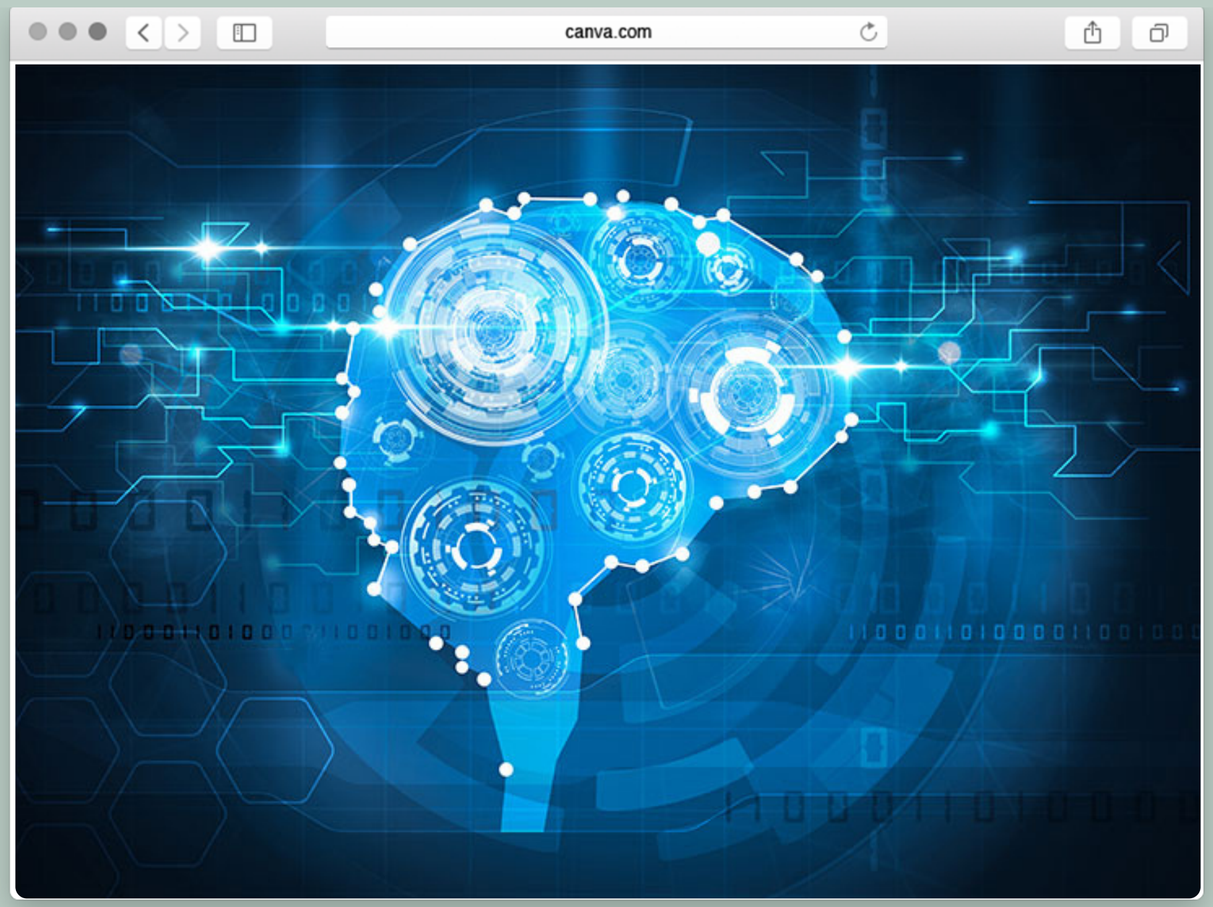
MACHINE LEARNING MODELS

GROUP 2

PUN YI JIE
FOO KE TING
GIAM HUI JIA
LEE LER YUNG
ANG SUN YEE
GAN YI HANG



Abstract

Heart-related diseases are the leading cause of death and have emerged as the most-threatening disease in the whole world.

To identify the most effective machine learning algorithm in predicting the occurrence of heart disease.

heart_disease dataset

age, sex, cp, trestbps → occurence of heart disease

3

K-Nearest Neighbours (KNN)
Decision Tree
Support Vector Machine (SVM)

K-Nearest Neighbours (KNN)

(lazy learner algorithm)

Supervised learning method

Classification Regression

(classifies the new data based on a similarity measure)

- Predicts the proper class for the test data by computing the distance between the test data and all of the training points

- Choosing the K number of points that are the most similar to the test data by assuming similar things exist in close proximity.

Decision Tree

Classification tree

Target variable can take a finite set of values or is categorical

Regression tree

Target variable can take continuous values

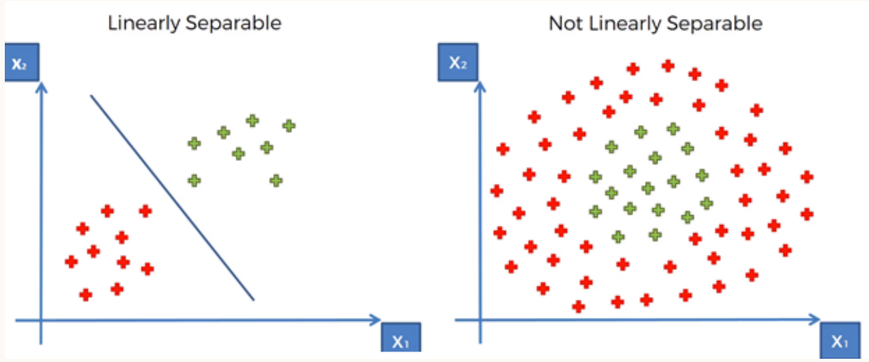
- Flow like a tree structure that works on the principle of conditions as well as a graphical representation of all the possible solutions.

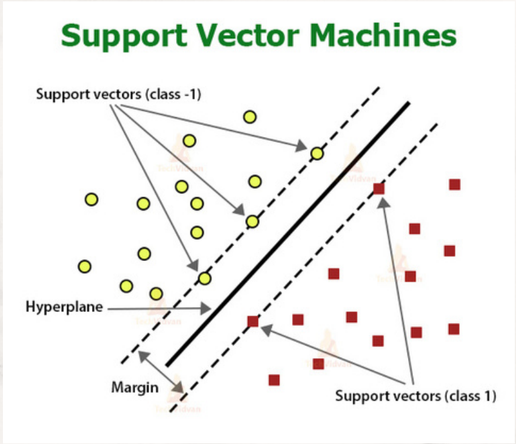
- We are able to pick the starting test condition, which is where an attribute should have the highest information gain to be selected for splitting.

  - Entropy
  - Information Gain

Support Vector Machine (SVM)

- Responsible for finding a hyperplane to separate different classes and maximize the margin.
- Support vectors - data points which can also be used to maximize the margin of the classifier.
- Classify non-linear data by using a kernel function to map the non-linear data to higher dimensions so that it becomes linear to find the decision boundary.





Objective

To identify the most effective machine learning model in predicting the occurrence of heart disease.

JUSTIFICATION OF CHOICES

K-Nearest Neighbours (KNN)

Decision Tree

Support Vector Machine (SVM)

K-NEAREST NEIGHBOURS (KNN)

- simplicity
- low processing time
- additional assumption is not needed

DECISION TREE

- not much data preparation
- handle non-linear dataset effectively
- no data assumption

SUPPORT VECTOR MACHINE (SVM)

- lower risk of overfitting
- robust to outliers
- effectively deal with complex problems or non-linear data

STEPS OF BUILDING MACHINE LEARNING MODELS

01 Data Exploration

02 Data Preparation

03 Machine Learning Modelling

04 Evaluation

01 DATA EXPLORATION

Exploratory data analysis (EDA)

Numbers of rows and columns

df.shape

Features

df.columns

Data types

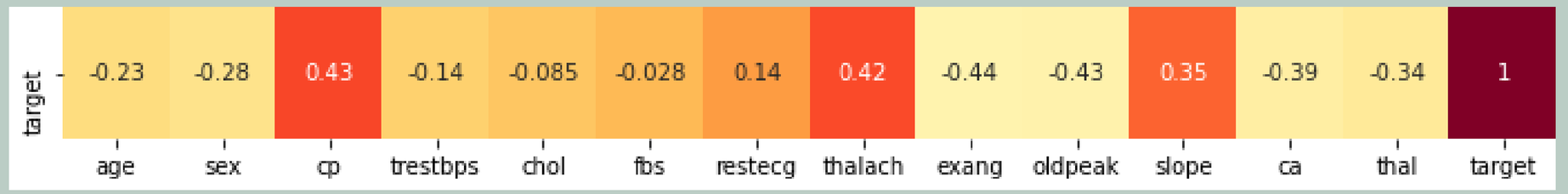
df.dtypes

Presence of null values

df.isnull().sum()

02 DATA PREPARATION

- **Feature Selection**
  Reduce overfitting and improves accuracy



Chosen features: exang, cp, oldpeak, thalch, ca, slope and thal

- **Data Splitting**
  60% Train Set/40% Test Set

  70% Train Set/30% Test Set

  80% Train Set/20% Test Set

- **Feature Scaling/Standardization**
  To normalize the range of features of data

03 MACHINE LEARNING MODELLING

- Determine value of k for KNN
- Determine maximum depth for Decision Tree
- Classification Methods
  KNeighborsClassifier
  DecisionTreeClassifier
  SVC
- Fit and Predict

```
knn3.fit(x_train80,y_train80)
knn3.predict(x_test20)
knn3.predict(x_train80)
```

04 EVALUATION

- **Confusion Matrix and Accuracy Score**

  Example: KNN

```
80% Train & 20% Test
Confusion Matrix:
 [[25  2]
 [ 4 30]]
Accuracy: 0.90164
```

- **Check for Overfitting**

```
80% Train & 20% Test
Accuracy for test set: 0.90164
Accuracy for train set: 0.86777
```

If the accuracy score for the training set is significantly better than the testing set, then the model is probably overfitting.

Comparison and Recommendation

Comparison

- **Number of rows of codes used**
  KNN and Decision Tree have longer codes compared to SVM as both algorithms need to calculate suitable k-value and maximum depth value respectively. KNN has a longer code compared to Decision Tree because KNN needs to plot a graph to determine a suitable k-value with the highest accuracy.

- **Assumption needed**
  KNN needs to assume the k-value whereas the other two machine learning algorithms do not need to make any assumption.

- **Accuracy**
  KNN has the highest accuracy for 60%, 70% and 80% training set compared to those of Decision Tree and SVM. KNN also has more stable accuracy than Decision Tree and SVM.

Recommendation

K-Nearest Neighbours (KNN)

Highest accuracy

Closest accuracy
among three divisions
of data

Shorter code

Results and Discussion

Accuracy of Each Machine Learning Model Built

	60% Train set 40% Test set	70% Train set 30% Test set	80% Train set 20% Test set
KNN	0.82787	0.86813	0.90164
Decision Tree	0.79508	0.81319	0.85246
SVM	0.81148	0.81319	0.85246

Comparisons between Accuracy of Train Set and Test Set for Overfitting Checking

	KNN		Decision Tree		SVM	
	Test set accuracy	Train set accuracy	Test set accuracy	Train set accuracy	Test set accuracy	Train set accuracy
60% Train set 40% Test set	0.82787	0.86188	0.79508	0.86188	0.81148	0.90055
70% Train set 30% Test set	0.86813	0.85849	0.81319	0.86792	0.81319	0.88208
80% Train set 20% Test set	0.90164	0.86777	0.85246	0.85124	0.85246	0.87603

Lesson learnt and conclusion

LESSON LEARNT

Experience

Time Management

Self-learning

Teamwork

CONCLUSION

- **Decision Tree and SVM are used for comparison**

  Both have obtain high accuracy and not overfitting
  Decision Tree requires less time and effort
  SVM perform well in complex datasets.

- **KNN as the most suitable algorithm**

  KNN has the highest accuracy for all set of data splitting.
  No sign of overfitting
  Easy to use

Thank you