Student's name: Ngoc Hai Nguyen
Student's ID: M23.ICT.004
Lecturer's name: Nhat Quang Doan (Dr.)
Course: Machine Learning

Report for the Machine Learning Project

I. Introduction

Topic: Heart Disease Prediction – predict heart disease based on health information.

Objective:

- Data analysis: explain attributes and their meanings, how they have been encoded, proving the data is linear, and the impacts of attributes on the final result.
- Dimensional Reduction: Use feature selection and PCA to reduce the dimension of the data, then apply the SVM model for single attribute, pairs of attributes, PCA-applied data and sets of attributes and analyze the result.
- Based on the property of the data, applying SVM (with 4 kinds of kernel), Logistic Regression, Perceptron, and Linear Discriminant Analysis and comparing the result and discussion.

Method:

Give an assumption (data can be separated linearly), then analyze the data. After analyzing the data, predict with SVM with different sets of data. After that, choose linear models and test with SVM models. From the comparison between models, we can conclude for data and models.

II. Project process and experiment

1. Analyzing the data

1.1. Data description

The data include 13 attributes describing the value of health information, and the final result is 0 (no heart disease) or 1 (heart disease).

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1

Figure 1: A part of the data

The attributes are described below:

1. *Age:* The age of the patient in years.

2. *Sex:* The gender of the patient. (0 = female, 1 = male)

3. *Chest Pain Type (cp):* The type of chest pain experienced by the patient. It's categorized into four types: (0: Typical angina); (1: Atypical angina); (2: Non-anginal pain); (3: Asymptomatic)

4. *Resting Blood Pressure (trestbps):* The resting blood pressure of the patient measured in mm Hg (millimeters of mercury) upon admission to the hospital.

5. *Cholesterol (chol):* The serum cholesterol level of the patient was measured in mg/dl (milligrams per deciliter).

6. *Fasting Blood Sugar (fbs):* Fasting blood sugar level of the patient. (0 = blood sugar < 120 mg/dl, 1 = blood sugar > 120 mg/dl)

7. *Resting Electrocardiographic Results (restecg):* Result of the resting electrocardiogram (ECG) of the patient. It's categorized into three types: (0: Normal); (1: Abnormality in ST-T wave (T wave inversions and/or ST elevation or depression of > 0.05 mV)); (2: Showing probable or definite left ventricular hypertrophy by Estes' criteria.)

8. *Maximum Heart Rate Achieved (thalach):* The maximum heart rate achieved by the patient during exercise.

9. *Exercise Induced Angina (exang):* Whether the patient experienced exercise-induced angina. (0 = no, 1 = yes).

10. *ST Depression Induced by Exercise Relative to Rest (oldpeak):* ST depression induced by exercise relative to rest. It indicates the degree of abnormality in electrocardiograms during exercise.

11. *Slope of the Peak Exercise ST Segment (slope):* The slope of the peak exercise ST segment. It's categorized into three types: (0: Upsloping); (1: Flat); (2: Downsloping)

12. *Number of Major Vessels Colored by Fluoroscopy (ca):* The number of major vessels (0-3) colored by fluoroscopy. These vessels typically refer to the major arteries of the heart.

13. *Thalassemia (thal):* A blood disorder affecting the amount of oxygen carried by the blood. It's categorized into three types: (1: Normal); (2: Fixed defect); (3: Reversible defect)

14. *Target:* The presence of heart disease. (0 = no heart disease, 1 = heart disease)

Observation: Data is encoded from health information. Theoretically, the data were encoded to accommodate the use of linear models for prediction. The parameters are gradually increasing or decreasing according to the degree of influence on the results.

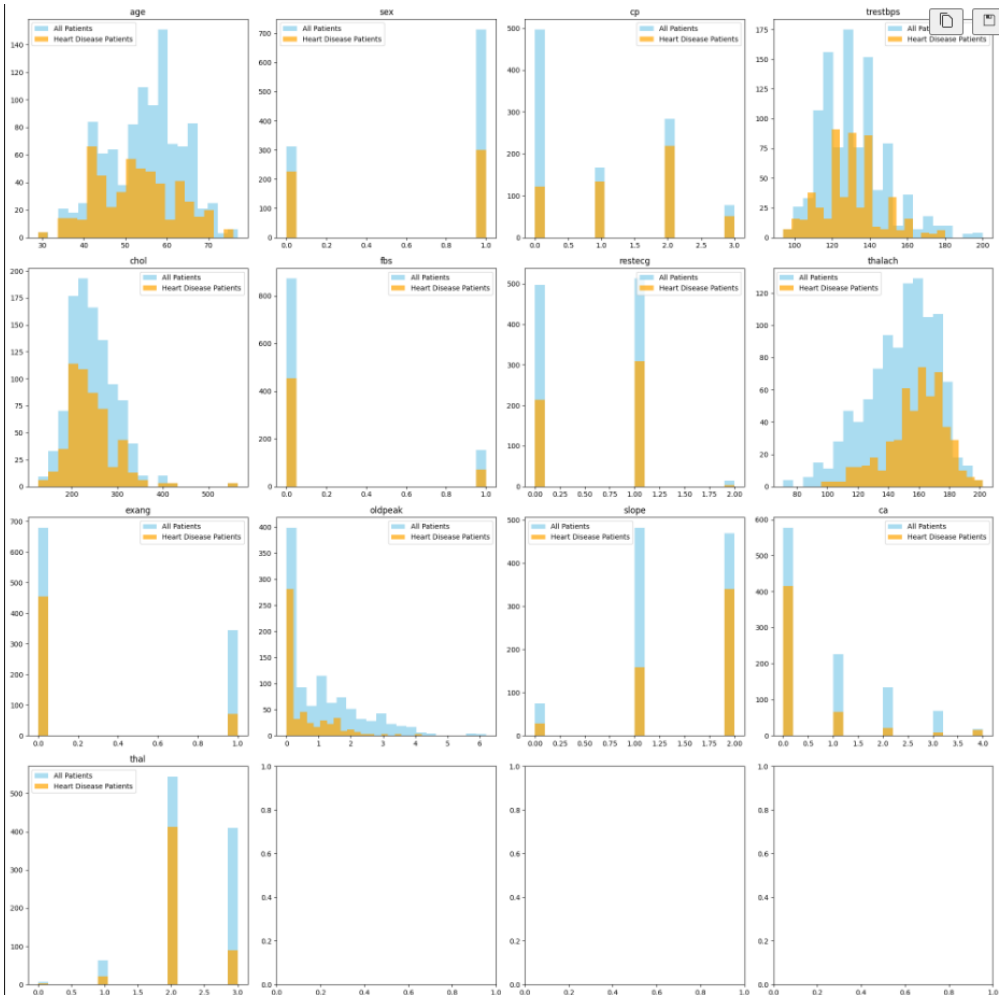

Figure 2: Attribute's distributions and heart disease percentages

1.2. Data analysis

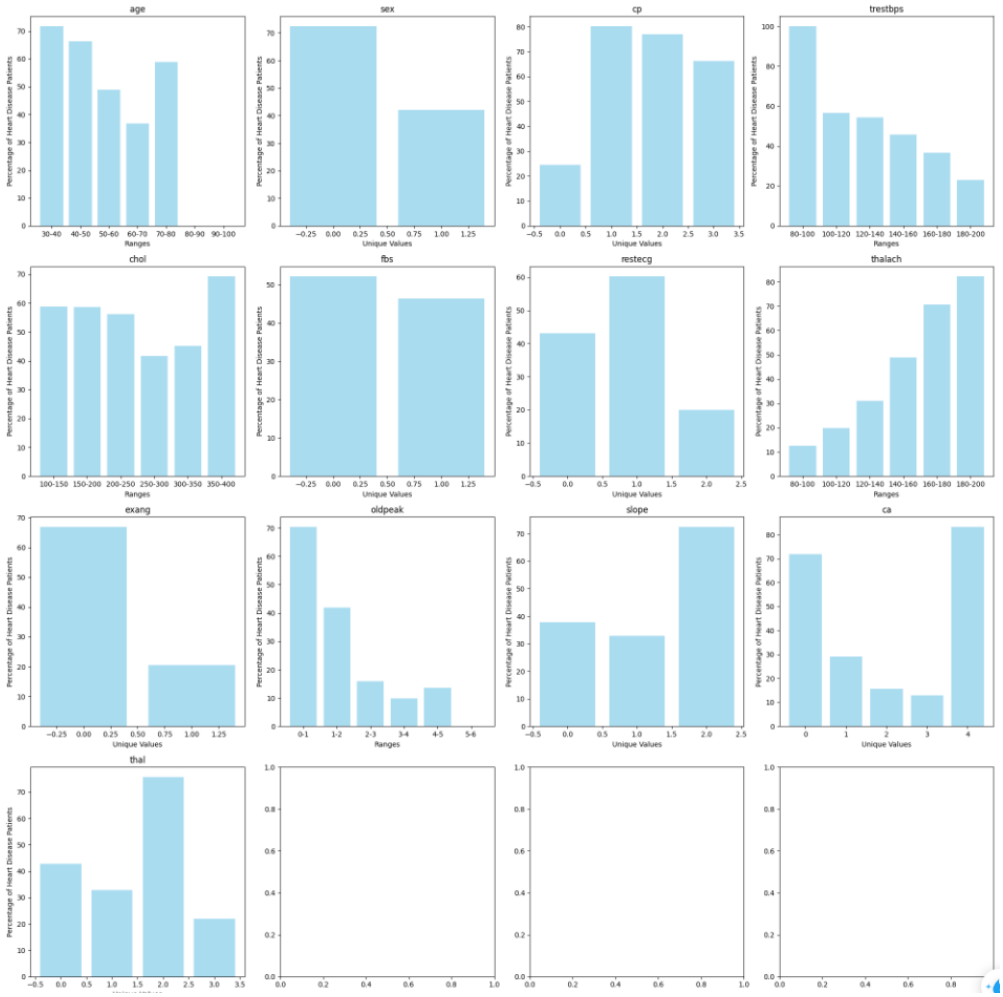

Figure 3: Percentage of heart disease patients in each attribute

From the above graphs, we can observe that many attributes can be used in linear models to predict the final result because with higher (or lower) values, the patient has higher accuracy to be heart disease. For example, The patient have higher accuracy to be heart disease if they have value of thalach > 140, so that 140 can be the decision boundary, combined with trestbps, we can have a linear line as the decision boundary, and it will be hyperplane when we apply linear model for many attributes.

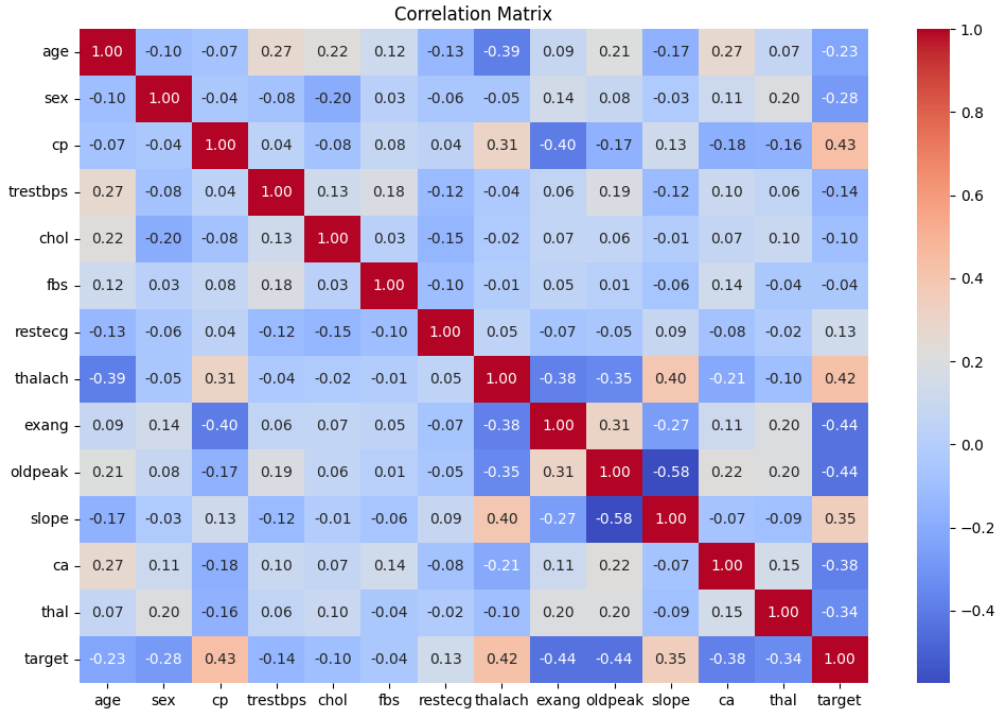

Figure 4: Correlation matrix

From the correlation matrix, we can see that cp with 0.43, thalach with 0.42, exang and oldpeak with -0.44, ca with -0.38, thal with 0.34 are the attributes having strong linear relationship to the final result. They have the biggest impact on the result and can be selected when applying dimensional reduction.

To prove it, we can use each attribute to predict final result using SVM model with kernel = ‘linear’, the result is showed below:

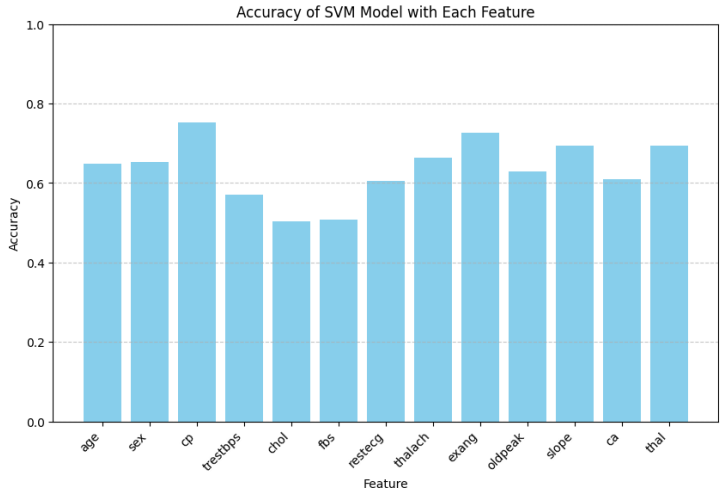

Figure 5: Accuracy of the SVM linear model with each attribute

Conclusion: From the data analysis, we can observe that the linear model can be applied to prediction for this dataset. Based on the correlation matrix, we can see which attributes have a strong linear relationship to the result. We can observe that if we use only 1 attribute for prediction, the accuracy is quite low, so this prediction just can be used to evaluate the level of impact of each attribute on the result.

2. Applying dimensional reduction with prediction

2.1. PCA

We will apply PCA to reduce the dimension of data to 2, then take the prediction. From the below result, we can see that the PCA method can keep the linearity of the data, and we can find the linear boundary for the applied PCA data, the accuracy is about 77%, it is equivalent to the highest accuracy of pair of attributes – presented below.

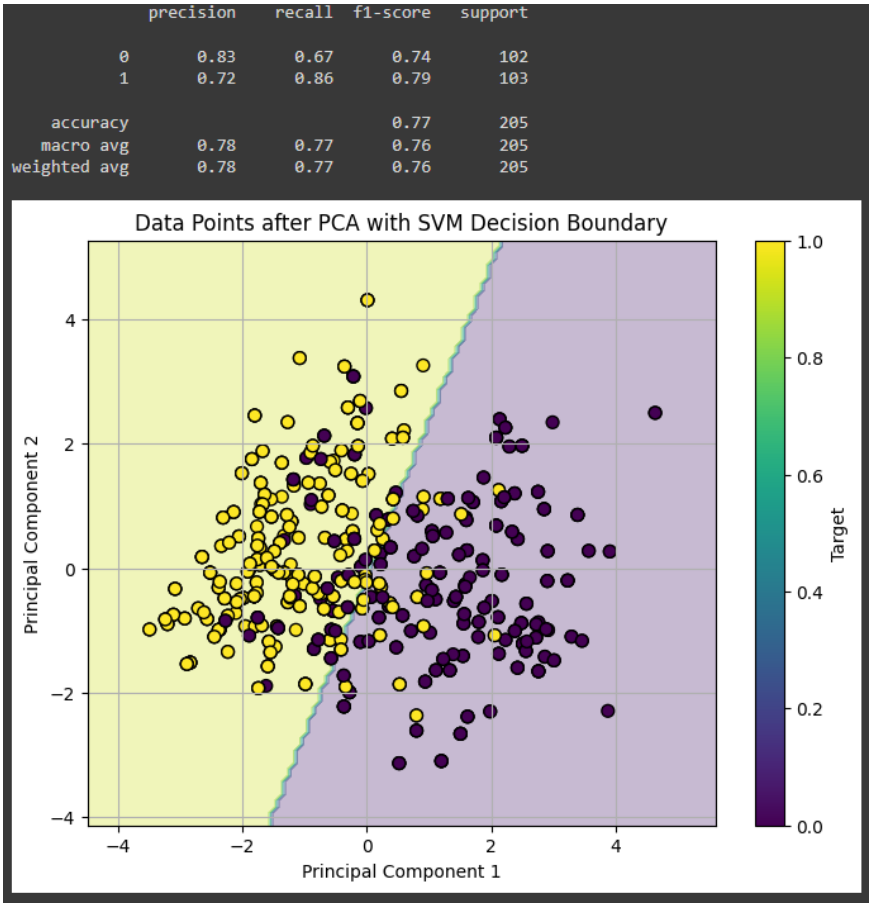

Figure 6: PCA-applied data prediction

2.2. Features selection

Based on the data analysis, we can see the attributes having the most effect on the result and select them to predict the final result. First, we will select the pairs of them and make predictions. To prove the linearity of the data, we will apply 2 kinds of kernel for SVM model – "linear" and "RBF – default if kernel is blank in code".

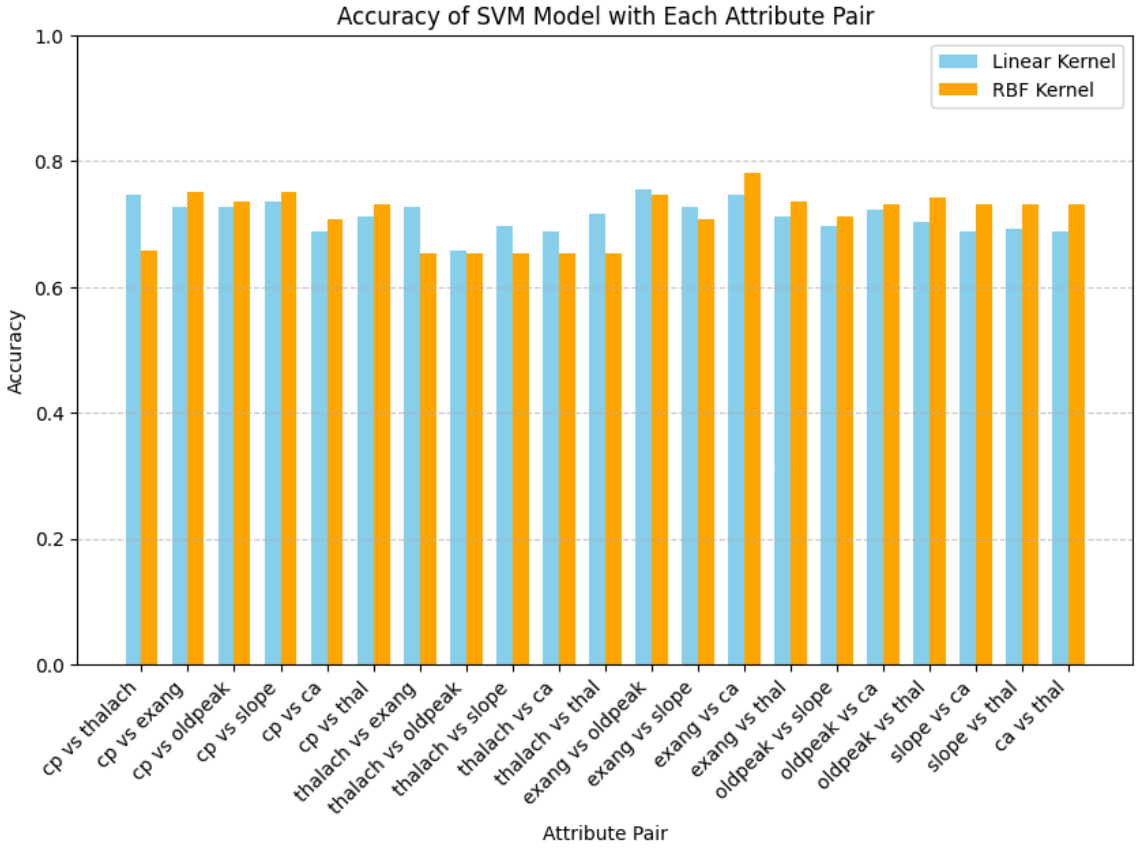

Figure 7: SVM model with pairs of attributes

Observation: for each pair, we have different results and we can conclude that linear model or non-linear is better in this case. We can observe that for linear model, pair of ca and thalach have the highest accuracy ~ 72% and for RBF, pair of exang and ca have the highest accuracy ~ 77%, but all SVM models above have accuracies lower than 80%, so we need more than 2 features to have a better model.

We can use 3 kinds of data: all of the highest impact attributes, PCA-applied data, and all attributes of data to predict. We apply 2 kinds of kernel for SVM – linear and RBF(non-linear). The results of these models are compared below:

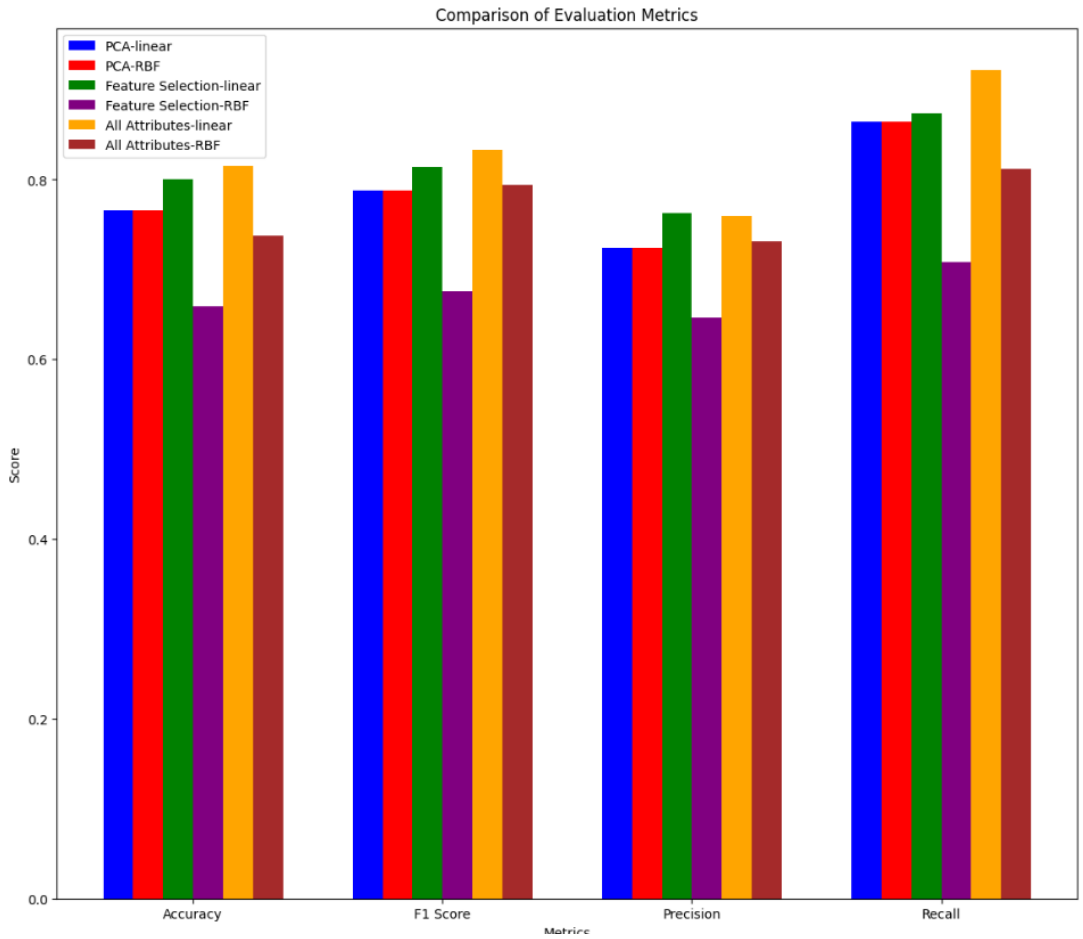

Figure 8: Comparison between models with 3 kinds of data

Conclusion: All attributes with the linear model give us the best model. Feature selection can be applied with lower performance, but we can see that with the linearity of data proved by above result, all linear model have better results than non-linear model. We can observe that we should use sets of attributes or all attributes will have better results than PCA or pairs of attributes. When we use pairs of attributes of PCA data, the dimension of data decreases too low, which means we lose a lot of features of initial data and leading to lower accuracy.

3. Applying different models and evaluation

For the SVM model, we have many kernels that can be used. For this comparison, linear, rbf, polynomial and sigmoid are used kernels. Not only SVM models but also other models

like Logistic Regression, Perceptron and Linear Discriminant Analysis are used for this data set, the results are showed below:

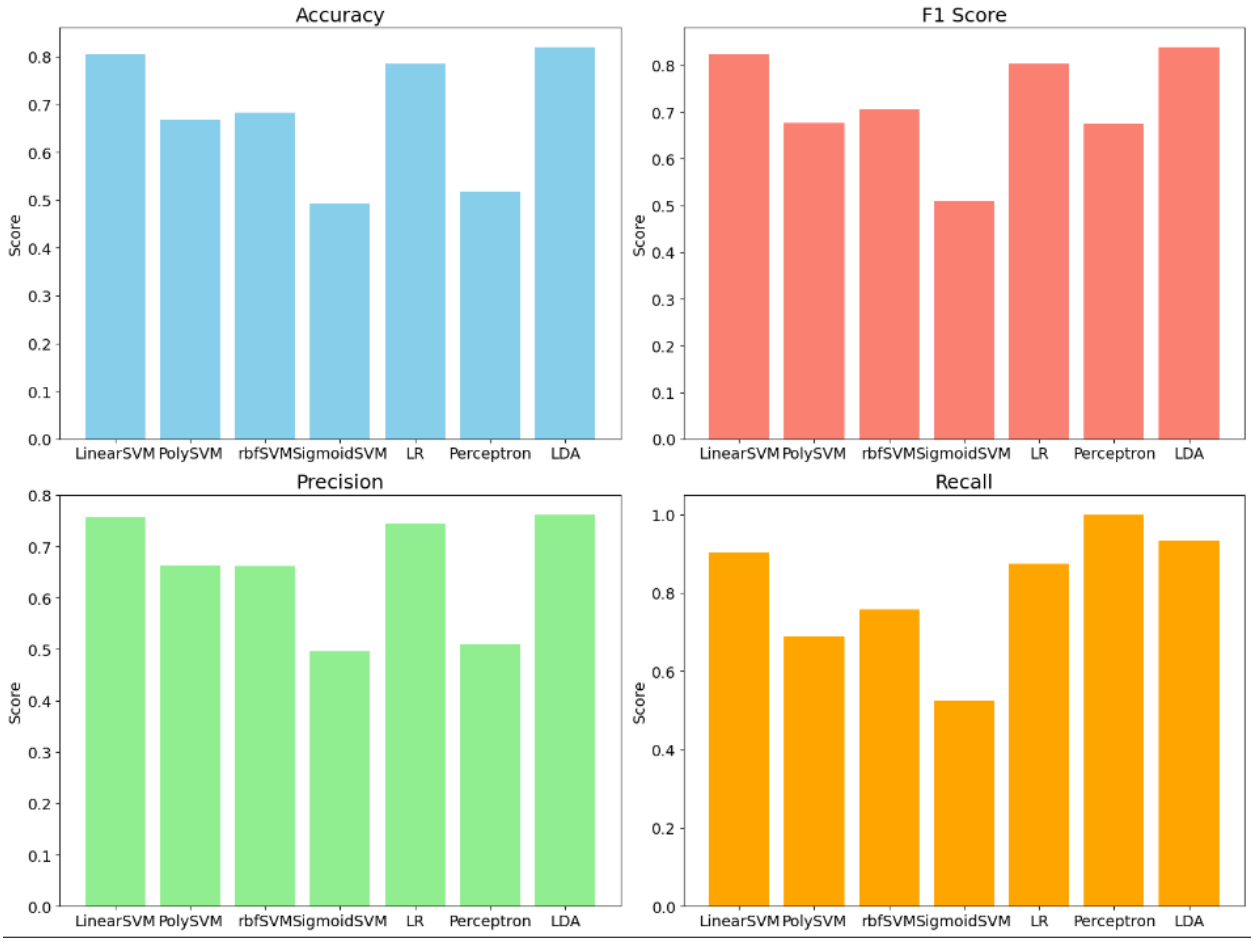

Figure 9: Comparison between models

Observation:

+ We have 3 best models that are Linear-SVM, Logistic Regression, and Linear Discriminant Analysis because they are appropriate with the property of the dataset.

+ While the remaining models have worse results, the decision boundary of Polynomial SVM and rbfSVM are not hyperplanes, which means that the decision boundary may fit too much to the data, it has low bias but leads to high variance, and finally overfitting. With SigmoidSVM, the shape of the decision boundary is unchanged, so It leads to high bias and high variance for this data set.

+ With Perceptron, this is an algorithm that uses a very simple learning approach, iteratively adjusting its weights whenever it misclassifies an example. This simplicity can

make it less robust compared to other linear models, particularly if the data is noisy or not perfectly linearly separable. The Perceptron can converge quickly to a suboptimal solution, especially if the data isn't perfectly linearly separable. We can see that this data is not linearly separable perfectly, it leads to low performance for Perceptron.

From this result, we can conclude that this dataset works well with Linear models, however, the linearity is not at all attributes, so we don’t have high values of evaluation metrics.

III. Conclusion and discussion

For heart disease dataset:

- The data is encoded well to fit some linear models, but it is not linearly separable perfectly.
- We can analyze each attribute and evaluate their importance to the final results, and give us information to select appropriate strategies for dimensional reduction. In this dataset, we tried dimensional reduction with PCA and feature selection, both work well on this dataset because they returned acceptable results.
- The highest impact attributes work well with linear models, and they are the most important aspect when predicting whether a patient has heart disease or not.
- PCA-applied data keeps the linearity of the data and it can be separated linearly.
- The Linear models are appropriate for this dataset, however, they can not have good evaluation metrics because the dataset is not linearly separable at all attributes, It lead to noises when applying linear models.
- The best models are linear models with accuracies are about 80%. In the medical aspect, they are acceptable, because they give the patient the potential to have heart disease with not too much health information. The patients and doctors can predict heart disease earlier and have appropriate heart examinations and appropriate treatments.

For a Machine Learning project:

- The way to encode data impacts the way to choose the model for the dataset. So we need to know how to have appropriate methods to encode data.
- Data analysis is very important. It can tell us so much information about the data (which attributes are most important, the attribute's effects on the final result, the properties of the data: linear or non-linear, ….). By analyzing the properties of the data set, we can easily choose set of appropriate models to test with the dataset.
- We can apply dimensional reduction to help our dataset better by removing the noise, … but we shouldn't reduce them too much, because it can lead to loss properties of the initial dataset.
- We should apply the dataset with many of the models and compare the results between them to choose the most appropriate model for the dataset. No models are perfect for any dataset, so we need to try with a set of suitable models to select the best one.