Capstone Two- Project Proposal

Problem Statement

Since their first use in the 18th century, vaccines have provided humans with a means of safeguarding the health of ourselves and others. Child routine immunization has become a systematic part of preventing disease, establishing herd immunity, and reducing costly hospitalizations, and preventing re-importation and transmission of disease. The development of vaccines has eliminated or reduced the spread of deadly diseases like polio, measles, and diphtheria, which once killed thousands of children in the United States annually.

Reluctance to use the COVID-19 vaccine swiftly developed in record-setting time in response to the 2020 pandemic sowed doubt regarding vaccine safety and re-ignited long standing anti-vaccine movements, contributing to a decline in child immunization rates across the US. While most kindergarteners in the US are vaccinated, an increasingly large number of exemptions from school immunization requirements have been made and as of 2023, immunization rates had not recovered to pre-pandemic levels (article).

Data from the National Immunization Survey for children (NIS-C) data are used to produce timely estimates of immunization coverage rates for all childhood immunizations recommended by the Advisory Committee on Immunization Practices (ACIP). I plan to use the NIS-C data on children aged 19 to 35 months to identify the factors most important in determining immunization coverage and how they may have changed over time. Given the NIS-Child survey's most recent data is from their 2023 survey, I can examine changes over the years prior to that year.

Questions:

1. **What are the strongest predictors of being up-to-date on the full vaccine series? Can one predict whether a child will be vaccinated based on parent vaccine hesitancy, family income, or other parental attributes?**
    a. Use predictive models such as logistic regression or random forest classifiers
2. **Where are future outbreaks most likely to occur? (if time allows)**
    a. Clustering by geographic units (e.g., zip code)
    b. Geospatial Analysis (e.g., GIS mapping) overlay immunization rates with complementary data such as the number of measles cases reported in the area or provider density maps.

Data wrangling

The NIS-C consists of two parts, a telephone interview and, if guardian/parent permits, a paper questionnaire mailed to the child's vaccinating health provider. These data sets may need to be merged and manipulated to facilitate analysis. The CDC provides a .dat file and an input file in R. The child .dat file will somehow need to be imported and possibly merged with a separate provider data set. The CDC data could be augmented with complementary sources such as data on provider density, the number of measles cases in that zip code, or the percentage of county residents that are insured in that area providing community-based characteristics that may drive improved model prediction.

Exploratory analysis will involve looking at a merged and enriched data set and visualizing trends for different vaccine antigens across several years (e.g. 2013-2023), trends in measles cases, insurance rates, the rate of timely vaccination, and the types of providers. EDA will then use scatter plots to visually examine the associations between MMR (measles, mumps, and rubella) immunization coverage and measles case rates as well as between immunization coverage rates and the variables below:

Health insurance cover
Mother's age
Hesitancy COVID-19
Hesitancy flu
Hesitancy routine immunizations
Household income

WIC participation
Child's insurance coverage
Type of vaccination provider
Whether the provider discussed recommended vaccines during visits.
Urban vs. rural

Pre-processing and training data development
This step will involve building a pipeline that specifies imputation and feature selection parameters for logistic regression or random forest regressions models while performing crossover validation of longitudinal training data.

Modeling
In this step, I will fit the final model on all available and compare model fit or error statistics to those produced in previous model training.

Documentation
I will finish by creating a short report or slide deck on findings.