VDS Application Write-Up
Lana Cartailler

1. Background information (1-2 paragraphs)
    a. Heart disease is a major public health concern, influenced by multiple risk factors such as high BMI, diabetes, smoking, and low physical activity. Predicting the likelihood of heart disease based on clinical features can help identify at-risk individuals early on. In this project, my goal was to build a machine learning model (a neural network, more specifically) to predict whether an individual has heart disease based on several clinical attributes.
2. Problem statement (1-2 sentences)
    a. The problem is to predict whether an individual has heart disease based on clinical data, using a dataset containing 17 attributes.
3. Hypothesis (Optional, but could make your write-up more cohesive)
    a. I hypothesized that a neural network model would be effective at predicting heart disease because of its ability to analyze complex, non-linear relationships. Thus, relationships between health factors like BMI and physical activity that are not modeled well by linear models would become more apparent after being trained on a neural network.
4. Methods (1 or more paragraphs, feel free to include different models you tried but emphasize on the model(s) you choose to report the results on)
    a. Data Preprocessing:
        i. The dataset was loaded and split into features (X) and the target variable (y), where y indicates whether an individual has heart disease.
        ii. Non-numeric features such as 'AgeCategory', 'Race', 'Diabetic', and 'GenHealth' were one-hot encoded, while binary features like 'Smoking' and 'Sex' were binary encoded.
        iii. Numerical features such as 'BMI', 'PhysicalHealth', and 'MentalHealth' were standardized using StandardScaler to normalize the data.
        iv. The data was then split into training (90%) and test (10%) sets.
    b. Model Architecture
        i. A neural network model was defined using Keras, with three layers:
            1. The first hidden layer used ReLU activation and contained a number of neurons equal to the average of the input and output layers.
            2. The second hidden layer contained 100 neurons and also used ReLU activation.
            3. The output layer contained a single neuron with a sigmoid activation function for binary classification (heart disease or no heart disease).
        ii. Class weights were applied to account for the imbalance in the target variable, where 'No' (no heart disease) was more prevalent than 'Yes'.
    c. Training

        i. The model was compiled using the Adam optimizer with a learning rate of 0.001, and the loss function was binary cross entropy. Metrics such as accuracy, precision, recall, and binary cross entropy were tracked.
        ii. The model was trained for 25 epochs with a batch size of 50.
    d. Evaluation
        i. The model was evaluated using the test data. Predictions were made based on a threshold of 0.5 for binary classification. An alternative scenario was also explored where the threshold was increased to 0.55.
5. Results and Discussion (One or more paragraphs, include any results).
    a. The model achieved reasonable accuracy but suffered from a high rate of false positives. Specifically, the model predicted that approximately 3 times more subjects had heart disease than was true.
    b. In terms of accuracy, the model correctly predicted around 80% of the cases, but its precision was much lower due to the imbalance in predictions.
    c. By adjusting the threshold for classification, the number of false positives could be reduced, but this might also lower recall (fewer true positive cases are identified).
6. Any outside resources that you use (cite your sources!)
    a. https://keras.io/guides/sequential_model/
    b. https://www.kdnuggets.com/2018/06/basic-keras-neural-network-sequential-model.html