95-891 Introduction to Artificial Intelligence

Homework 1: Clustering and Classification

Due 11:59 PM EST September 16, 2025

Overview

On County Health Rankings, the University of Wisconsin's Population Health Institute has published rankings of all 3142 counties in the US by their health outcomes and behaviors. In this assignment, you will apply clustering and classification techniques to the data from 2025 to determine if there are interesting groupings of counties with similar health outcomes and behaviors and develop a predictive model to see which factors influence health outcomes. Individuals and governments can use such a model to improve public health.

This assignment must be completed individually. General discussion with other students is permitted, but discussing or sharing code or text submitted for grading is not and will be considered an Academic Integrity Violation.

We strongly discourage the use of generative AI for this assignment to help build your machine learning skills. If you must use generative AI, you need to submit a .pdf file showing your interaction with the generative AI tool, including prompts you entered and results. Use of generative AI tools without appropriate documentation will be considered an Academic Integrity Violation.

Tasks

1. **Data Source and Preparation:**
    - The data file for analysis is available at: 2025 CHR CSV Analytic Data.
    - Documentation for measures in the data file can be found here:
        - 2025 CHR CSV/SAS Analytic Data Documentation
        - 2025 Data Dictionary (PDF)
    - Use only the "Select Measures" from the 2025 Measures for modeling, not the "Additional Measures."
2. **Feature Selection:**
    - **Use only columns with the words "rawvalue" (e.g., `v001_rawvalue`) as features.** Exclude columns related to numerators, denominators, confidence intervals, etc.

   - Include as predictors in your models the variables under “Community Conditions” factors (e.g., “Access to exercise opportunities”) and exclude “Population Health and Well-Being” outcome variables (e.g., premature death, poor or fair health). Using “Population Health and Well-being” as input features would constitute data leakage and invalidate results. The categorizaqtion is shown in the appendix to this assignment and is based on the 2025 Technical Documentation.
   - You may omit Community Conditions variables that are not actionable through public policy, but you must explain your assumptions.
3. **Exploratory Data Analysis (EDA):**
   - Summarize your steps for handling missing data, identifying outliers, and computing summary statistics. Visualizations (e.g., histograms, boxplots) are encouraged.
4. **Clustering:**
   - Identify noteworthy groupings of counties with similar health outcomes and behaviors using unsupervised learning techniques.
   - Justify the method used to determine the number of clusters (e.g., elbow method, silhouette score).
5. **Supervised Learning Models:**
   - Develop **two supervised learning models** to predict premature death, defined as years of potential life lost before age 75 per 100,000 population.
   - List the **five most important factors** influencing premature death as shown by your models.
   - Evaluate the accuracy of both models and justify which one is more accurate.
6. **Recommendations:**
   - Provide actionable recommendations to reduce premature death in Allegheny County based on your findings.

Submission Guidelines

1. Submit the following to Gradescope:
   - A Jupyter Notebook named `IAI_HW1_<your Andrew id>.ipynb`.
   - An HTML export of the notebook.
2. Ensure all visualizations and code outputs are clearly labeled.
3. Do not rename data files; use relative paths while importing data.

Rubric

Item	Points
Exploratory data analysis and data preparation	1
Development of clustering model	1

Item	**Points**
Justification of the number of clusters	1
Development of first supervised learning model predicting premature death	1
Development of second supervised learning model predicting premature death	1
Identification of five most important factors influencing premature death	1
Evaluating accuracy of the two supervised learning models	1
Recommendations for reducing premature death in Allegheny County	1
Total	**8**

Hints/Suggestions

1. **Data Preparation:**
   - Clearly describe how you handle missing values and outliers.
   - Consider imputation or exclusion strategies and justify your choices.
2. **Clustering:**
   - Suggested techniques: K-means, hierarchical clustering.
   - Suggested methods for determining the number of clusters: elbow method, silhouette score.
3. **Visualizations:**
   - Use Python libraries like Matplotlib or Seaborn for histograms, boxplots, and scatter plots to enhance EDA.
4. **Supervised Learning Models:**
   - Consider using models like linear regression, decision trees, or random forests.
   - Evaluate models using metrics such as RMSE, R^2, or accuracy.
   - Use cross-validation to prevent overfitting.
5. **Recommendations:**
   - Focus on actionable and cost-effective measures.
   - Address both immediate and long-term strategies for reducing premature deaths.

Academic Integrity

- Ensure all work is your own.
- Clearly document any external resources or libraries used.

Appendix: Outcomes and Factors

Documented in 2025 Technical Documentation:
https://www.countyhealthrankings.org/sites/default/files/media/document/CHRR%20Technical%20Documentation%202025_2.pdf

Population Health and Well-being (Outcomes)

- Premature Death
- Poor Physical Health Days
- Poor Mental Health Days
- Low Birthweight 20%
- Poor or Fair Health

Community Conditions (Factors)

- Access to Exercise Opportunities
- Flu Vaccinations
- Food Environment Index
- Preventable Hospital Stays
- Uninsured
- Primary Care Physicians
- Mammography Screening
- Mental Health Providers
- Dentists
- Air Pollution: Particulate Matter
- Drinking Water Violations
- Broadband Access
- Library Access
- Severe Housing Problems
- Driving Alone to Work
- Long Commute - Driving Alone
- High School Completion
- Some College
- Unemployment
- Children in Poverty
- Income Inequality
- Child Care Cost Burden
- Injury Deaths
- Social Associations