the top–10 feature importances.

4. Report the adjusted rand index between the top–10 features according to random forest and the features selected by Lasso with the optimal C parameter.

5. Plot ROC curves and compute AUC for all the models. Which model would you deploy and why? (max 5 sentences).