**Machine Learning Final Project**

**Project name / title:**

Predicting Well-being Disparities Across Socio-Economic Groups and in Relation to Climate and Environmental Risk.

**Group Members:**

Armande Aboudrar-Méda, Henry Baker, Giorgio Coppola, Lino Hans Julian Zurmuehl

**Theoretical Mechanism**

Individuals from lower SES backgrounds disproportionately reside and work in areas and sectors with poorer environmental conditions due to historical, economic, and social factors. These areas are characterized by higher pollution levels, inadequate housing, and limited access to clean water and air. Such environmental conditions can lead to increased exposure to pollutants and environmental hazards, contributing to higher rates of health issues, including respiratory conditions, cardiovascular diseases, and stress-related illnesses. This creates compounding effect on top other socio-economic vulnerabilities (such as limited access to healthcare, poor nutrition, and stress from economic instability) and environmental exposures.

**What are the units of analysis?**

Each row in the dataset represents an individual respondent EU resident aged 50+ whose health and wellbeing.

**What are we trying to predict?**

The label in the dataset would be a specific health-related outcome. Based on the SHARE-ENV documentation, we can use **self-perceived health status (`*sphus*`, five levels) as a proxy of well-being.** Self-perceived health status is a categorical variable; therefore, this would be a multi-class classification problem.

**What will we use to make that prediction?**

Features for predicting **self-perceived health status** **disparities** due to differences in environmental quality could include:

* Health Variables:
  + Presence of chronic diseases (e.g., heart disease, diabetes, respiratory diseases)
  + Mental health indicators (e.g., depression scores)
  + Illness and Health Conditions Related to Environmental Exposure:

`Ill\_any\_env\_related\_issue` [1/2/3]: Indicates any environment-related issue in illness periods 1 , 2, or 3.

Environment-related health issues include:

angina or heart attack

stroke

asthma

other respiratory problems

migraines

emotional distress

fatigue

infectious diseases

allergies

* Environmental Exposure Variables:
  + - Average or median concentrations of pollutants (e.g., PM2.5, NO2) at the regional level
    - Exposure to extreme temperatures and radiation
    - Cumulative environmental hazard variables
    - Foods events data
* Socio-Economic Variables:
  + - Household Net Worth (Household Net worth): Indicates the total net worth of the household, which can serve as a proxy for economic status.
    - Household Income (Household Income (current/average)): Reflects the income level of the household, providing insight into the economic well-being of respondents.
    - Education Level (ISCED educ. level): The International Standard Classification of Education (ISCED) level can indicate the educational attainment of respondents, which is often correlated with SES (socio-economic status).
    - Housing quality: Lower socio-economic groups often live in areas with poorer housing quality and higher environmental risks. Socio-economic factors, including income and material well-being, influence housing quality, which in turn affects vulnerability to environmental risks.
    - Location of dwelling: living in a big city, the suburbs or outskirts of a big city, a large town, a small town, a rural area or village
    - Work conditions
* Demographic Variables

**How would this prediction be used in a decision-making context?**

Predictions from this analysis could inform public health policies aimed at reducing health disparities by identifying those groups most vulnerable to poor environmental quality. This information could guide targeted interventions, resource allocation, and policy adjustments to improve environmental conditions in areas with significant health disparities. It can also support advocacy for environmental justice and inform strategies to mitigate the impacts of environmental risks on health, particularly among socio-economically disadvantaged groups.

**What is the best model to predict well-being?**

When selecting the optimal model to understand the role of environmental factors in predicting inequality in well-being, defined by self-perceived health status, it's crucial to evaluate models with and without environmental variables.

A model without environmental variables may capture basic relationships between socioeconomic factors, health behaviors, and outcomes but overlook the significant influence of environmental exposures on health inequalities. Conversely, a model that includes environmental variables provides a holistic view, taking into account pollution levels, extreme temperatures, and other hazards, and thus provides a deeper understanding of how environmental conditions affect health inequalities and, thus, the perception of the well-being of individuals across European countries.

By comparing these models, we can measure the improvement in predictive accuracy conferred by environmental variables, highlighting their crucial role in identifying and addressing health inequalities among older adults. This underscores the need to include environmental factors in decision-making processes, particularly in formulating public health policies and interventions aimed at reducing health inequalities.