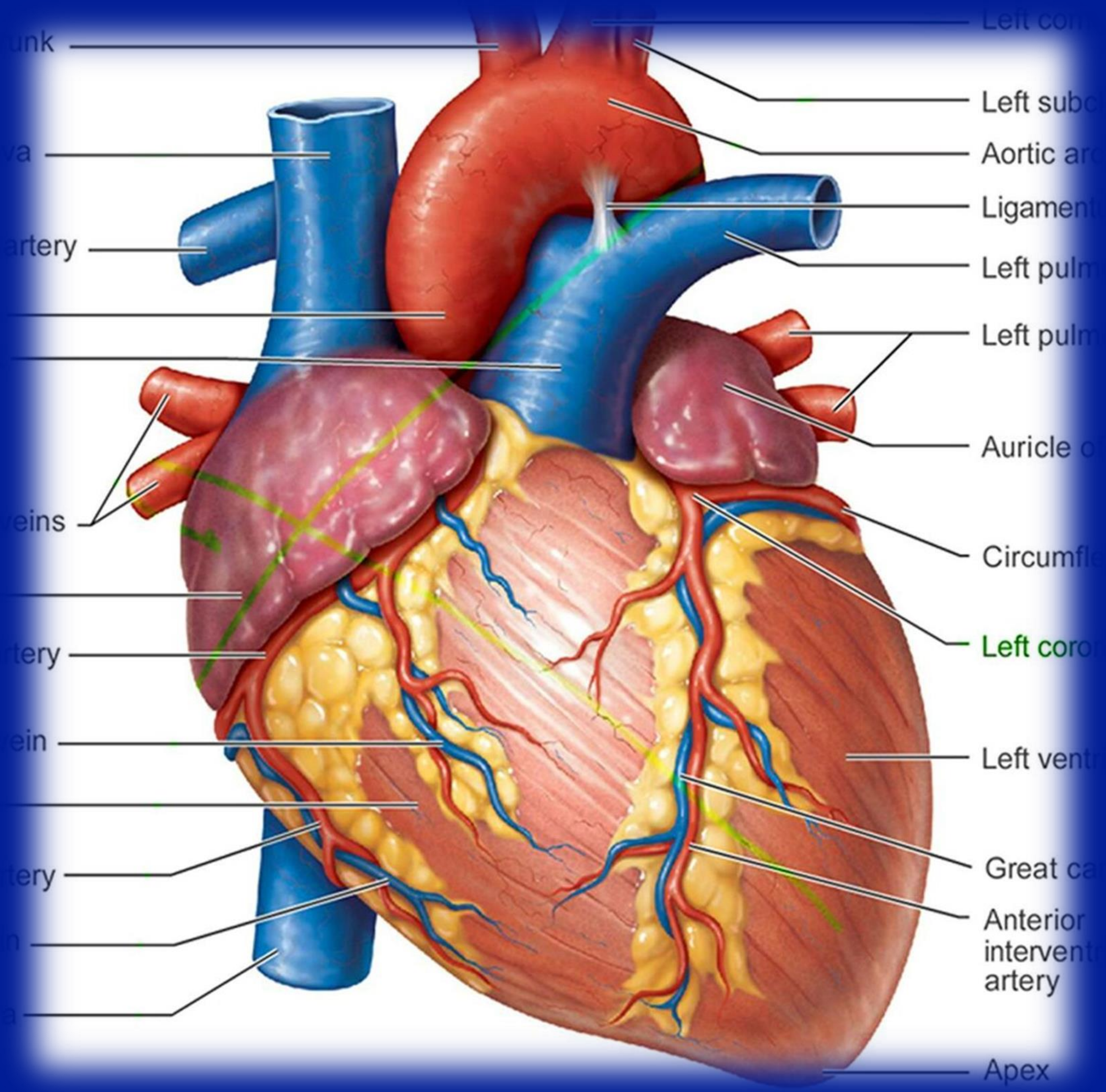
Hearth Attack Analysis & Prediction
Machine Learning Approach

Alexandre Gonçalves

Francisco Pinto



Methodology

01 Exploratory Data Analysis

02 Machine Learning

03 Fuzzy Inference System

04 Deep Learning

INTRODUCTION

In medical applications, predicting heart attacks is crucial, requiring models that are both accurate and interpretable

- ”Heart Attack Analysis & Prediction Dataset” from Kaggle,
- 303 records and 14 clinical features.
- The target variable is binary, indicating whether a patient is at risk of a heart attack;

This project aims to compare different types of machine learning models, focusing on both performance and interpretability.

01

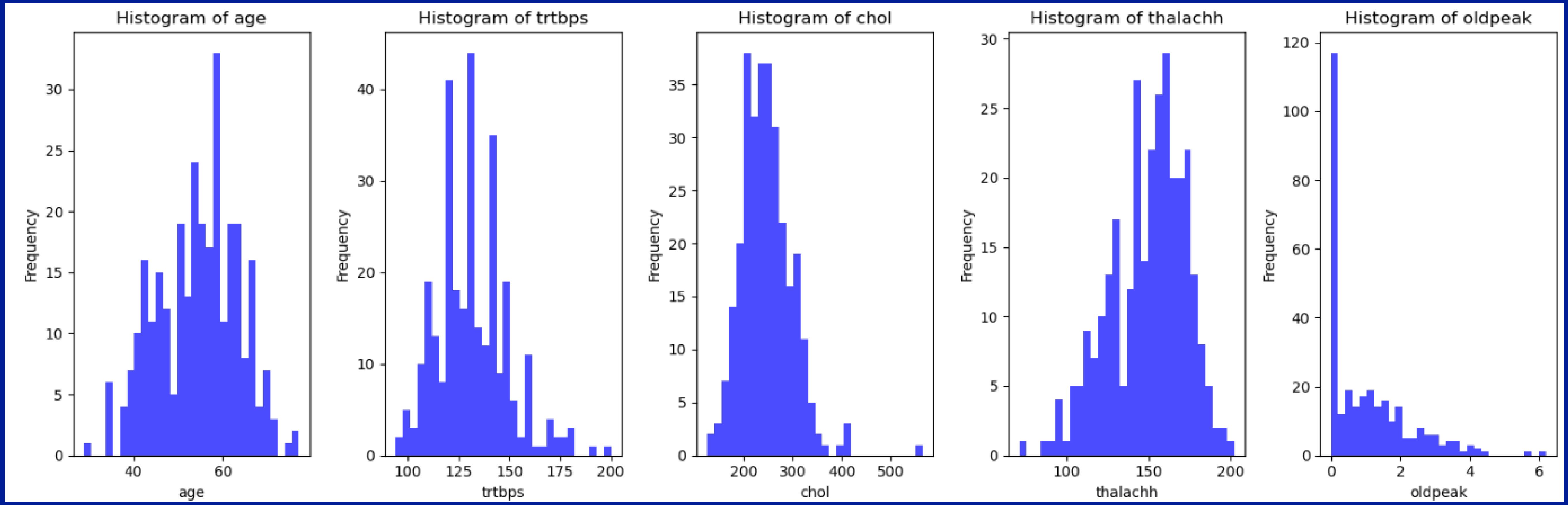
Exploratory Data Analysis

Exploratory Data Analysis

Variable	Type	Description
Age	Numerical	Patient's age
Sex	Categorical (0/1)	Patient's sex
Cp	Categorical (0-3)	Degree of chest pain symptoms
Trestbps	Numerical	Resting blood pressure (mm Hg)
Chol	Numerical	Patient's cholesterol (mg/dl)
Fbs	Categorical (0/1)	Fasting blood sugar >120 mg/dL
Restecg	Categorical (0-2)	Resting ECG results
Thalach	Numerical	Maximum heart rate achieved
Exang	Categorical (0/1)	Exercise induced angina
Oldpeak	Categorical (0/1)	ECG depression induced by exercise
Slope	Categorical (0-2)	Slope used during the exercise
Ca	Categorical (0-3)	Number of major vessels
Thal	Categorical (0-3)	Degree of thalassemia symptoms

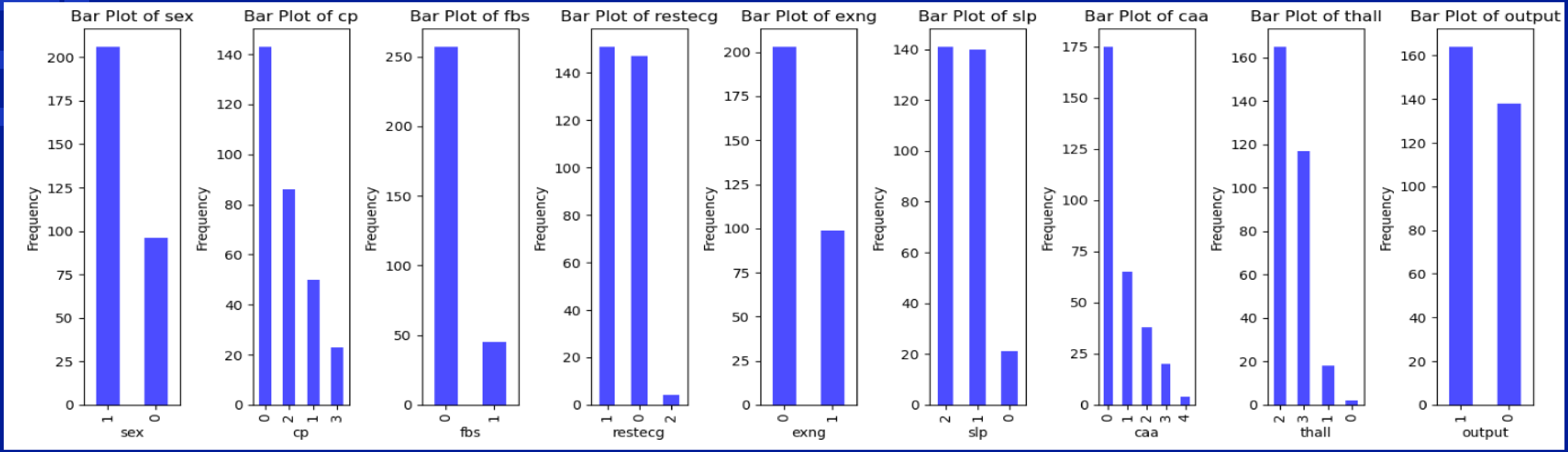
Exploratory Data Analysis

Dataset contains 303 observations, each with 14 features;



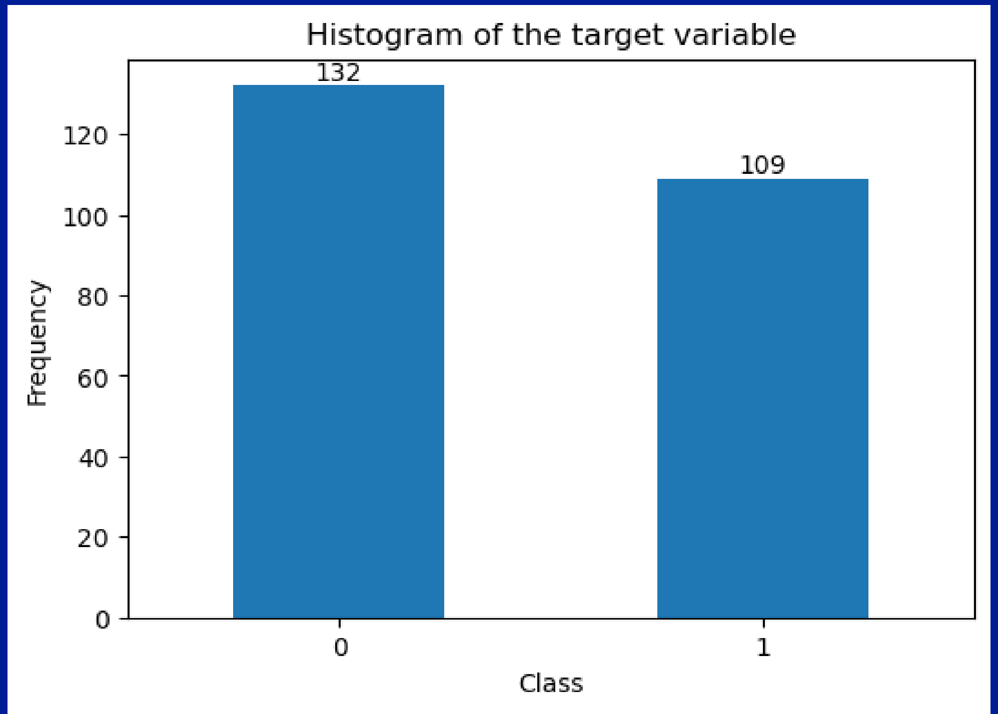
Exploratory Data Analysis

Dataset contains 303 observations, each with 14 features;



Feature Selection and Engineering

- IQR used to identify and treat outliers;
- Pearson correlation matrix used to identify correlated pairs of features;
- The target variable wasn't extremely unbalanced
- RFE used to recursively identify relevant subsets of features;
- Standard Scaler used to scale the data and adjust its distribution.



Important Considerations

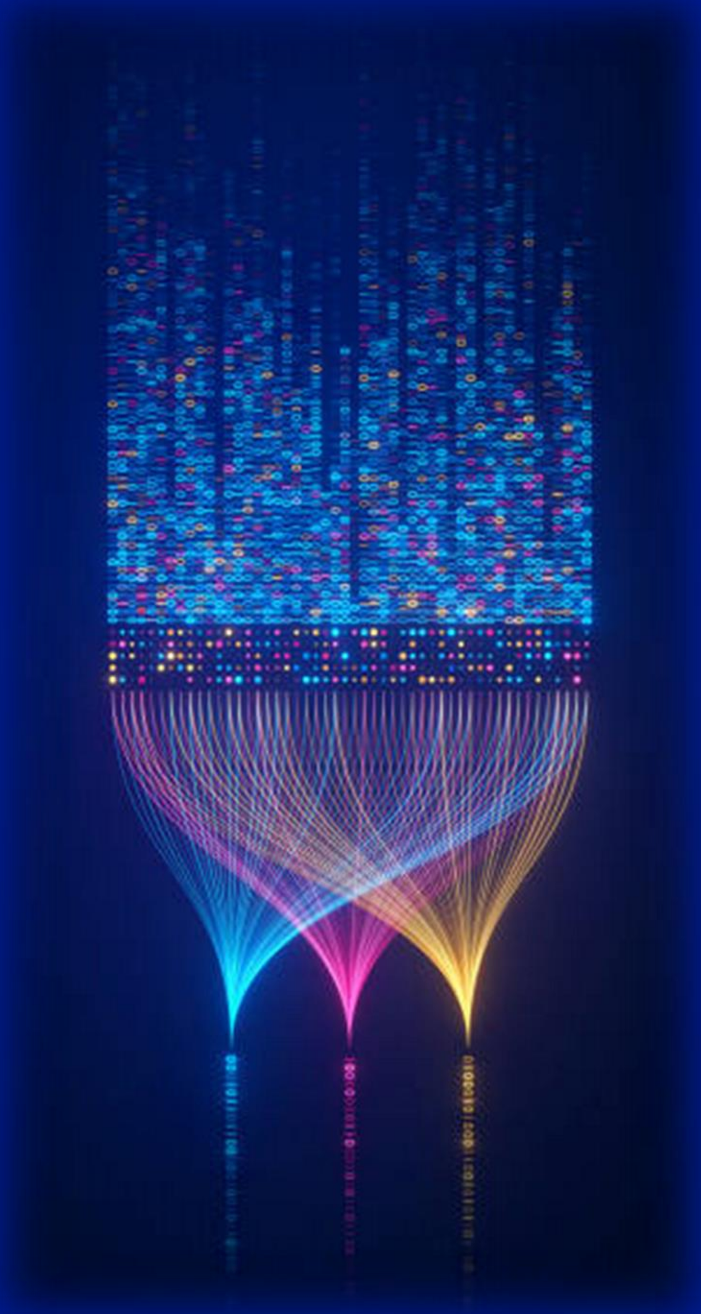
- Our recall should be as high as possible;
- Due to the size of the dataset – K-Fold CV

02

Machine Learning

Different Methods

- Random Forest Classifier
- XGBoost Classifier
- Support Vector Machine

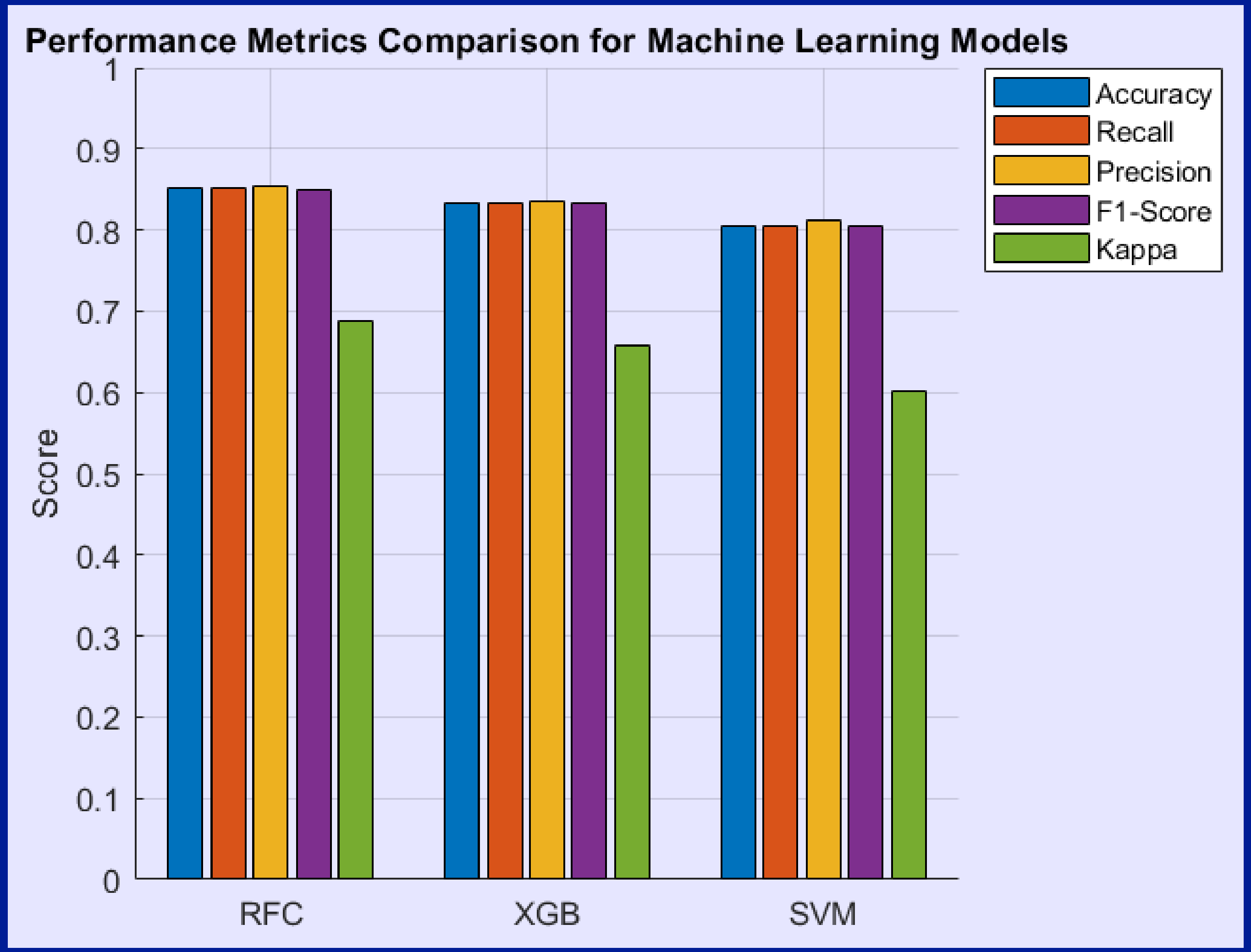


Results

RFC
- Min_samples_split = 2
- Min_samples_leaf = 5

XGB
- Min_child_weight = 10
- Mas_depth = 3

SVM
- C = 1
- Gamma = ‘scale’



03

Fuzzy Inference System

Fuzzy Inference System

- Simpler and more explainable

- Building steps
  1. Fuzzy Clustering;
  2. Approximate membership functions;
  3. Estimate consequents.

- Since the data used to build the model was scaled, we will adapt the final model so it can receive and work with real world values, making it more interpretable:

$$\textbf{Gaussian Membership Functions}: \quad \sigma' = \frac{\sigma}{\sigma_{\text{data}}}, \quad \mu' = \mu\sigma_{data} + \mu_{data}$$

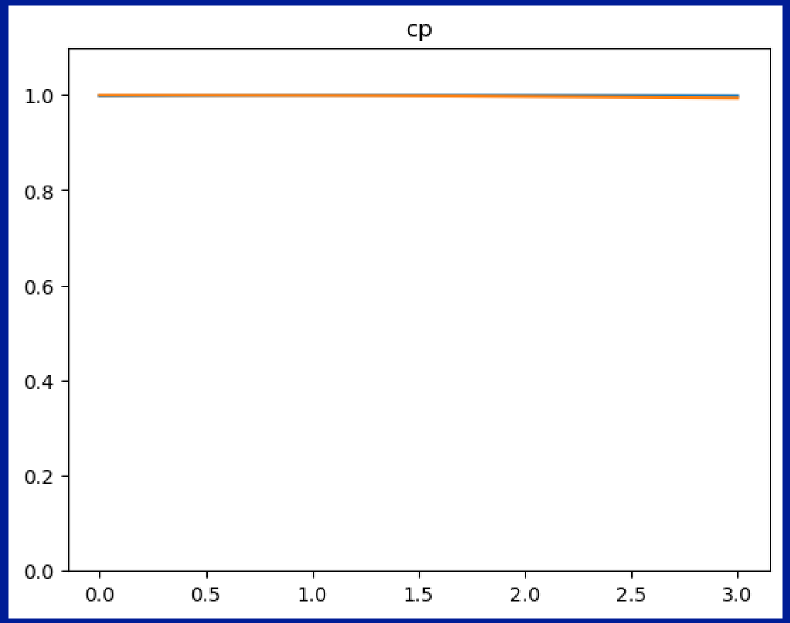
$$\textbf{Consequents}: \quad Y = aZ + c \Rightarrow Y = aX + \left(c - \frac{\mu_{data}}{\sigma_{\text{data}}}\right)$$

Fuzzy Inference System

- Fuzzy c-means clustering delivered bad results, cluster centers were almost coinciding, and membership functions were almost one in all the universe of discourse (the clusters were “competing” with each other.

- FST-PSO clustering was used, as offered by pyFUME. This method is better at handling high-dimensional or complex datasets.

- This method fixed our problem in most of the membership functions.



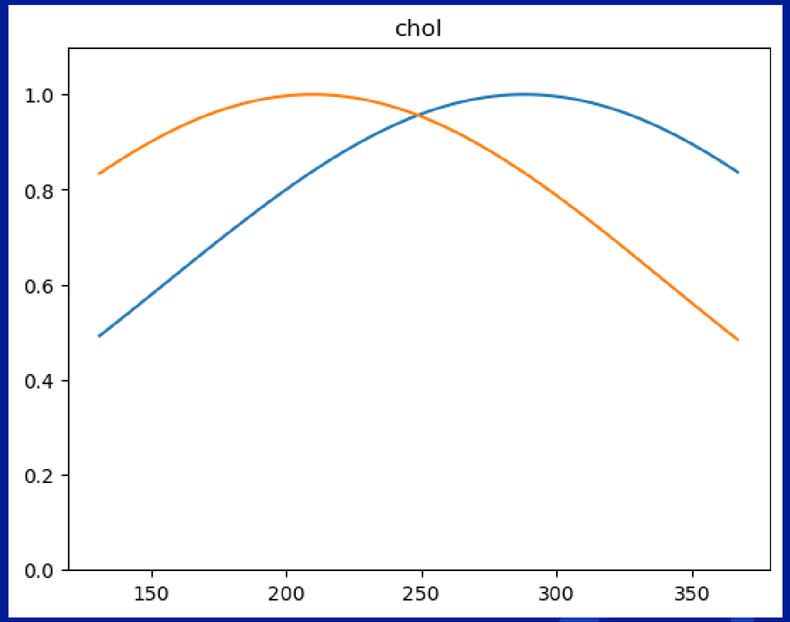


Figure: Chest pain and cholesterol membership functions

Fuzzy Inference System

- The obtained results are on-par with other used methods;
- This model is highly interpretable, we can easily deduct why it reached a specific result by looking at each variable's membership functions and at the consequents it uses.

Accuracy	Precision	Recall	F1-Score	Kappa Score
0.787 / 1	0.828 / 1	0.750 / 1	0.787 / 1	0.575 / 1

Table: Model's performance metrics

	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	constant
0	-0.262711	0.097674	-0.002009	-0.000513	0.025580	0.143485	0.002075	-0.178311	-0.075183	0.046707	-0.107495	-0.080919	0.972633
1	-0.152544	0.070931	-0.003319	-0.000658	0.188784	-0.018731	-0.002600	-0.120633	-0.051297	0.163675	-0.111168	-0.252651	1.317023

Table: Consequents used by the model

04

Deep Learning

Different Methods

Deep Neural Network

Three hidden layers :

64, 32 and 16 neurons

Long Short-Term Memory

Reshape and 2 layers:

64 and 32 neurons

Convolutional Neural Network

Reshape and two convolutional layers

Multi-Layer Perceptron

Different architectures:

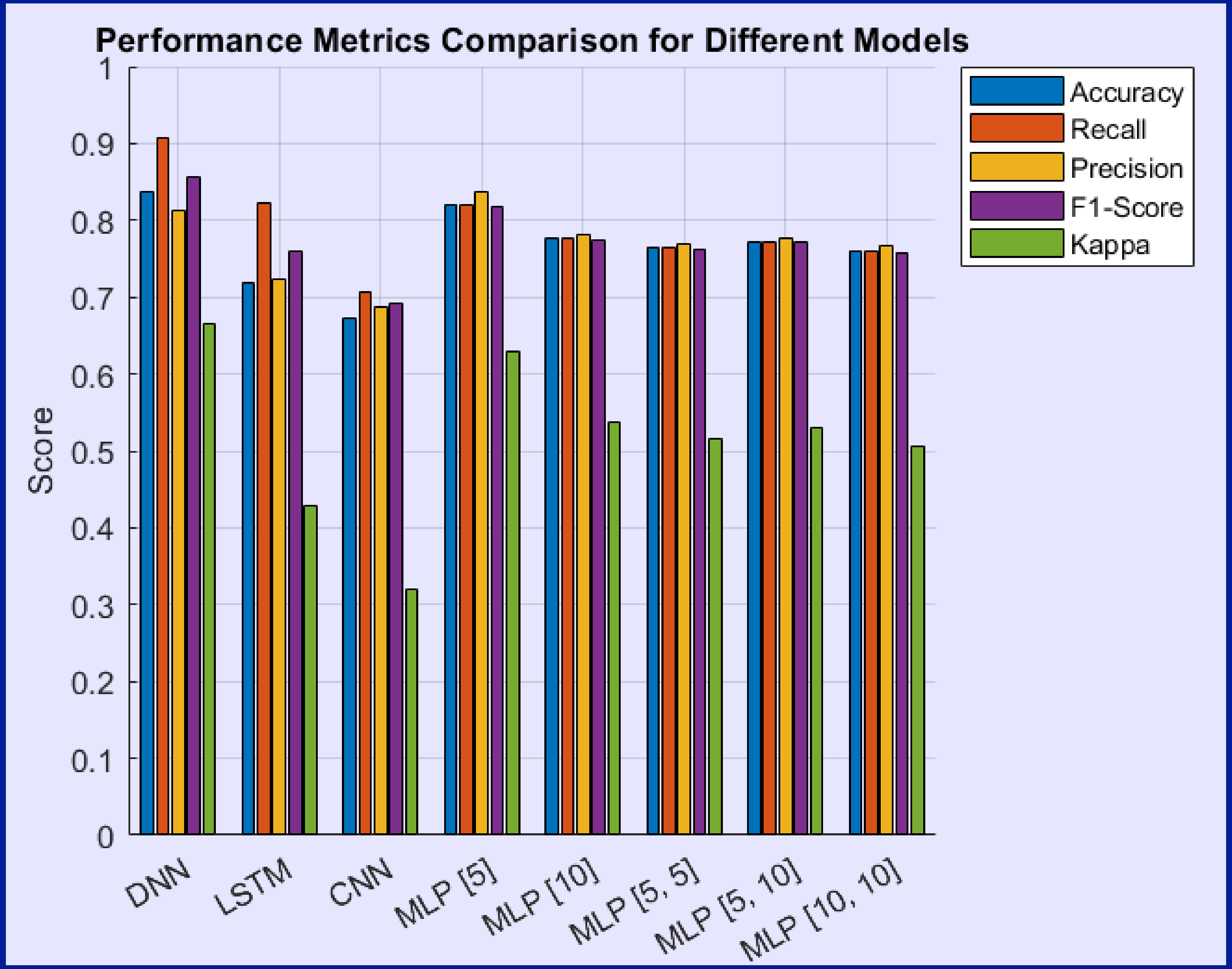
[5], [10], [5,5], [5, ,10], [10, 10]

Results

DNN outperformed the other models

MLP [5] has the higher Precision

Suggesting that less complex architectures can perform well



05

Final Results

Results

RFC shows a robust performance:

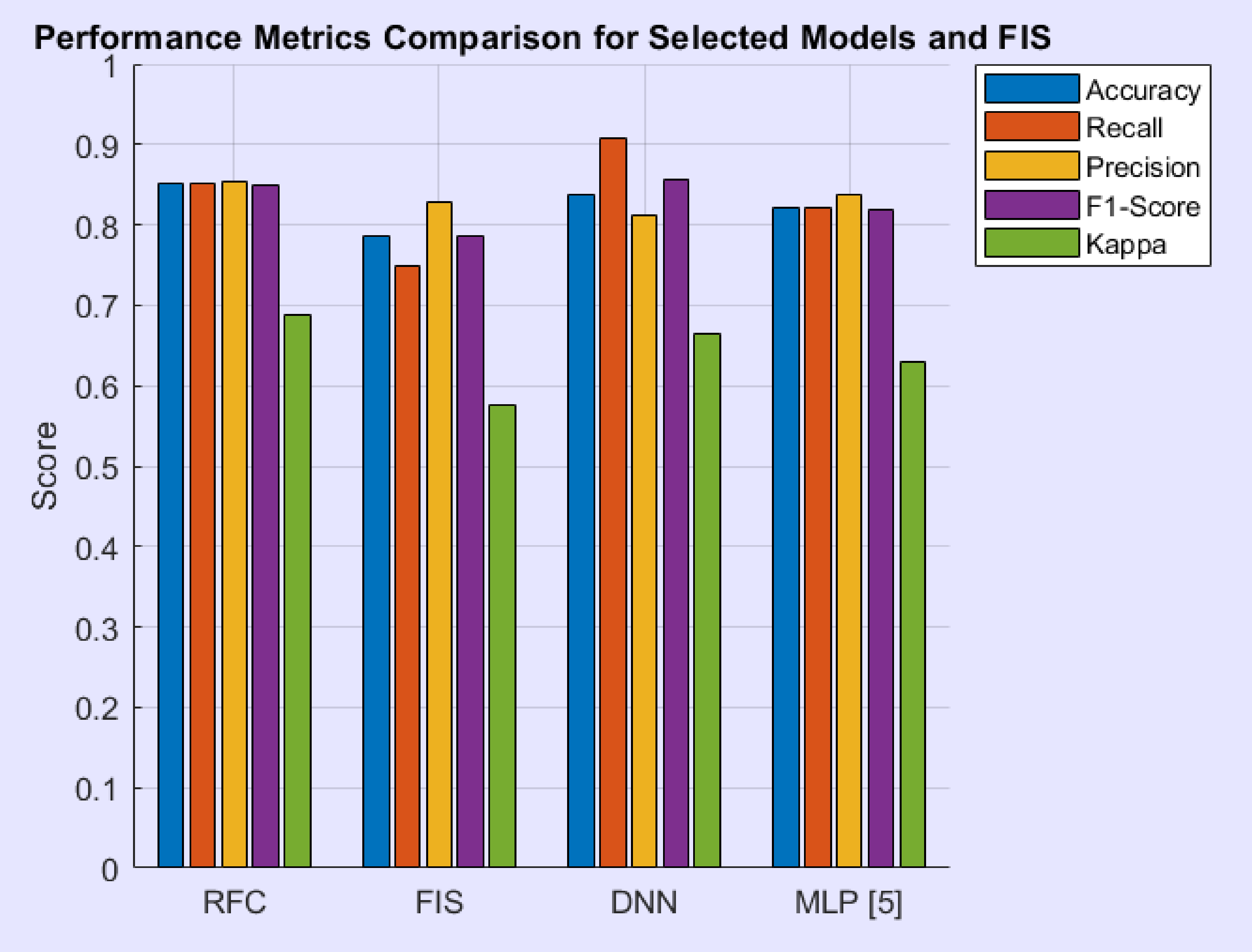
- Accuracy
- Precision
- Kappa Score

DNN also viable:

- Recall
- F1- Score

FIS – High Precision but Low Recall:

- Avoidable In a scenario like ours



Thank You