ANNUAL CLINICAL SURVEY IN INDIA
Using survey data to predict lifestyle diseases in Uttar Pradesh
Sophie Gräfnitz, Evelin Aasna, Marianne Dengo

Introduction
In this project, we used the data from an annual clinical survey in India to identify patterns and predict lifestyle diseases in the most populous state, Uttar Pradesh. The original dataset consisted of 9 states and included a total of 1.89 million observations and 53 variables. The chosen subset of adults from Uttar Pradesh had health data from 300,000 individuals.

Methods
In particular, we were interested in exploring three diseases:

- **HEART DISEASE**
- **DIABETES**
- **ANEMIA**

We were able to achieve desired results by implementing various methods of data analysis in the process, including:

**DATA CLEANING AND PREPARATION**
- Dropping unnecessary features and outliers
- Data classification, scaling, and centralizing

**EXPLORATORY ANALYSIS**
- Data examination, feature engineering
- Additional data cleaning for each disease

**MODELING**
- Building and evaluating models

Results

**Data Analysis**

Pairwise correlations among features that had an above 0.4 correlation with at least one other feature

Frequencies of blood glucose, pulse rate, and haemoglobin values among men and women in the dataset

**Conclusion**

In the course of the project, we succeeded to:

- Explore the dataset to find relationships and correlations between features
- Create new features, as well as use the results of data exploration to clean the dataset and prepare it for modeling
- Train models to sufficiently predict heart disease, diabetes, and anemia among the patients of Uttar Pradesh
- Assess the performance of said models

**LIMITATIONS**

Our limitations included:

- A lot of missing data
- Unbalanced features
- Several unclear features that weren’t fully explained in the data dictionary
- Strange or uncommon values and trends

**Sources**

Original dataset:
https://www.kaggle.com/rajanand/cab-survey

Our Kaggle kernel:
https://www.kaggle.com/sophiegrae/summary-of-project-kaggle-clinical

Project repository:
https://github.com/mdengo/cab-survey

"Uttar Pradesh is the most populous state in the Republic of India as well as the most populous country subdivision in the world."