REPORT

- Problem description
  - Overview:
    Develop a medical diagnosis system that compares two machine learning algorithms namely; **KNN and SVM.** Use patient data such as symptoms, medical history, and test results to predict the likelihood of various diseases or conditions. If possible, add a deep learning model towards the end and compare it to see if it's better.

  - Dataset used: Heart Disease

- The dataset for this project was pre-processed in R Studio and later uploaded to Google Drive in order to use it on Google Colab.

- Suggested ML model's overview

- **Note: for KNN and SVM, there were two types of datasets. One was hot encoded(21 columns) and the other was label encoded(11 columns). Both datasets were used in the notebook. The report here just shows the best between the both of them.**
  - **KNN:**
    - **Description:** K-Nearest Neighbors is a simple, non-parametric, and lazy learning algorithm used for classification and regression tasks. It doesn't involve any training phase as it stores all the available data and classifies new data points based on similarity measures.(e.g distance functions).
    - **Methodology:** The model takes a new patient's data and calculates the distance (commonly Euclidean distance) between the patient's data and all other data points in the training set. It then identifies the K closest data points to the new patient's data based on the calculated distances
    - **Results:**
      - **Accuracy:** 80.9%
      - **Classification Report:**

	precision	recall	f1-score	support
0	0.85	0.83	0.84	64

	precision	recall	f1-score	support
1	0.77	0.80	0.78	45
accuracy			0.82	109
macro avg	0.81	0.81	0.81	109
weighted avg	0.82	0.82	0.82	109

- **Confusion Matrix:**

Predicted	Heart Disease	Normal
Actual		
Heart Disease	71	15
Normal	20	78

- **SVM:**
  - **Description:** It is a supervised learning algorithm used for classification and regression tasks. It works best by finding the optimal hyperplane that separates the data into different classes.
  - **Methodology**: It aims to find the hyperplane that separates the two classes (heart disease and Normal) in the feature space. It does that by identifying the data points (support vectors) that are closest to the hyperplane. For non-linearly separable data, SVM uses kernel functions (e.g. linear, radial basis, polynomial) to transform the input into higher-dimensional spaces where a linear hyperplane can effectively separate the classes.
  - **Results**
    - **Accuracy:**78%
    - **Classification Report:**

	precision	recall	f1-score	support
Heart Disease	0.74	0.81	0.77	86
Normal	0.82	0.74	0.78	98
accuracy			0.78	184
macro avg	0.78	0.78	0.78	184
weighted avg	0.78	0.78	0.78	184

    - **Confusion Matrix:**

Predicted	Heart Disease	Normal
Actual		
Heart Disease	70	16
Normal	25	73

- **Deep Learning:**

- **Description:** Deep Learning is a subset of machine learning that involves neural networks with many layers (deep neural networks). It excels at processing and learning from large amounts of data by automatically discovering patterns and features. Deep learning models, such as Convolutional Neural Networks (CNNs)is particularly effective in tasks like image recognition, natural language processing, and complex pattern recognition. These models require substantial computational power and large datasets but are capable of achieving high accuracy and performance in various applications.
- **Methodology:** The process begins by taking an input and passing it through the layers of the model, each containing activation functions. These activation functions calculate the output of each neuron based on a weighted sum of the inputs. The model's output is then compared with the desired output, generating a loss score. This loss score is sent to the optimizer, which adjusts the weights to minimize the loss. By adjusting the weights, the model fine-tunes the weighted sums to achieve the desired output more accurately.
- **Results:**
  - **Training Accuracy:**98.43%
  - **Validation Accuracy:**86.2%
  - **Testing Accuracy:** 80.7%
  - **Classification Report:**

	precision	recall	f1-score	support
0	0.85	0.83	0.84	64
1	0.77	0.80	0.78	45
accuracy			0.82	109
macro avg	0.81	0.81	0.81	109
weighted avg	0.82	0.82	0.82	109

  - **Confusion Matrix:**

Predicted	Heart Disease	Normal
Actual		
Heart Disease	53	11
Normal	9	36

- Conclusion
  - **Model Performance Results comparison:** Among the three models evaluated (KNN, SVM, and Deep Learning), the Deep Learning model demonstrated the highest accuracy. Additionally, it provided greater flexibility for parameter adjustments to enhance learning and accuracy.
  - **Future work:** The Deep Learning model has the potential to achieve even higher accuracy. By further adjusting and fine-tuning the parameters, it is anticipated that an optimal and more accurate model can be developed.