|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **HAZAR UNIVERSITY**  **FACULITY OF ENGINEERING**  **MIDTERM QUESTIONS** | | | **EXAM** | **Type** | 1st Midterm Exam 2nd Midterm Exam  Final | | | |
| **Date** | 13.04.2025 | | **Location** | Online (Take Home) |
| **Time** | 09:00 AM – 09:00 AM | | **Time** | 24 hour |
| **COURSE** | **Semester** | 2024-2025 Spring Semester | | | **STUDENT** | **Name Surname** | **Sabuhi Gasimzada** | |
| **Code, Name** | Artificial Intelligence | | | **Number** | **B.01.631.21.044** | |
| **Teaching staff** | Assoc. Prof. Dr. Adem Tekerek | | | **Signature** | **S.Q** | |

|  |
| --- |
| **Total Score of the Student:** |

QUESTIONS

### **Question 1.** **Feature Selection (6 Point)**

**Dataset:** Heart Disease UCI Dataset (<https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data> )

**Assignment:**  
You are provided with a dataset containing various medical measurements to predict the presence of heart disease.

1. After necessary preprocessing, apply **at least three different feature selection techniques** (e.g., Recursive Feature Elimination, Mutual Information, L1 Regularization).
2. Train a logistic regression model using the selected features from each method and compare their performance using metrics such as AUC, Accuracy, and F1-score.
3. Analyze and interpret the results: Which method yields the most effective feature subset and why?

### **Question 2.** **Linear Regression (6 Point)**

**Dataset:** Ames Housing Dataset – (<https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset> )

**Assignment:**  
You are tasked with building a model to predict house prices based on numerous numeric and categorical features.

1. Investigate multicollinearity using techniques like Variance Inflation Factor (VIF), and explain how you handled it.
2. Train both **Ridge** and **Lasso** regression models.
3. Compare their performance using metrics like RMSE and R², and analyze the learned coefficients.
4. Discuss which method is more appropriate for this problem, considering performance and interpretability.

### **Question 3.** **Logistic Regression (6 Point)**

**Dataset:** Give Me Some Credit – (<https://www.kaggle.com/c/GiveMeSomeCredit> )

**Assignment:**  
Using this dataset of financial and personal attributes, predict whether a person is at risk of defaulting on their credit.

* a) Train a logistic regression model and evaluate it using AUC, Precision, Recall, and F1-score.
* b) Address class imbalance using techniques such as SMOTE or class weighting.
* c) Calculate and interpret the **odds ratios** of each feature. Which features increase or decrease the odds of default?

### **Question 4.** **Support Vector Machines (SVM) (6 Point)**

**Dataset:** [Spambase – UCI ML Repository](https://archive.ics.uci.edu/ml/datasets/spambase) (<https://archive.ics.uci.edu/dataset/94/spambase> )

**Assignment:**  
Your task is to build a spam classifier using email data.

* a) Train two SVM classifiers: one using a **linear kernel**, and the other using an **RBF kernel**.
* b) Compare their performance using Accuracy, F1-score, and AUC.
* c) Use **Grid Search** to optimize hyperparameters (C and gamma) for the RBF kernel.
* d) Evaluate which kernel is more suitable for the dataset and explain why.

### **Question 5.** **Decision Tree (8 Point)**

**Dataset:** Pima Indians Diabetes Dataset (<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> )

**Assignment:**  
Develop a model to predict whether a patient has diabetes based on health measurements.

* a) Train both a **Decision Tree** and a **Random Forest** classifier.
* b) Analyze the effect of varying tree depth and other hyperparameters on overfitting and underfitting.
* c) Extract and visualize feature importances from each model.
* d) Compare the models in terms of accuracy, generalization, and interpretability.

### **Question 6.** **Unsupervised Learning (8 Point)**

**Dataset:** Mall Customers Dataset (<https://www.kaggle.com/datasets/simtoor/mall-customers> )

**Assignment:**  
Perform customer segmentation using unsupervised learning techniques.

* a) Apply both **K-Means** and **Hierarchical Clustering** algorithms.
* b) Determine the optimal number of clusters using the **Elbow Method** and **Silhouette Score**.
* c) Visualize the clusters and compare the results of both methods.
* d) Interpret each customer segment and suggest how this segmentation could inform marketing strategy.