IND522
Advanced Statistical Modelling

Final Exam - Fall 2025

Professor Orhan FEYZİOĞLU

Due: 12.01.2025

Instructions

Welcome to the culmination of your journey in Advanced Statistical Modelling! This final exam is your opportunity to showcase your mastery of statistical techniques by diving deep into a real-world dataset tailored just for you. Think of it as your personal statistical adventure: uncovering hidden patterns, building predictive models, and drawing meaningful insights that could inform decisions in fields from healthcare to environmental science. Your mission: Analyze your assigned dataset comprehensively, tackling regression and classification challenges where they naturally arise, and demonstrate how statistical modeling can illuminate complex problems.

- **Dataset Assignment and Focus:** Each of you has received a unique dataset from the list provided, complete with its domain context, size, variable types, and suggested tasks for regression (predicting continuous outcomes) and classification (categorizing discrete outcomes). Tailor your analysis to the dataset's strengths—whether it's predicting survival rates on the Titanic or forecasting energy efficiency in buildings. If your dataset lacks a natural classification or regression task, derive one logically from the variables (e.g., binning a continuous variable into categories). Emphasize both exploratory insights and predictive power, linking your findings back to the dataset's real-world implications.

- **Tools and Analysis Requirements:** Harness the power of statistical software like MATLAB (or equivalents such as R or Python if you're more comfortable, but justify your choice) to conduct all computations. Your submission should weave together clear, narrative explanations of your methods and results, supported by code snippets, output logs, and vivid visualizations (e.g., plots, charts, and tables). Don't just crunch numbers—interpret them! For instance, explain why a scatter plot reveals multicollinearity or how a confusion matrix highlights model biases. Aim for reproducibility: Ensure your code is well-commented and could be rerun by a peer.

- **Academic Integrity and Independence:** This is your solo quest—no collaborations, consultations, or external aids beyond cited references. All work must be original, reflecting your own understanding and creativity. Plagiarism or unauthorized assistance will result in severe penalties, as per university policy. Remember, true learning comes from grappling with the data yourself!

- **Submission Format and Deadline:** Compile everything into a single, polished PDF or DOCX file for seamless review. Structure it logically: Start with an executive summary of your dataset and key findings, followed by sections addressing each exam question, and end with a discussion of broader implications. Include an appendix for full code scripts, raw outputs, and any supplementary materials. Submit via MS Teams by January 15, 2026. Please plan ahead to avoid last-minute statistical storms.

Exam Questions

(1) Exploratory Data Analysis (20 Points)

- Summarize your dataset with descriptive statistics and plots:
  (a) Univariate analysis: For each key variable (both numerical and categorical), compute summary statistics such as mean, median, mode, standard deviation, quartiles, and range for numerical variables; frequency counts and percentages for categorical variables. Generate histograms or density plots for numerical variables, and bar charts or pie charts for categorical ones to visualize distributions.
  (b) Bivariate analysis: Explore relationships between pairs of variables using scatter plots for numerical-numerical pairs (include trend lines if applicable), box plots or violin plots for numerical-categorical pairs, and contingency tables or stacked bar charts for categorical-categorical pairs. Compute correlation coefficients (e.g., Pearson for numerical, or Cramér's V for categorical) where relevant.
- Discuss patterns, trends, and anomalies: Identify any noticeable trends (e.g., positive/negative correlations), patterns (e.g., clustering in scatter plots), or anomalies (e.g., unexpected spikes in histograms). Explain potential real-world implications based on the dataset's domain.
- Identify potential challenges, such as missing data or outliers: Quantify missing values (e.g., percentage per variable), detect outliers using methods like IQR or z-scores, and suggest handling strategies (e.g., imputation, removal). Discuss how these issues might affect downstream analyses.

(2) Probability and Sampling (15 Points)

- Select a key numerical variable from your dataset:
  (a) Assume it follows a known distribution (e.g., normal, exponential, Poisson, or another appropriate based on your data). Estimate its parameters using Maximum Likelihood Estimation (MLE). Show the likelihood function, derive the estimators if possible, and compute them using your software. Justify your distribution choice with goodness-of-fit tests (e.g., Kolmogorov-Smirnov or QQ plots).
  (b) Simulate a sample of size 1,000 based on the estimated distribution parameters. Compare the empirical distribution (from the simulated data) to the theoretical one using overlaid histograms, empirical CDF vs. theoretical CDF plots, and statistical tests (e.g., chi-square goodness-of-fit).
- Construct a 95% confidence interval for the variable's mean: Use both parametric (e.g., t-interval if normal) and non-parametric (e.g., bootstrap) methods. Compare the intervals and discuss any assumptions violated in your data.

(3) Dimensionality Reduction (15 Points)

- If your dataset contains multiple numerical variables:
  (a) Perform Principal Component Analysis (PCA) and report the variance explained by the first three components, including individual and cumulative proportions. Provide a scree plot to justify the number of components retained (e.g., using the elbow method or Kaiser criterion).

(b) Visualize the first two principal components using a scatter plot (color-coded by a categorical variable if available, such as the target class). Interpret the results by examining loadings: identify which original variables contribute most to each component and explain what these components might represent in the dataset's context.

- Discuss how dimensionality reduction could impact model performance: Address benefits like reduced computational cost, multicollinearity mitigation, and noise reduction, as well as drawbacks such as information loss or interpretability challenges. Suggest when PCA might be preferable over other methods (e.g., t-SNE for non-linear reduction).

(4) Classification (20 Points) Develop a classification model to predict a categorical outcome (use the canonical or derived classification task from your assigned dataset):

(a) Split the data into training (70%) and testing (30%) sets using stratified sampling to maintain class balance. Preprocess the data as needed (e.g., handle missing values, encode categoricals, scale numericals).

(b) Train at least two classification models (e.g., logistic regression, decision tree, SVM, or random forest) and evaluate them using precision, recall, F1-score, accuracy, ROC-AUC (for binary), and a confusion matrix. Compare models via cross-validation (e.g., 5-fold) on the training set and select the best one for final testing.

(c) Discuss overfitting and suggest techniques to address it: Identify signs of overfitting (e.g., high training accuracy vs. low test accuracy). Propose methods like regularization (e.g., L1/L2 in logistic regression), pruning (for trees), early stopping, or ensemble techniques (e.g., bagging/boosting). Apply one technique and report improved results.

(5) Regression (20 Points) Develop a regression model to predict a numerical outcome (use the canonical or derived regression task from your assigned dataset):

(a) Fit at least two regression models (e.g., linear regression, generalized linear model, or ridge/lasso for regularization) after preprocessing (e.g., feature selection, transformations for non-linearity). Evaluate performance using R^2, adjusted R^2, mean squared error (MSE), root mean squared error (RMSE), and residual plots (e.g., residuals vs. fitted, QQ plot for normality).

(b) Interpret the coefficients of the model and discuss their significance: For the best model, report p-values, confidence intervals, and effect sizes. Explain practical meaning (e.g., "a one-unit increase in X is associated with a Y-unit change in the outcome, holding others constant"). Test for multicollinearity (e.g., VIF) and address if necessary.

(c) Use cross-validation to assess model robustness and compare its performance across folds: Implement k-fold cross-validation (e.g., 10-fold) and report average metrics (e.g., mean CV-MSE). Compare to a baseline model (e.g., mean prediction) and discuss variability across folds, suggesting improvements like feature engineering.

(6) Monte Carlo Methods (10 Points) Use Monte Carlo simulation to solve an inferential problem related to your dataset:

(a) Clearly define the problem (e.g., estimating a population mean under uncertainty, testing a hypothesis about a parameter, or approximating a complex integral related to your model's predictions). Specify assumptions and the simulation setup (e.g., number of iterations, random variable generation).

(b) Run the simulation (e.g., 10,000 iterations) and compare the results to traditional statistical methods (e.g., t-test for means or analytical confidence intervals). Visualize results with histograms of simulated values and overlay the empirical estimate.

(c) Discuss the advantages and limitations of the Monte Carlo approach: Highlight benefits like flexibility for non-standard distributions or complex models, and computational efficiency for large samples. Note limitations such as dependence on random seed, computational intensity, and potential bias from poor simulation design. Suggest when it might outperform parametric methods in your dataset's context.

Dataset Assignment

1. **Titanic Survival Dataset** - *Melike MİRAN*
   Domain: Social Science, History, Maritime Safety; **Source:** Kaggle "Titanic: Machine Learning from Disaster" (originally British Board of Trade passenger records); **Size/Type:** 890 rows, 12–15 features depending on preprocessing (mix of numeric & categorical); **Typical Variables:** Passenger class, sex, age, family relations, fare, embarkation port; **Regression Task:** Predict fare or age from socio-demographic and travel-related variables; **Classification Task:** Predict survival (Yes/No) based on passenger characteristics.

2. **Heart Disease Dataset** - *Bedirhan DEMİRCİ*
   Domain: Medicine, Cardiology, Healthcare Analytics; **Source:** UCI Machine Learning Repository (e.g., Cleveland dataset); **Size/Type:** 303 patients, 13–14 features (mixture of numeric & categorical); **Typical Variables:** Age, sex, chest pain type, blood pressure, cholesterol, ECG results, max heart rate; **Regression Task:** Predict cholesterol level, maximum heart rate, or a continuous risk score from the other features; **Classification Task (canonical):** Predict presence/absence of heart disease, possibly with graded severity classes.

3. **Wine Quality Dataset** - *Büşra SAVRAN*
   Domain: Food Science, Chemistry, Sensory Analysis; **Source:** UCI Machine Learning Repository ("Wine Quality" red/white variants); **Size/Type:** 1,599 red wine or 4,898 white wine samples; 11 physicochemical features, 1 quality score (0–10); **Typical Variables:** Acidity measures, residual sugar, chlorides, sulphates, alcohol; **Regression Task (canonical):** Predict wine quality score as a continuous variable from chemical properties; **Classification Task (derived):** Group wines into low/medium/high quality or good vs. not good based on the score threshold.

4. **Student Performance Dataset** - *Sinan Doğukan YILMAZ*
   Domain: Education, Social Science; **Source:** UCI Machine Learning Repository ("Student Performance" from Portuguese schools); **Size/Type:** 649 students, 30 attributes (numeric + categorical); **Typical Variables:** Study time, past failures, family support, absences, grades from previous terms; **Regression Task (canonical):** Predict final grade as a continuous score; **Classification Task (derived):** Classify students as pass/fail, low/medium/high achievers, or at-risk vs. not at-risk.

5. **Pima Indians Diabetes Dataset** - *Mahmut Selim ÇOBAN*
   Domain: Public Health, Epidemiology; **Source:** UCI "Pima Indians Diabetes"; **Size/Type:** 768 instances, 8 numeric features; **Typical Variables:** Pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age; **Regression Task:** Predict glucose or BMI using the remaining features; **Classification Task (canonical):** Predict whether a patient has diabetes (Yes/No).

6. **California Housing Dataset** - *Zülal DAĞLI*
   Domain: Economics, Urban Planning, Real Estate; **Source:** Originally StatLib; bundled with scikit-learn as fetch_california_housing; **Size/Type:** 20,640 instances, 8 numerical predictors; **Typical Variables:** Median income, house age, rooms per household, population, latitude/longitude; **Regression Task (canonical):** Predict median house value for a block group; **Classification Task (derived):** Turn house value into high vs. low value (e.g., above/below median or specific threshold) and classify.

7. **Adult Income Dataset** - *Caner SEZGİN*
   Domain: Economics, Labour Market, Sociology; **Source:** UCI Machine Learning Repository ("Adult" or "Census Income"); **Size/Type:** 48,842 records, 14 features (mixed categorical & numerical); **Typical Variables:** Age, education, occupation, hours per week, marital status, native country; **Regression Task:** Predict hours worked per week or alternatively a derived continuous score (e.g., total income if constructed); **Classification Task (canonical):** Predict whether income is $>$ 50K USD/year vs. $\leq$ 50K.

8. **Energy Efficiency Dataset** - *Alp Burak KAN*
   Domain: Building Physics, Energy Engineering; **Source:** UCI Machine Learning Repository ("Energy efficiency"); **Size/Type:** 768 samples, 8 input features, 2 continuous targets; **Typical Variables:** Wall area, roof area, overall height, glazing area, orientation, relative compactness; **Regression Task (canonical):** Predict heating load or cooling load as a function of building design features; **Classification Task (derived):** Convert loads into low/medium/high energy consumption or efficient vs. inefficient building categories.

9. **Bank Marketing Dataset** - *Meral GENÇ*
   Domain: Marketing Analytics, Banking, CRM; **Source:** UCI "Bank Marketing" (Portuguese bank direct marketing campaigns); **Size/Type:** 45,211 observations, 16–20 features (mostly categorical); **Typical Variables:** Contact type, campaign duration, previous contacts, economic indicators, client attributes; **Regression Task:** Predict campaign duration or customer balance from other attributes; **Classification Task (canonical):** Predict if a client will subscribe to a term deposit (yes/no).

10. **Breast Cancer Wisconsin (Diagnostic) Dataset** - *Şeyda KANTAR*
    Domain: Medical Diagnosis, Oncology; **Source:** UCI / scikit-learn (load_breast_cancer); **Size/Type:** 569 instances, 30 numeric features; **Typical Variables:** Mean, worst and standard error of radius, texture, perimeter, area, smoothness of cell nuclei; **Regression Task:** Predict one continuous cell characteristic (e.g., radius_mean) using the remaining features; **Classification Task (canonical):** Predict diagnosis: malignant vs. benign.

11. **Abalone Dataset** - *Ecem Sıla AŞICI*
Domain: Marine Biology, Fisheries; **Source:** UCI "Abalone" dataset; **Size/Type:** 4,177 rows, 8 continuous features + 1 categorical (sex); **Typical Variables:** Length, diameter, height, whole weight, shucked weight, shell weight, sex; **Regression Task (canonical):** Predict age (via number of rings) from physical measurements; **Classification Task (derived):** Classify abalones as young / adult / old or old vs. not old based on age thresholds.

12. **Concrete Compressive Strength Dataset** - *Gül Mine AK*
Domain: Civil Engineering, Materials Science; **Source:** UCI "Concrete Compressive Strength"; **Size/Type:** 1,030 observations, 8 input variables, 1 continuous target; **Typical Variables:** Cement, slag, fly ash, water, superplasticizer, coarse and fine aggregates, age (days); **Regression Task (canonical):** Predict compressive strength of concrete; **Classification Task (derived):** Categorize concrete into low/medium/high strength or structural vs. non-structural based on thresholds.

13. **Algerian Forest Fires Dataset** - *Ekin MADAN*
Domain: Environmental Science, Wildfire Risk; **Source:** UCI "Algerian Forest Fires Dataset"; **Size/Type:** 244 instances, 13 meteorological and fire weather attributes + fire/no-fire label; **Typical Variables:** Temperature, relative humidity, wind speed, rain, Fine Fuel Moisture Code, Drought Code, Fire Weather Index; **Regression Task:** Predict a continuous fire index or burned area (depending on chosen version/label) from weather variables; **Classification Task (canonical):** Predict fire vs. no fire days.

14. **Medical Cost Personal Dataset (Insurance Charges)** - *Büşra AVCU*
Domain: Health Economics, Insurance Pricing; **Source:** Kaggle "Medical Cost Personal Datasets"; **Size/Type:** 1,338 observations, 7 features + continuous charge variable; **Typical Variables:** Age, sex, BMI, children, smoker status, region; **Regression Task (canonical):** Predict medical charges (charges) based on demographic and lifestyle factors; **Classification Task (derived):** Define high-cost vs. low-cost patients (e.g., top quartile of charges) and classify.

15. **Bike Sharing Dataset** - *Melisa YILDIRIM*
Domain: Transportation, Smart Cities; **Source:** UCI "Bike Sharing" or Kaggle derivative (Capital Bikeshare); **Size/Type:** Daily data: 731 rows, 13 features; Hourly data: 17,000+ rows, 16 features; **Typical Variables:** Weather conditions, season, holiday indicator, time of day, registered/casual counts; **Regression Task (canonical):** Predict total number of rented bikes (cnt) for a given hour/day; **Classification Task (derived):** Classify demand as high vs. low or into multiple categories based on cnt quantiles.

16. **Online Shoppers Purchasing Intention Dataset** - *Emre BİLİR*
Domain: E-commerce Analytics, Web Behaviour; **Source:** UCI "Online Shoppers Purchasing Intention"; **Size/Type:** 12,330 sessions, 17 features (numeric + categorical); **Typical Variables:** Administrative/product-related pages visited, visit duration, bounce rate, exit rate, month, visitor type; **Regression Task:** Predict total session duration or number of pages visited from other variables; **Classification Task (canonical):** Predict whether the session ends with a purchase (Revenue True/False).

17. **Rain in Australia (RainTomorrow) Dataset** - *Merve Nur KARABULUT*
Domain: Meteorology, Climate, Environmental Modelling; **Source:** Popular Kaggle dataset "Weather in Australia"; **Size/Type:** 145,000+ daily records, 20–25 features (after cleaning); **Typical Variables:** Min/Max temperature, rainfall, wind speed/direction, humidity, pressure, cloud, location, date; **Regression Task:** Predict tomorrow's rainfall amount or maximum temperature using today's conditions; **Classification Task (canonical):** Predict RainTomorrow (Yes/No).

18. **German Credit Dataset (Statlog Credit Risk)** - *Fatih Emre DURGUN*
Domain: Finance, Credit Risk, Banking; **Source:** UCI "Statlog (German Credit Data)"; **Size/Type:** 1,000 instances, 20 attributes (categorical + numerical), binary target; **Typical Variables:** Account status, credit history, loan purpose, credit amount, duration, personal status, property, age; **Regression Task:** Predict credit amount or loan duration from the other applicant characteristics; **Classification Task (canonical):** Predict good vs. bad credit risk.