Final Project — Statistical Models for Artificial Intelligence

From Classical Statistics to Artificial Intelligence: understanding, modeling, and predicting with data.

We live in an era defined by an explosion of data. Every second, massive amounts of information are generated across diverse fields such as health, economics, education, environment, and social media. In this context, Statistics plays a fundamental role as the science that allows us to extract meaningful knowledge from data, transforming raw numbers into evidence-based insights and informed decision-making.

Classical Statistics provides the foundation for understanding variability, identifying relationships, and quantifying uncertainty.

Through techniques such as descriptive analysis, inference, regression, and multivariate modeling, we can describe patterns, test hypotheses, and predict behaviors.

These methods are essential for making rational, evidence-driven decisions — a principle that also underlies modern Artificial Intelligence.

With the rise of Machine Learning and Artificial Intelligence (AI), statistical models have gained new dimensions:

- Regression analysis evolved into supervised learning algorithms;
- Factor analysis and clustering became key techniques for dimensionality reduction and data segmentation in AI;
- Probabilistic models form the basis of classifiers such as Naive Bayes;
- And core statistical concepts such as fit, error, significance, and variance remain crucial to understanding, validating, and interpreting AI models.

Therefore, the connection between Statistics and AI is deep and complementary: Statistics provides theoretical rigor, interpretability, and explainability, while AI extends our capacity for prediction, automation, and large-scale data analysis.

This final project invites students to integrate both perspectives — using Statistics as a foundation and Python as a computational tool — to develop intelligent, critical, and well-supported analyses. Beyond the mechanical application of methods, the true goal is to understand data behavior, extract meaningful insights, and promote explainable and responsible AI grounded in solid statistical reasoning.

General Objective

The aim of this project is to apply, in an integrated way, the concepts, techniques, and statistical methodologies taught in the course, using **Python** as a tool for data analysis and modeling in the context of **Artificial Intelligence**.

The project should follow all stages of the analytical cycle:

Collect → Prepare → Explore → Model → Conclude

Structure and Phases of the Project

Phase 1 — Data Collection and Selection (Initial Research)
Objective: Choose and justify a dataset that is appropriate for the chosen research topic.

Students should:

1. Select a real-world dataset related to an area of personal or professional interest, such as:
    - Health: Kaggle "Heart Disease", WHO Open Data, Global Health Observatory, Our World in Data.
    - Economy and Finance: World Bank Open Data, IMF Data, Yahoo Finance API, OECD Data.
    - Education: UNESCO Data, PISA datasets.
    - Environment and Energy: Climate Data Portal, Eurostat Environment, NASA Earth Data.
    - Society and Behavior: Kaggle "Mental Health", Pew Research, European Social Survey.

Examples of open data sources:

- Kaggle: https://www.kaggle.com/datasets
- UCI Machine Learning Repository: https://archive.ics.uci.edu/?utm_source=chatgpt.com
- Google Dataset Search: https://datasetsearch.research.google.com/
- Data.gov: https://data.gov/
- Our World in Data. https://ourworldindata.org/

Describe briefly:

- The topic: What is the main subject or problem the project will address?
- The origin of the dataset and its source: Indicate where the data comes from (organization, platform, or repository) and provide the access link.
- Justification for the choice: Explain why the dataset was selected — its relevance, interest, and applicability to the research or field of study.

- Type of data: Specify whether the variables are quantitative, qualitative, or mixed, and briefly describe the structure of the dataset (number of observations, variables, etc.).

Expected output:
A short descriptive text presenting the dataset, including the theme, source, rationale for selection, and a verifiable link to the data source.

Phase 2 — Data Preparation and Cleaning

Objective: Prepare the dataset for statistical analysis and modeling.

Students should:

1. Identify the variables:
    - Define dependent and independent variables;
    - Classify each as numerical, categorical, or ordinal.
2. Handle missing and inconsistent data:
    - Replace, remove, or impute missing values appropriately;
    - Correct typing errors, inconsistencies, or scaling mistakes.
3. Detect and treat outliers:
    - Use visualization tools such as boxplots;
    - Apply statistical methods (e.g., Z-score, IQR).
4. Transform variables, if necessary:
    - Apply normalization or standardization;
    - Encode categorical variables using one-hot encoding or similar techniques;
    - Create derived variables when useful for analysis.

Expected output:

A Python notebook containing well-commented code and a clean, finalized dataset ready for analysis.

Phase 3 — Exploratory Analysis and Descriptive Statistics
Objective: Understand the behavior of the variables and characterize the dataset.

Students should:

1. Compute descriptive statistics — means, standard deviations, medians, quartiles, etc.;
2. Create visual representations, such as:
    - Histograms, boxplots, scatter plots, and heatmaps;

3. Explore bivariate relationships:
    - Use Pearson or Spearman correlations (for numerical variables);
    - Build contingency tables for categorical relationships (if applicable);
4. Write a descriptive analysis interpreting the main results and insights.

Expected output:
An intermediate report including visualizations and descriptive summaries (to be integrated into the Python notebook).

Phase 4 — Application of Statistical Models and Techniques

Objective: Apply and interpret statistical models appropriate to the research problem.

Students should select and apply at least two different techniques from those studied throughout the course, depending on the type of variables and research objectives.

Examples of Possible Approaches

Research Objective	**Possible Techniques**
Compare groups	*t-test*, One-way or Multi-factor ANOVA
Explain / predict a continuous variable	Simple or Multiple Linear Regression
Predict a categorical variable	Binary or Multinomial Logistic Regression, Naive Bayes
Reduce dimensionality	PCA (Principal Component Analysis) or Factor Analysis
Cluster / group individuals	Clustering (K-Means or Hierarchical)
Classify predefined groups	Discriminant Analysis, Decision Trees
Improve model generalization	Regularization (Ridge, Lasso, ElasticNet)

Each group must:

1. Justify the choice of methods, explaining why each technique was selected for the research problem;
2. Present results, including tables, model outputs, and evaluation metrics;
3. Interpret the findings, discussing coefficients, statistical significance, and conclusions.

Expected Deliverable:

A Python notebook with clearly commented code, result tables, and a written interpretation of the findings.

Phase 5 — Writing and Presentation of the Final Report

Objective: To synthesize the entire analytical process and the obtained results in the form of a technical–scientific article.
The report should clearly demonstrate the students' ability to design, analyze, interpret, and communicate statistical findings in a coherent and professional manner.

Recommended Structure of the Report

1. Introduction

- Contextualization of the research topic and problem;
- Definition of the general objective and specific hypotheses;
- Relevance of the study in the context of Artificial Intelligence and data analysis.

2. Methodology

- Description of the chosen dataset (source, number of observations, variables, etc.);
- Data preparation process: cleaning, transformations, and justifications for preprocessing steps;
- Statistical models and tests applied, with rationale for each selection.

3. Results

- Presentation of descriptive and inferential statistics;
- Visualizations (plots, graphs, heatmaps, etc.) to illustrate findings;
- Model outputs (e.g., regression summaries, confusion matrices, metrics);
- Clear and concise interpretation of numerical and graphical results.

4. Discussion

- Critical reflection on the results and their meaning;
- Comparison with findings from existing literature or related studies (if applicable);
- Identification of methodological limitations and suggestions for further research or improvements.

5. Conclusions

- Summary of main findings and insights;
- Implications for practice, decision-making, or future studies;
- Relation of the work's outcomes to the objectives and hypotheses defined.

6. References

- Citation of all data sources (including dataset links);
- Bibliography of books, articles, and websites used throughout the project, formatted in APA style or equivalent.

Format and Deliverables

- Final Report: Article-style document (5–10 pages, excluding references and appendices);
- Python Notebook: Complete, well-commented .ipynb file showing all analyses, visualizations, and results;
- Group Work: The project may be developed in teams of 2–3 students;
- Final Submission:
  → One PDF document (report) + one Jupyter Notebook (.ipynb) file.

Evaluation Criteria

Criterion	Description	Weight
Relevance and quality of the dataset	Adequacy of the chosen data to the research topic; justification of its importance and applicability.	**10%**
Data preparation and cleaning	Correct handling of missing values, outliers, transformations, and variable coding; reproducible preprocessing steps.	**15%**
Exploratory analysis and interpretation	Depth of descriptive statistics, graphical representations, and preliminary insights drawn from the data.	**15%**
Correct application of statistical methods	Proper choice and implementation of at least two appropriate statistical or machine learning methods.	**25%**
Interpretation and discussion of results	Accuracy, critical thinking, and clarity in explaining the obtained outcomes; connection to research objectives.	**20%**
Report clarity and visual/code quality	Organization, readability, structure, and aesthetics of the report and Python code (comments, plots, and layout).	**15%**

Submission Deadline: 🗓️ ***December 31, 2025***

*All deliverables (**report in PDF + Python notebook .ipynb**) must be submitted by **23:59 (local time)** on the due date.*
*Late submissions will **not be accepted** unless previously justified and approved by the instructor.*

⚠️ ***Important:*** *Ensure that your final version is properly named and includes all authors' names.*
Example file naming convention:
AI_StatisticalModels_GroupX_Lastname1_Lastname2.pdf
AI_StatisticalModels_GroupX_Lastname1_Lastname2.ipynb