Heart Disease Prediction

Using Supervised Machine Learning Models

by
Vaishak S

Contents

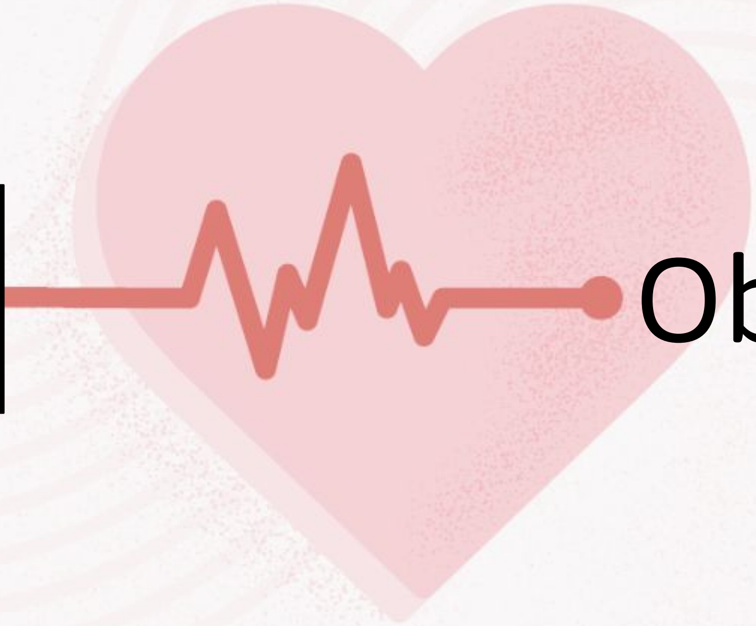
01

Problem Statement

Problem Statement

Heart disease is a leading cause of death, yet predicting it early remains difficult due to diverse risk factors like age, BMI, and lifestyle. Our ensemble learning approach combines multiple models to enhance accuracy and support timely interventions.

02



Objective

Objective

The goal is to enhance heart disease prediction accuracy by applying supervised machine learning models such as Logistic Regression, Decision Trees, Random Forests, Ensemble Methods, and Artificial Neural Networks. The models are optimized using feature engineering, data preprocessing, and class balancing, and evaluated with metrics like Accuracy, Precision, Recall, and F1 Score to support early diagnosis.

03 Approach

DATA	
Features	**Values**
General_Health	Excellent', 'Fair', 'Very Good', 'Poor', 'Good'
Checkup	Within the past year', 'Within the past 2 years', 'Within the past 5 years', '5 or more years ago', 'Never'
Exercise	Yes', 'No'
Heart_Disease	Yes', 'No'
Skin_Cancer	0, 1
Other_Cancer	0, 1
Depression	Yes', 'No'
Diabetes	No', 'Yes', 'No, pre-diabetes or borderline diabetes', 'Yes, but female told only during pregnancy'
Arthritis	Yes', 'No'
Sex	Male', 'Female'
Age_Category	40-44', '80+', '30-34', '55-59', '65-69', '75-79', '70-74', '50-54', '25-29', '60-64', '45-49', '18-24', '35-39'
Height_(cm)	Height in centimeters
Weight in kilograms	Weight in kilograms
BMI	Numerical value
Smoking_History	Yes', 'No'
Alcohol_Consumption	Numerical value
Fruit_Consumption	Numerical value
Green_Vegetables_Consumption	Numerical value
FriedPotato_Consumption	Numerical value

Data Insights

- Total samples: **308854**

- Total Features: **19** (Numerical - **9**, Categorical - **10**)

- Target: Heart_Disease - No(**0**) & Yes(**1**)
  - 0 : **283883**
  - 1 : **24971**

Data Preprocessing

- ❑ Feature Construction: Derived **4** new features
- ❑ Feature Scaling: **StandardScaler** on numerical attributes
- ❑ Feature Encoding: **OnehotEncoder** on Categorical attributes
- ❑ Handling Imbalanced Target Class: **Undersampling** and **SMOTE**

04

Results and Inferences

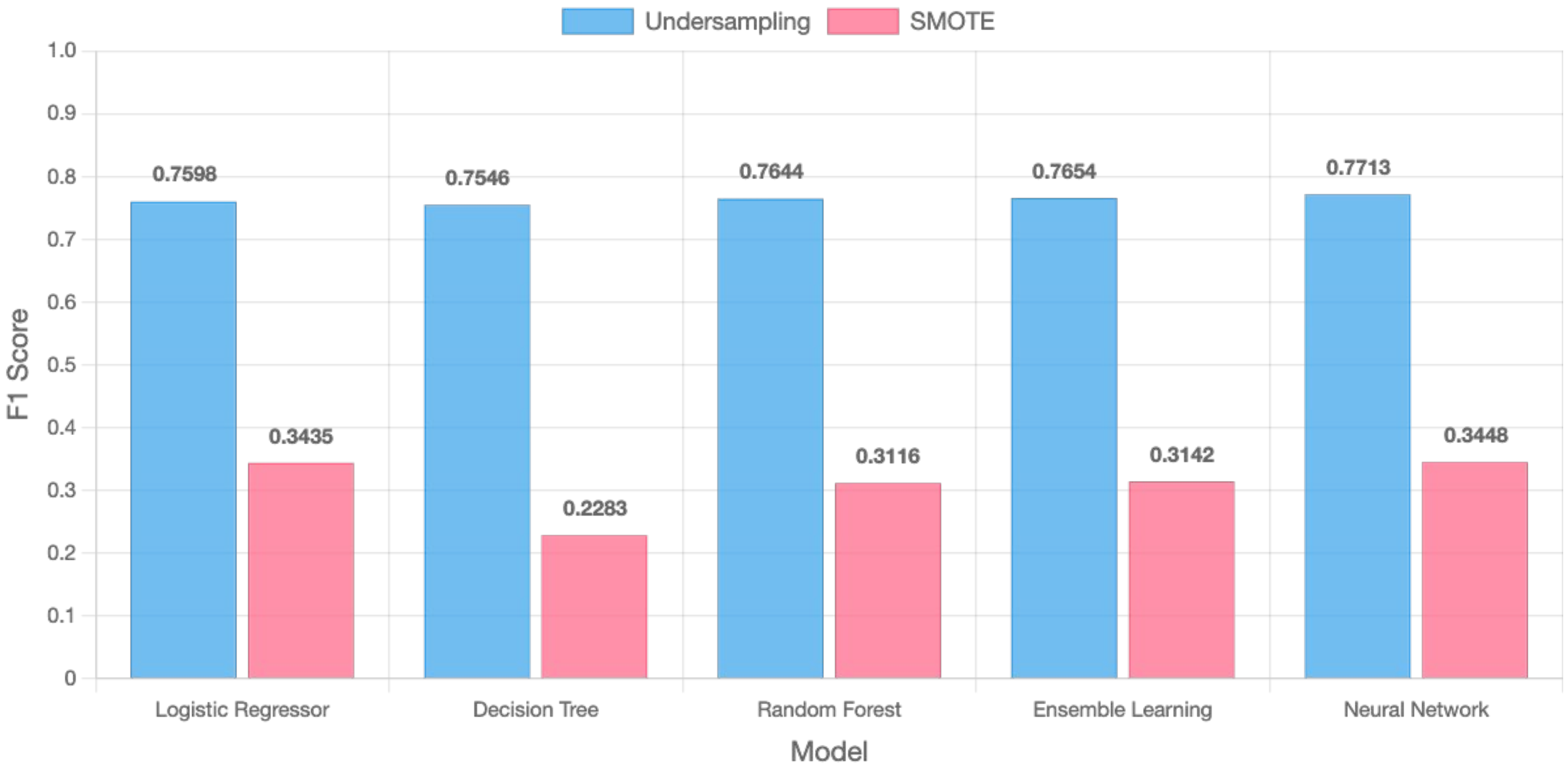
Following are the observations when fit( ) method is applied on the Test dataset.

Model Used	Sampling technique	Accuracy	Precision	Recall	F1 Score
Logistic Regression	Undersampling	0.7522	0.7371	0.7839	0.7598
	SMOTE	0.743	0.2202	0.7805	0.3435
Decision Tree	Undersampling	0.7429	0.7218	0.7905	0.7546
	SMOTE	0.4705	0.1306	0.9099	0.2283
Random Forest	Undersampling	0.7487	0.7195	0.8152	0.7644
	SMOTE	0.6907	0.1927	0.813	0.3116
Ensemble Learning	Undersampling	0.7526	0.7277	0.8072	0.7654
	SMOTE	0.6881	0.1938	0.8296	0.3142
Neural Network	Undersampling	0.7543	0.7214	0.8286	0.7713
	SMOTE	0.8065	0.2434	0.5913	0.3448

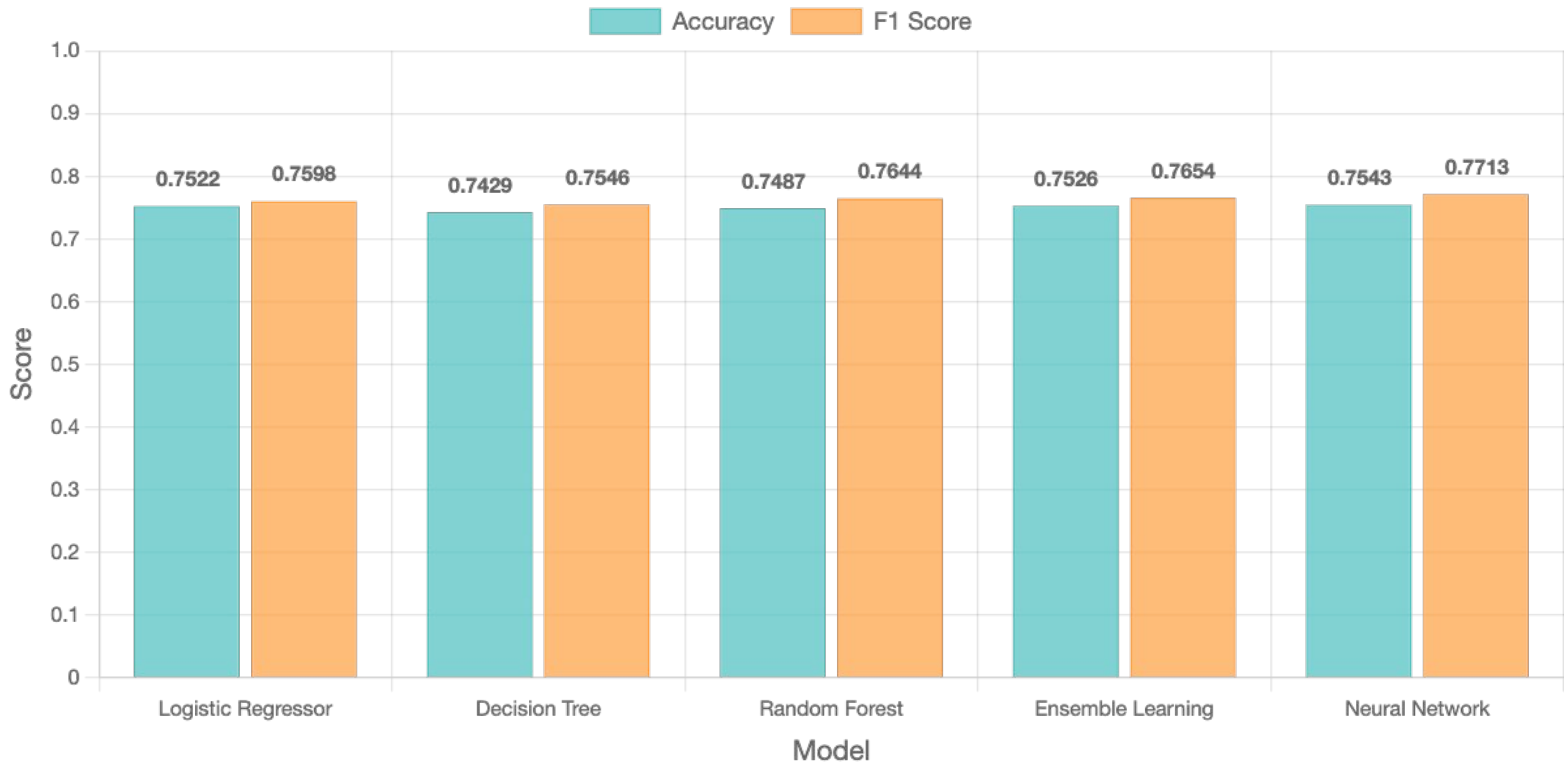
Inferences

- **Neural Network** with Undersampling emerged as the **top performer**, achieving the **highest F1 Score** (**0.7713**), indicating the best balance between precision and recall.
- **Ensemble Learning** and **Random Forest** (with **Undersampling**) followed closely, showing competitive F1 scores (0.7654 and 0.7644 respectively), making them reliable alternatives.
- Across all models, **Undersampling** consistently **outperformed SMOTE**, which suffered from very low precision values, significantly reducing F1 Scores.
- **Accuracy alone was not a sufficient indicator of model performance** - SMOTE showed decent accuracy in some cases but very **poor F1 Scores** due to imbalance in precision and recall.
- Overall Model Ranking: **Neural Network** > **Ensemble Learning** > **Random Forest** > **Logistic Regression** > **Decision Tree**

F1 Score Comparison: Undersampling vs SMOTE



Undersampling: Accuracy, F1 score plot



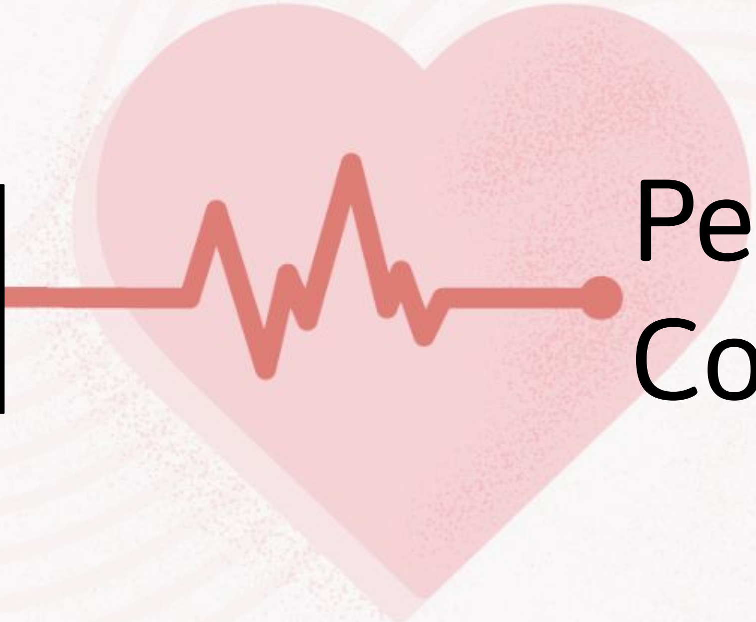
Neural Network Model Summary

- **Loss function**: binary_crossentropy
- **Regularization**: L2 regularization
- **Activation function**: selu, relu, sigmoid
- **Metrics**: Accuracy, Precision, Recall and F1 score
- **Early Stopping**: val_accuracy
- **fit( )**:
  - epochs = 100,
  - batch_size = 32,
  - validation_split = 0.2,
  - early_stopping = val_accuracy

```
Model: "sequential_16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_52 (Dense)             (None, 90)                3060
_________________________________________________________________
batch_normalization_24 (Batc (None, 90)                360
_________________________________________________________________
dropout_32 (Dropout)         (None, 90)                0
_________________________________________________________________
dense_53 (Dense)             (None, 40)                3640
_________________________________________________________________
batch_normalization_25 (Batc (None, 40)                160
_________________________________________________________________
dropout_33 (Dropout)         (None, 40)                0
_________________________________________________________________
dense_54 (Dense)             (None, 1)                 41
=================================================================
Total params: 7,261
Trainable params: 7,001
Non-trainable params: 260
```

- **Hyperparameter tuning and Cross Validation**: GridSearchCV

05

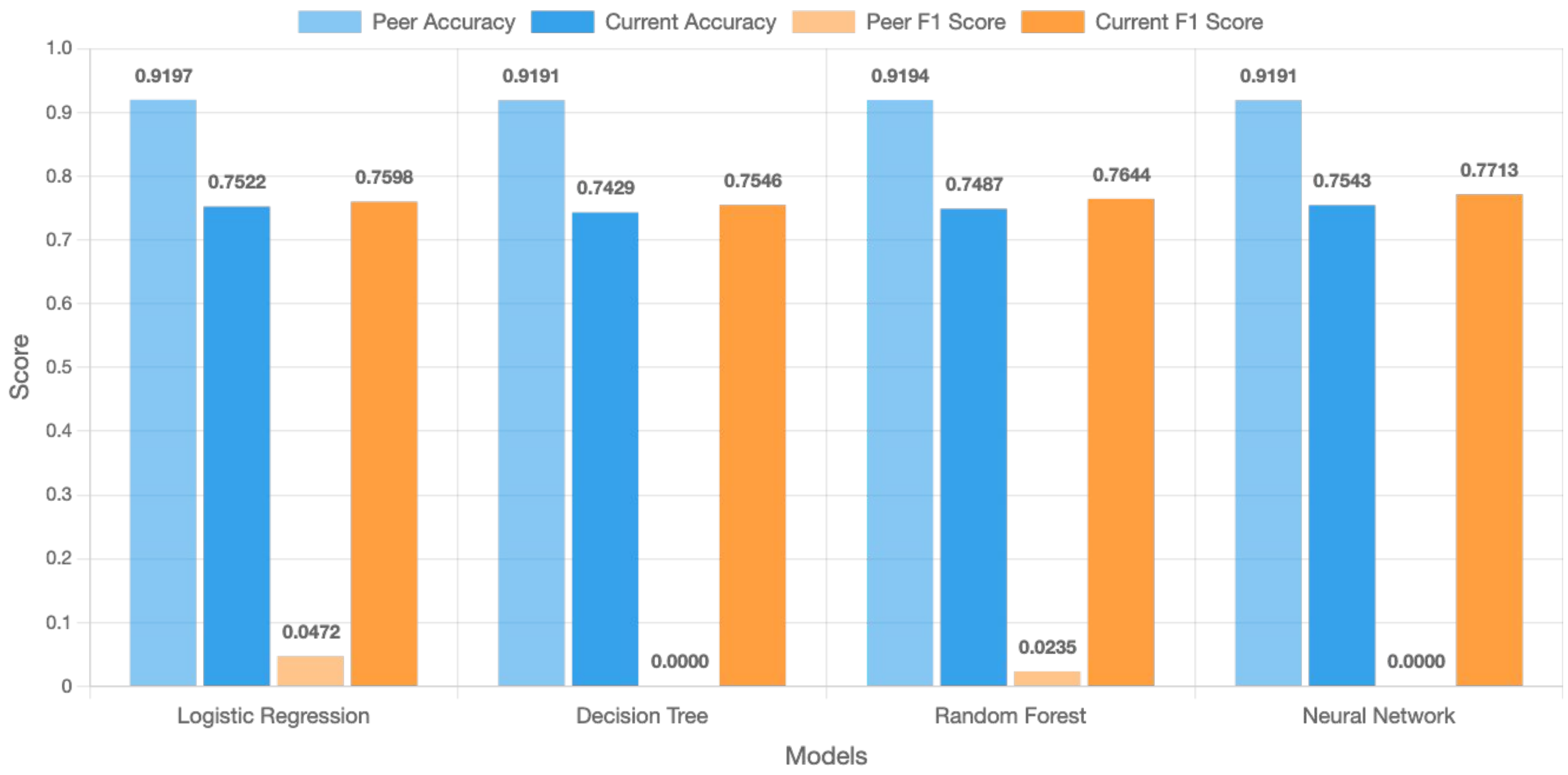


Peer Comparison

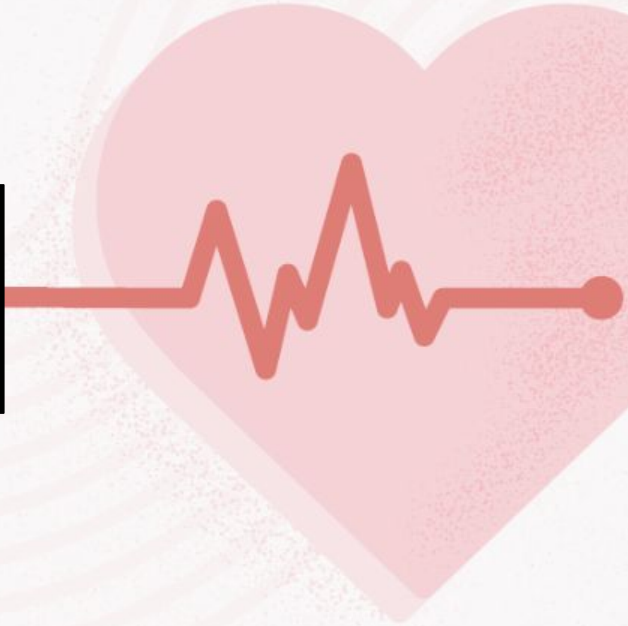
Peer Comparison

Model Used	Metric	Peer Benchmark	Current Benchmark
Logistic Regression	Accuracy	0.9197	0.7522
	F1 Score	0.0472	0.7598
Decision Tree	Accuracy	0.9191	0.7429
	F1 Score	0	0.7546
Random Forest	Accuracy	0.9194	0.7487
	F1 Score	0.0235	0.7644
Neural Network	Accuracy	0.9191	0.7543
	F1 Score	0	0.7713

Peer vs Current Benchmarks



06



Proposed Novelty in Approach

Proposed Novelty in Approach

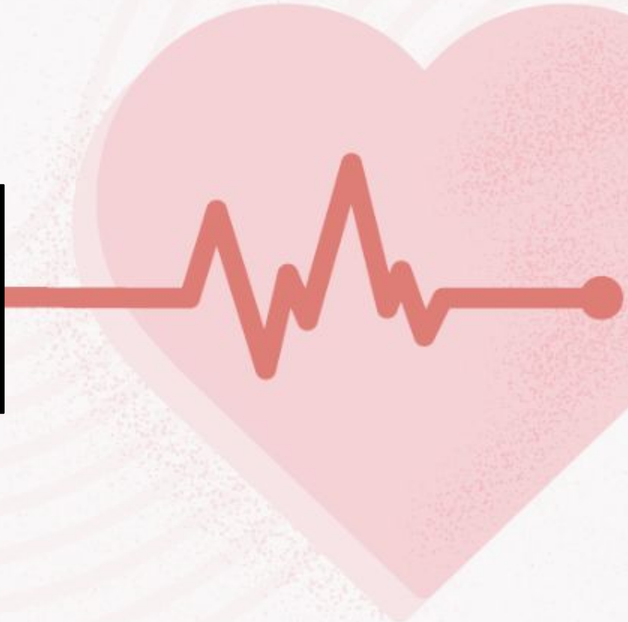
- Introduced **4** new features from existing features.
  - **BMI_Category**: categorize into Underweight, Normal, Overweight, Obese
  - **Healthy_eating_ratio**: FriedPotato_Consumption - (Fruit_Consumption + Green_Vegetables_Consumption)/2
  - **Age_numeric**: taking the mean of Age_Category
  - **Cancer_Risk**: columns grouped into one

- Overcoming target class imbalance by **undersampling** the data for training. Only **49942** samples were considered.

- Columns neglected as part of Feature selection are: **Age_Category**, **Skin_Cancer**, **Other_Cancer** and **Checkup**

07

Areas for Refinement



Areas of Refinement

- Dataset needs to have more laboratory result data like **Cholestrol level**, **Blood Pressure**, **Blood Sugar** and **ECG reports**.

- **Lack of instances** or **samples** to overcome class imbalances

08 Conclusion

Conclusion

In this project, we explored and compared various machine learning models to predict the likelihood of heart disease, with the aim of identifying the most effective approach for early detection. By analyzing model performance through metrics like Accuracy, Precision, Recall, and F1 Score, we found that neural networks, in particular, offer promising results when paired with proper data balancing techniques like undersampling.

We hope this work contributes meaningfully toward building smarter, data-driven solutions in the healthcare space—especially in aiding early diagnosis and timely intervention for heart disease. With further optimization and real-world validation, such predictive models can play a crucial role in reducing the burden of cardiovascular conditions and ultimately saving lives.

THANK YOU

Link to code:
https://jupyter.e.cloudxlab.com/user/vaishaksatheesh6467/notebooks/keras_sample_projects/CVD_Final_Project.ipynb#

Google Drive Link:
https://drive.google.com/drive/folders/1IGt3J6xJu9D_McUo9bg3MKKiDgQj4szG?usp=drive_link