# Project 1 Milestone 1 – Proposal and Data Selection

Astrid Fuentes, MS

Bellevue University

DSC 680: Applied Data Science

Prof. Catherine Williams

September 4th, 2022

# Project 1 Milestone 1 – Proposal and Data Selection

**Milestone 1 - Proposal**

- Topic:
  Use Machine Learning methods to predict the mortality status (either Dead or Alive) of individuals based on certain underlying health conditions like Diabetes and Hypertension.

- Business Problem:
  Describe the business problem your project is trying to solve and/or the research questions you will explore
  As a data scientist working for a life insurance company, I have been assigned to evaluate the different causes of deaths and how these have been changing throughout the years. The idea is to be able to predict if an individual with certain health conditions is more likely to die.

- Datasets:
  I will be Using the National Center for Health Statistics (NHCS) 2019 Public-Use Linked Mortality Files available for 1986-2018 NHIS, 1999-2018 NHANES, and NHANES III

- Methods:
  I will start using some statistical analysis to identify correlation and other statistical measurements. After that, I will train a model using Machine Learning to be able to predict the leading cause of death of an individual based on the presence of underlying health issues like diabetes and hypertension. I will also implement a classification model using logistic regression to predict the outcome of an individual (either Dead or Alive) based on underlying health conditions.

- Ethical Considerations:
  The files include a limited set of mortality variables for adult participants only. The public-use versions of the NCHS Linked Mortality Files were subjected to data perturbation techniques to reduce the risk of participant re-identification. For select records, synthetic data were substituted for follow-up time or underlying cause of death. This may limit the accuracy of the data.
  Another ethical consideration is the fact that people could have an underlying condition of diabetes and/or hypertension and die of an unrelated cause, for example an accident. These circumstances are not being considered in my project.

- Challenges/Issues:
  Finding the right type of model to use might probably be the most difficult challenge. I will need to review my books and past assignments to select the methods that might be a better fit for this project. I might also need to do different attempts on modelling and see which one works best and produces a higher accuracy. I will split my data sets in "training" and "testing" so that I can use real data to validate my models. This is a good way to validate my results and support.

Project 1 Milestone 1 – Proposal and Data Selection

**References**


Center For Disease Control and Prevention (CDC). 2019 Public-Use Linked Mortality Files. https://www.cdc.gov/nchs/data-linkage/mortality-public.htm

Center For Disease Control and Prevention (CDC). 2019 Public-Use Linked Mortality Files Data Dictionary. https://www.cdc.gov/nchs/data/datalinkage/public-use-linked-mortality-files-data-dictionary.pdf

Center For Disease Control and Prevention (CDC). 2019 Public-Use Linked Mortality Files Data Files. https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/linked_mortality/