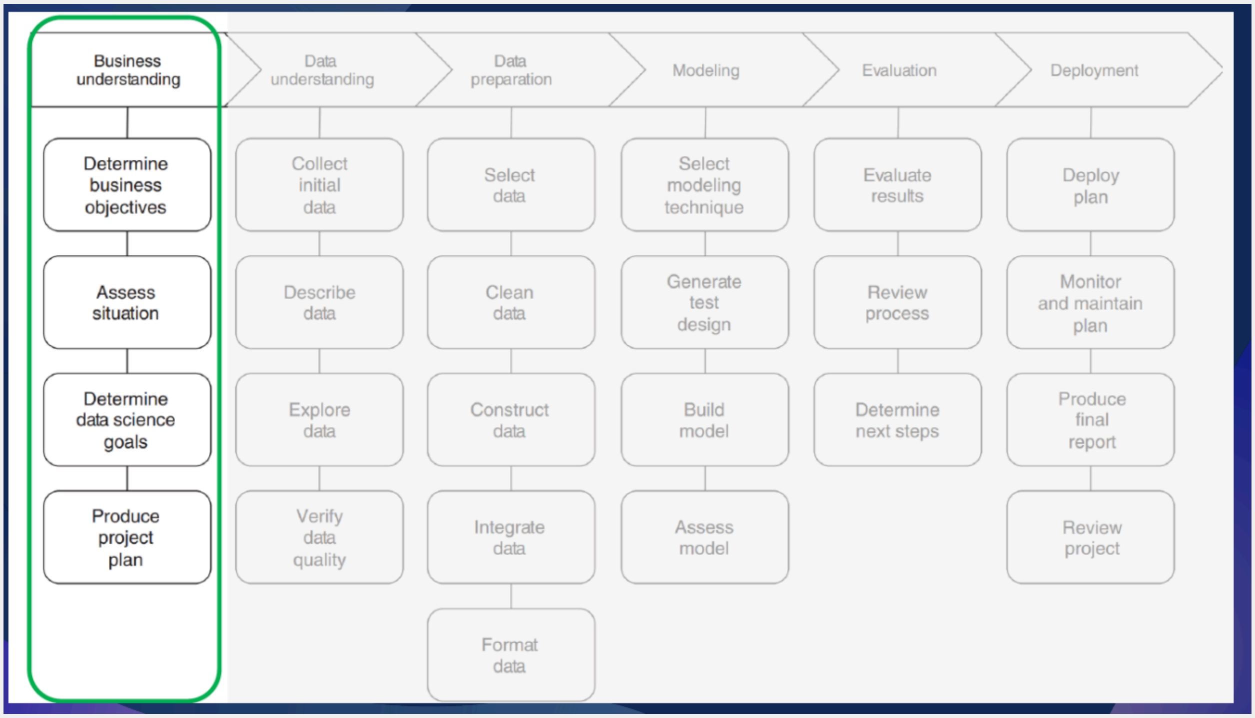
https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

https://github.com/feldmamg/ds201-final-project

^ README just quick explanation about proclaim we're trying to solve, findings

README can have pictures as it's like a webpage, not just text



- Business Understanding
  - Determine Business Objectives:
    - If we were a gym business, we would want to know if poor health habits correlate with diabetes or mental health. We would like to see if we can motivate more people to work.
  - Assess Situation
    - We have had a decrease in customers showing up at the gym.
  - Determine Data Science Goals
    - Use health metrics to incentivize more people to work out
  - Produce Project Goals
    - Get more gym membership subscriptions, especially from the people who would benefit most.
- Data Understanding
  - Collect Initial Data

        - The original data is from a telephone survey that finished in 2015. It is the “Diabetes Health Indicators Dataset” collected from the Behavioral Risk Factor Surveillance System, which is submitted to the CDC.
        - The variables we will be using are GenHlth, MntlHlth, PhysHlth, and BMI from the following website (this is just a rough draft, we can expand on this later)
    - Describe Data
        - The CDC Diabetes Health Indicators dataset was pulled from the CDC.gov website through the UCI Machine Learning Repository. This is the 2014 Behavioral Risk Factor Surveillance System (BRFSS) data. This aggregates the BRFSS, combines landline and cell phone datasets from both landline and cell data. It includes all 50 states, DC, Guam, and Puerto Rico. The data includes variables on diabetes diagnosis, demographics (race, sex), personal information (income, education), and health history (drinking, smoking, mental health, physical health).
    - Explore Data
        - The dataset was originally created to understand the relationship between lifestyle and diabetes within the country and was funded by the CDC. Each row represents an individual who participated in the study, and the only data preprocessing performed was the bucketing of age.
    - Verify data quality
        - We anticipate this data to be high-quality and accurate due to the source (the BRFSS / CDC).
- Data Preparation
    - The website that we used had the data already cleaned, and we did not need to do anything when we imported the data into our Google Colab.
- Modeling + Evaluating (EDA+MachineLearningModelTraining)
    - Select Modeling Technique
        - First, we fit a simple logistic regression model using self-reported general, mental, and physical health to predict whether a person is obese, which in our case means that the person’s BMI is over 30. We did this because it is easy to interpret the model in this context and it provided a simple baseline for us. After this, we trained another type of model, being decision tree classifiers, to predict if someone has poor mental health, which we defined as having 10 or more “bad days” in a month. We did this using the variables BMI_40plus, GenHlth, and PhysHlth, because they allow the decision trees to create human readable if-then rules within the model.
    - Generate test design

        - For each model mentioned above, we randomly split the data into 80% training and 20% testing sets using skicit-learn's train test split with stratification, which is the tool that splits the data into these sets. We did this so that the proportion of obese vs non-obese, or the proportion of patients with poor mental health vs good mental health in the train and test sets match the full dataset. Stratification in this case helps us test how well the models generalize to new people rather than just memorizing training data, which in this case shows overfitting.
    - Build model
        - To build our logistic regression model, we trained it on 80% of the data using the self-reported general, mental, and physical health to predict our baseline for obesity, which as we defined earlier meant that the person's BMI is over 30. We followed this up by building two decision-tree classifiers for poor mental health, which again as we defined earlier was if the person had at least 10 bad days in a month. The first one was a BMI only "stump" using the risk flag BMI_40plus. We also built a depth-3 decision tree using three yes/no variables that acted as risk flags, which were bad general health, bad physical health, and BMI_40plus and used class weights to ensure the model gave enough attention to the smaller group of people with poor mental health.
    - Assess model
        - The logistic regression model has about 65% accuracy, which indicates that self-reported health does not reliably identify high BMI individuals. The BMI-only showed that if the BMI is greater than or equal to 40, the probability of poor mental health doubles (from about 12% to 23%). The three-flag decision tree using bad physical health, bad general health, and high BMI achieved a similar accuracy while giving "if-then" rules easier to act on. The decision tree mainly puts people into the mostly good and mostly poor groups.
- Evaluation
    - Evaluate Results
        - Since our regression was only at 65% accuracy, which is relatively weak, it shows that self-reported health alone is not a good predictor of BMI status. The decision-tree model for poor mental health had similar accuracy but highlighted that people with bad physical health, bad general health, and BMI greater than or equal to 40 have a much higher rate of 10 or more bad mental health days. The flags are useful for identifying groups with a high-risk, but the individual risks are noisy
    - Review Process

        - We pulled CDC diabetes health indicators through the UCI database, checked to make sure there weren't any missing values, explored our BMI and various health variables, and created binary flags. We used stratified 80/20 train/test splits and simple models like regressions and decision trees to pull our insights and included cutoffs to keep outliers out of our results. A limitation we faced during this was that we only had access to self-reported survey data, and we cannot verify how accurate any measure might be.
    - Determine next steps
        - In the future, we could potentially add predictors like age, sex, activity level, and education. We could also use other machine learning models to compare and cross-validate results. From a business perspective, we could design a program that targets people with risk flags and collect follow up data to see whether targeted engagement actually reduces the number of poor mental and physical health days.
- Deployment
    - Deploy plan
        - We will use the model as a scoring tool on theoretical health survey responses. Looking at each person's general, physical, and mental health as well as each person's BMI can give us insight into risk flags. With that, the decision tree will output a probability of that person having poor mental health given those. People who have risk scores that exceed a threshold will be flagged and we will offer them financial incentives to join our gym.
    - Monitor and maintain plan
        - We plan to look at the deployed model by tracking performance metrics such as accuracy and the rate of flags and by looking at how different groups are being targeted by the model. In theory, we would retrain the model regularly to inform our gym membership acquisition process. This would allow us to use the most recent data and compare the new model to our current one.
    - Produce final report
    - Review project

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
GenHlth	Feature	Integer		Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor		no
MentHlth	Feature	Integer		Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days		no
PhysHlth	Feature	Integer		Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days		no
BMI	Feature	Integer		Body Mass Index		no