- **What types of model or models do you plan to use and why?**

  - A logistic model will be used on the dataset to determine which features are most related to or most predictive of heart disease, where the HeartDisease feature is the target. This will be a supervised learning model where "the supervisor is the target variable, a column in the data representing values to predict from other columns in the data. The target variable is chosen to represent the answer to a question the organization would like to answer or a value unknown at the time the model is used that would help in decisions. Sometimes supervised learning is also called predictive modeling. The primary predictive modeling algorithms are classification for categorical target variables or regression for continuous target variables." (Abbott, 2014, p. 5).

  - In addition to the logistic model with 1 target and 17 features, a second logistic model will be used with the 5 best features only, where the 5 features are derived with highest chi-squared statistics.

- **How do you plan to evaluate your results?**

  - I plan to calculate the accuracy, precision, recall and F1 score of both the logistic regression model with 17 features and best 5 features, as well as visualize a confusion matrix and an ROC curve where "The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6." (El Khouli, 2009). All of these together will be used to evaluate the results of the logistic regression models.

- **What do you hope to learn? Why is this data useful to solve the problem?**

- o   This data references individual lifestyle and other disease information, which is useful as
      they could have an impact on heart disease risk potential. I hope to learn what features in
      the dataset, if any, are predictive of heart disease.

- **Assess any risks with your proposal**

  - o   Logistic regression models are not perfect and have their disadvantages, "Logistic
        Regression is a statistical analysis model that attempts to predict precise probabilistic
        outcomes based on independent features. On high dimensional datasets, this may lead to
        the model being over-fit on the training set, which means overstating the accuracy of
        predictions on the training set and thus the model may not be able to predict accurate
        results on the test set. This usually happens in the case when the model is trained on little
        training data with lots of features. So on high dimensional datasets, Regularization
        techniques should be considered to avoid over-fitting (but this makes the model complex).
        Very high regularization factors may even lead to the model being under-fit on the training
        data." (Grover, 2020). If overfitting is identified as an issue via cross-validation, two types of
        regularization techniques will be used: Lasso (L1 Regularization) or Ridge (L2 Regularization),
        where λ is called the regularization parameter and controls the trade-off between fitting the
        training data well and keeping the parameters small to avoid overfitting.

- **Identify a contingency plan if your original project plan does not work out**

  - o   If this plan does not work out, the contingency plan will be to use a similar medical dataset
        on HCC (Hepatocellular Carcinoma dataset) survival. This dataset has more features (50)
        than the selected Heart Disease dataset (17), with less samples (165). The target variable
        would be the "Class" feature (nominal - 1 if patient survives, 0 if patient died), and a logistic
        regression model would be used to predict patient survival from HCC.

Include anything else you believe is important:

- **Problem Statement**

  - Which dataset feature(s) are most related to or most predictive of heart disease (single characteristics and interactions)?

- **Explain why the problem is important/interesting**

  - This problem is important because many people die of heart disease every year "Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the United States from cardiovascular disease. About 659,000 people in the United States die from heart disease each year—that's 1 in every 4 deaths. Heart disease costs the United States about $363 billion each year from 2016 to 2017. This includes the cost of health care services, medicines, and lost productivity due to death." (CDC, 2022). Understanding what factors contribute to heart disease could save lives and healthcare costs.

- **Who would be interested in solving this problem, i.e., who would you be trying to sell this project to?**

  - Doctors, healthcare professionals, patients, and life & health insurance companies would all be interested in understanding what factors are most predictive of heart disease. Life & health insurance companies would be interested in solving this problem to better understand how risky a particular customer would be to insure. Doctors, healthcare professionals and patients would be interested to ensure to be able to advise on heart disease risk factors and reduce heart disease events (coronary heart disease (CHD) or myocardial infarction (MI)).

- **Where did you get your data?**

  o This dataset was obtained from an open source data website, Kaggle

    (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-

    disease?resource=download). The dataset contains 319,795 rows and 18 columns or

    features (9 booleans, 5 strings and 4 decimals), which are:

    ▪ HeartDisease: Respondents that have ever reported having coronary heart disease

      (CHD) or myocardial infarction (MI) - either yes or no

    ▪ BMI: Body Mass Index (BMI) - float

    ▪ Smoking: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs

      = 100 cigarettes] - either yes or no

    ▪ AlcoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week

      and adult women having more than 7 drinks per week - either yes or no

    ▪ Stroke: (Ever told) (you had) a stroke? - either yes or no

    ▪ PhysicalHealth: Now thinking about your physical health, which includes physical

      illness and injury, for how many days during the past 30 days have you exercised -

      integer

    ▪ MentalHealth: Thinking about your mental health, for how many days during the

      past 30 days was your mental health not good? - integer

    ▪ DiffWalking: Do you have serious difficulty walking or climbing stairs? - either yes or

      no

    ▪ Sex: Are you male or female? - string

    ▪ AgeCategory: Fourteen-level age category - string

    ▪ Race: Imputed race/ethnicity value- string

- Diabetic: (Ever told) (you had) diabetes? - either yes, no, no-borderline diabetes, or yes pregnancy

- PhysicalActivity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job - either yes or no

- GenHealth: Would you say that in general your health is... - Excellent, Very Good, Good, Fair, Poor

- SleepTime: On average, how many hours of sleep do you get in a 24-hour period? - integer

- Asthma: (Ever told) (you had) asthma? - either yes or no

- KidneyDisease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? - either yes or no

- SkinCancer: (Ever told) (you had) skin cancer? - either yes or no

References

1. CDC. (2022, February 7). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved March 21, 2022, from https://www.cdc.gov/heartdisease/facts.htm

2. Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. Wiley.

3. El Khouli, R. H., Macura, K. J., Barker, P. B., Habba, M. R., Jacobs, M. A., & Bluemke, D. A. (2009, November). Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. Journal of magnetic resonance imaging : JMRI. Retrieved February 20, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935260/

4. Grover, K. (2020, June 23). *Advantages and disadvantages of logistic regression*. OpenGenus IQ:

   Computing Expertise & Legacy. Retrieved March 21, 2022, from

   https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/