Faculty of Engineering
Cairo University


---

**Artificial Intellegence in Medicine - Final Project**

# 10-Year Risk of Death of Individuals from the NHANES I Epidemiology Dataset

---

Submitted to:
Dr. Inas Yassin
Eng. Christeen Adly

Submitted by:

| Name | Sec | BN |
|---|---|---|
| Mohamed Ahmed Abdelaziz | 2 | 14 |
| Mohamed Khaled Galloul | 2 | 15 |
| Mohamed Abdelkareem Seyam | 2 | 18 |
| Ahmed Mohamed Mohamed | 1 | 7 |

Cairo University
Please let us know if we need to provide any further details.
mohamed.ahmed997@eng-st.cu.edu.eg
mohamed.attia99@eng-st.cu.edu.eg

# 1 Data Science Problem

10-year risk of death of individuals from the NHANES I epidemiology dataset.

The NHANES I Epidemiologic Follow-up Study (NHEFS) is a national longitudinal study that was jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the Public Health Service. The NHEFS was designed to investigate the relationships between clinical, nutritional, and behavioral factors assessed in the first National Health and Nutrition Examination Survey NHANES I and subsequent morbidity, mortality, and hospital utilization, as well as changes in risk factors, functional limitation, and institutionalization.

The NHEFS cohort includes all persons 25-74 years of age who completed a medical examination at NHANES I in 1971-75 (n = 14,407). It is comprised of a series of follow-up studies, four of which have been conducted to date. The first wave of data collection was conducted for all members of the NHEFS cohort from 1982 through 1984. It included tracing the cohort; conducting personal interviews with subjects or their proxies; measuring pulse rate, weight, and blood pressure of surviving participants; collecting hospital and nursing home records of overnight stays; and collecting death certificates of decedents.

## 1.1 Data File Description

The NHEFS public use data files are divided into four components. These are:

- **Vital and Tracing Status Data**
  The 1992 Vital and Tracing Status contains tracing, vital status, and demographic data for all 14,407 subjects for each wave of data collection.
- **Interview Data**
  The 1982-1984, 1986, 1987, and 1992 Interview data contain information collected from the subject and proxy interviews conducted during each follow-up period. When merged together, these files provide a complete follow-up history for each subject from baseline through the last completed interview.
- **Health Care Facility Stay Data**
  The 1982-1984, 1986, 1987, and 1992 Health Care Facility Stay data contain information collected during each follow-up period regarding overnight stays in health care facilities, including diagnostic and summary information abstracted from hospital and nursing home records. Supplemental Health Care Facility Stay data is used to provide information on overnight facility stays that occurred outside the reported follow-up period. When these five Health Care Facility Stay data files are merged together, they provide a history of all reported hospitalizations and institutionalizations from baseline through 1992.
- **Mortality Data**
  The 1992 Mortality Data contains death certificate information collected during each follow-up period coded according to ICD-9 multiple-cause-of-death procedures for all 4,497 decedents identified through 1992.

## 1.2 Geographic Coverage

In the first National Health and Nutrition Examination Survey (NHANES I), data were collected from a national probability sample of the civilian non-institutionalized population. The NHANES I Epidemiologic Follow-up Study attempts to trace and interview all study subjects at their current location.

### 1.3 Features in Dataset

- Age
- Diastolic BP
- Poverty index
- Race
- Red blood cells
- Sedimentation rate
- Serum Albumin
- Serum Cholesterol
- Serum Iron
- Serum Magnesium
- Serum Protein
- Sex
- Systolic BP
- TS
- White blood cells
- BMI
- TIBC
- Pulse pressure

## 2 Available Dataset

Epidemiology Dataset

## 3 Preprocessing

### 3.1 Dealing with Missing Data

Missing data is a common occurrence in data analysis, that can be due to a variety of reasons, such as measuring instrument malfunction, respondents not willing or not able to supply information, and errors in the data collection process.

### 3.2 Feature Normalization / Scaling

Feature scaling is essential for machine learning algorithms that calculate distances between data. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. It also results in faster convergence.

### 3.3 Imputation Approaches

Handle the missing values by replacing them with substituted values based on the other values that we have.

### 3.4 Handling class imbalance

## 4 Candidate Features Proposed Methodologies

The importance of a feature is the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.

Feature importance is calculated after the model has fitted. In the features columns, it randomly shuffles/permutate each column without touching the target and other feature column and calculates how it affects the accuracy of the prediction of new shuffled data.

- applied after the model has been fitted. It selects each feature column one at a time and shuffles/permutate that column randomly
- Then it calculates how that affect the prediction/accuracy of the model
- If changing value of a feature column affect the accuracy of the model heavily then the importance of that column feature is higher. In this way, feature importance is calculated.

PCA is also a very good option to consider, generally called data reduction technique, is useful feature selection technique as it uses linear algebra to transform the dataset into a compressed form.

**Proposed Methods:**

1. **We will implement PCA feature selection technique with the help of PCA class of scikit-learn Python library.**
2. **we will implement 'Permutation importance' by the eli5 library.**

# 5 Candidate Classification and Regression techniques

- Train the following models: SVM, Decision Tree, Random Forest, KNN and XGBoost.
- Select best three promising algorithms from the above.
- Apply hyperparameter grid search on each one of the three above and get each one's best parameters.
- Apply ensemble method(Bagging) on the three classifiers above.
- Apply a Grid search on the three classifiers above to get each one's best bagging parameters.
- Apply ensemble method(Boosting) on the three classifiers above.
- Apply ensemble method(voting) on the three classifier outputs above to get one optimum classifier.
- Apply PCA on the data and retry the same steps above.
- Compare the results with PCA and without PCA.

## 5.1 Explore How a single feature affect our prediction

After we get the most important features of our model by our proposed methods in section 4; We need to know how each feature affecting accuracy.

For example, let's use the features 'Age'. We know that age is an important feature. But we don't know that the chance of death is increasing with age or decreasing. For knowing how single features affect our prediction we need to use a different technique called "Partial Dependence Plots".

Partial Dependence Plots or PDP is also a very popular method. PDP is calculated after the model is fitted. We then use a single row from test data to predict the outcome. Instead of predicting one prediction, we repeatedly alter one variable of the row to make a series of prediction. For example, for our model. we take on a row from test data and repeatedly alter a single variable value like age, and then make a series of prediction. And we do these for multiple rows, then plot average predicted the outcome on vertical axes.

Key things to remember for PDP:

- PDP calculated after a model is fitted.
- Use single row for prediction.
- Repeatedly alter a variable value to make a series prediction.
- Do that for multiple rows and plot the average prediction on the vertical axes.

## 5.2 Explore Why our model predicted its predictions

now we want to know why an individual prediction is made. Why our model made such a prediction? We will know that by using a method named **'SHAP Values'**.

SHAP Values (an acronym from SHapley Additive exPlanations) break down a prediction to show the impact of each feature . It explains why a model made a certain prediction. Based on certain features, we want to know how each feature together contribute to the prediction

# 6 Error Analysis

Find a subgroup of the test data on which the model performs poorly.