

University of Tehran  
Department of Electrical and Computer Engineering  
**Data Science - Final Assignment**

Lead Teaching Assistant Team

Spring 2025

**Analyzing Global Tech Talent Migration:** You are provided with a dataset of 50,000 tech professionals. The goal is to predict `Migration_Status` (1 if they migrated to another country for work, 0 otherwise). Features include `GitHub_Activity`, `Research_Citations`, `Industry_Experience`, and categorical data such as `Education_Level`.

## 1 Question 1: Advanced Data Engineering & SQL

### Part A: Time-Series Trends via Window Functions

Write a single SQL query to compute the 3-year moving average of `Research_Citations` for each user, partitioned by `Country_Origin`, and rank users inside each country by this moving average.

### Part B: Diagnostic Identification of Data Leakage

Given proposed features below, identify which would introduce leakage and justify your answer:

- `Years_Since_Degree`
- `Visa_Approval_Date`
- `Last_Login_Region`
- `Passport_Renewal_Status`

## 2 Question 2: Statistical Inference & Linear Models

### Part A: Elastic Net Derivation

Given

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda_1 \sum_{j=1}^n |\theta_j| + \frac{\lambda_2}{2} \sum_{j=1}^n \theta_j^2,$$

derive  $\nabla J(\theta)$  with respect to  $\theta_j$ , and explain subgradient behavior at  $\theta_j = 0$ .

## Part B: Interpreting Regression Outputs

For `GitHub_Activity`:

- Coefficient: 0.52
- P-value: 0.003
- 95% CI: [0.18, 0.86]

State whether it is statistically significant under  $H_0 : \beta = 0$  and interpret both p-value and CI.

## 3 Question 3: Optimization & Gradient Descent

Explain the ravine phenomenon and compare Momentum vs Adam for optimization in this geometry.

## 4 Question 4: Non-Linear Models & Kernels

### Part A: SVM RBF Parameterization

If RBF-SVM overfits, how should  $\gamma$  be adjusted and why?

### Part B: Cost-Complexity Pruning

Given  $R_\alpha(T) = R(T) + \alpha|T|$ , explain how  $\alpha$  controls pruning and bias-variance behavior.

## 5 Question 5: Unsupervised Learning

### Part A: PCA Explained Variance

For a  $3 \times 3$  covariance matrix, show how to compute explained variance ratios for PC1 and PC2 and interpret eigenvalues.

### Part B: K-Means Elbow Method

Give a mathematical/geometric argument for why WCSS elbow is a useful heuristic for choosing  $K$ .

## 6 Question 6: Capstone Explainability

Using SHAP for an XGBoost model, explain the difference between `base_value` and `output_value` for a high-citation candidate predicted as No Migration.