PRESENTATION

FINAL PROJECT REPORT

BY GROUP 1 - DAP391- A1703
LECTURER: HAODD

Welcome To Our PRESENTATION

TEAM MEMBERS

1. Nguyen Ngoc Vu Thong
2. Pham Huynh Quy An
3. Bui Nhat Tan
4. Huynh Duc Tinh
5. Nguyen Van Thanh Thong

TEAM MEMBERS

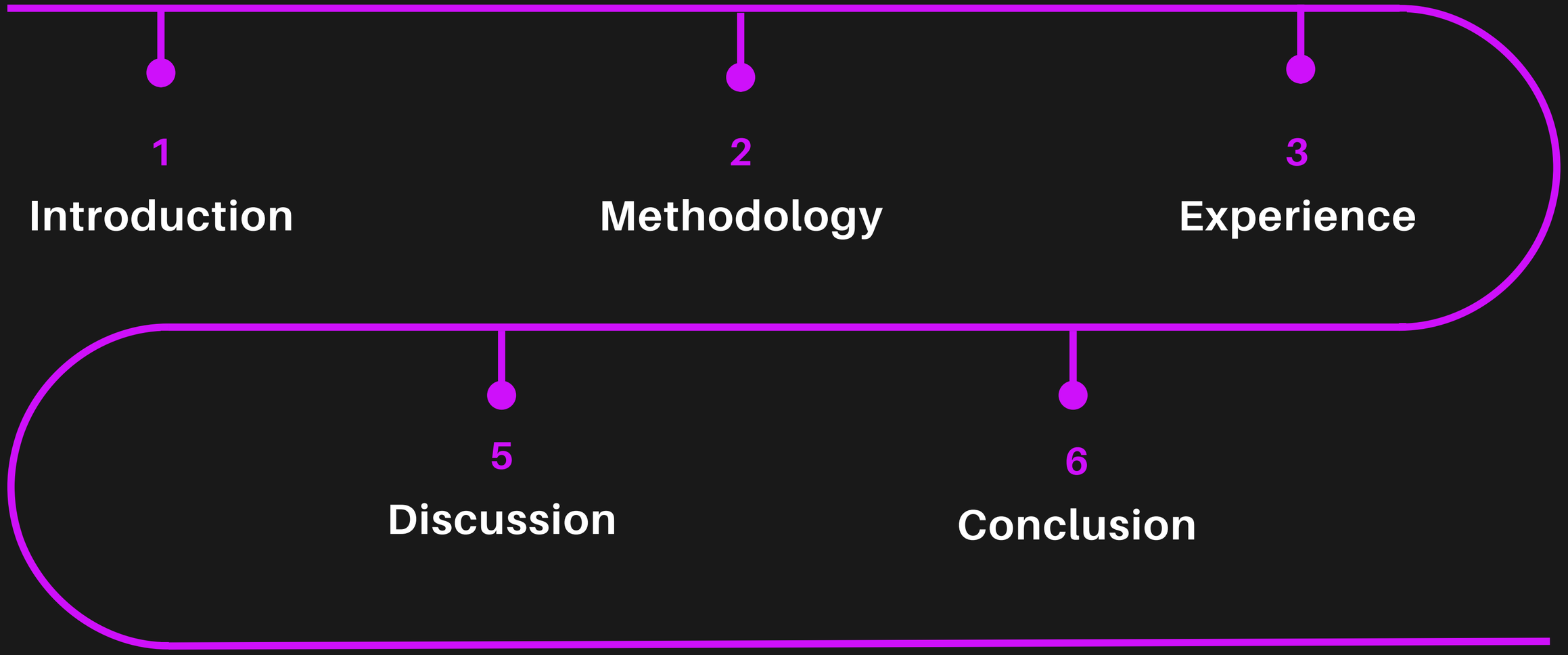
6. Nguyen Hai Long
7. Tran Duyen Hong Minh
8. Nguyen Quoc Anh
9. Nguyen Khac Anh Duc

Topic

APPLYING REGRESSION MODEL TO PREDICT HEART DISEASE

Tue 18/07/2023

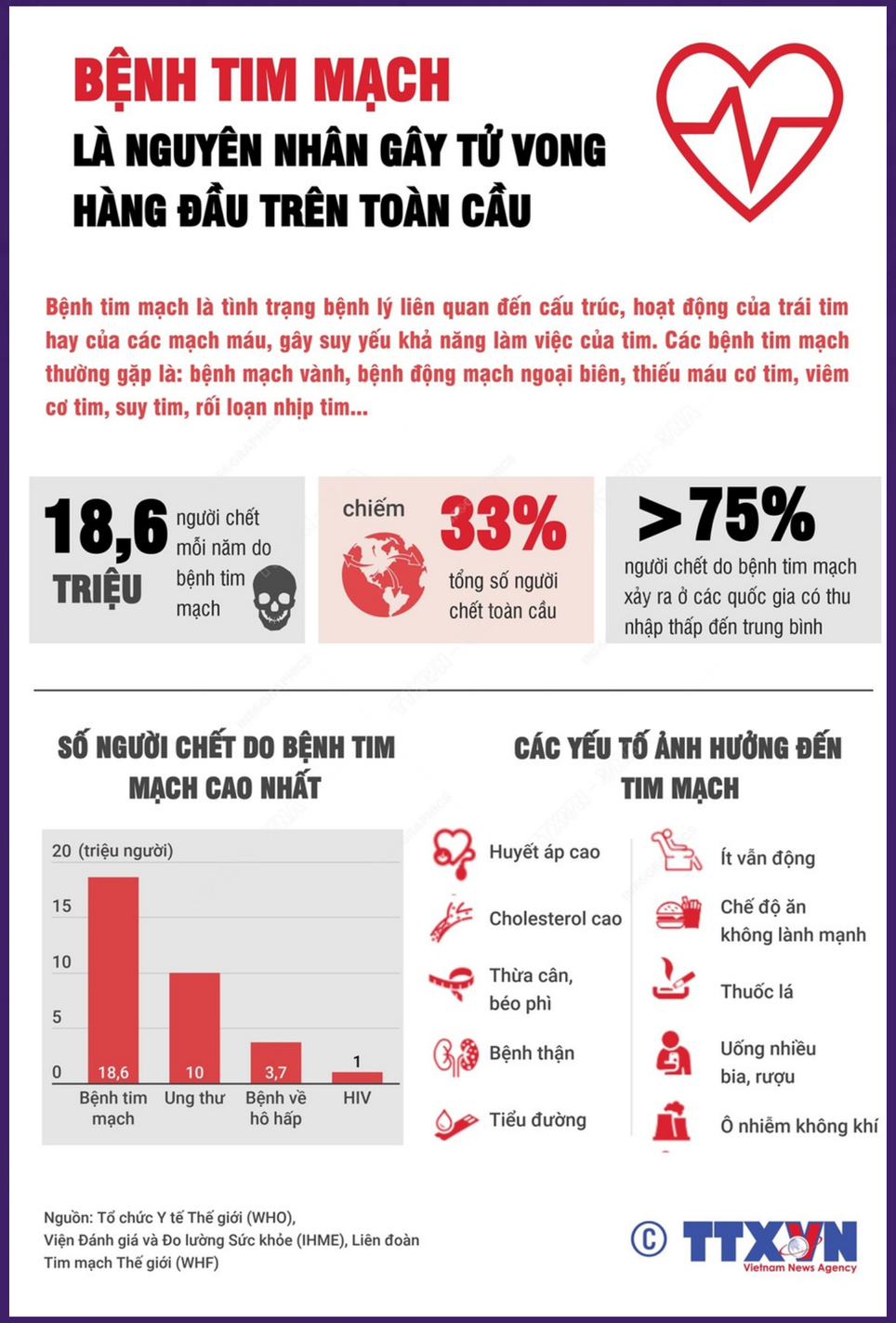
OUR CONTENTS



1. Introduction

- Why heart disease?
- Associated with several contributory risk factors, difficult to diagnose on time.
- Huge data, very noisy, hard for humans to understand. But easy for machine, algorithm.

=> Focus to develop a predictive model of heart disease using multiple prediction algorithms
=> Compare and evaluate
=> Apply best model to our GUI that allows input factors then predict



2. Methodology

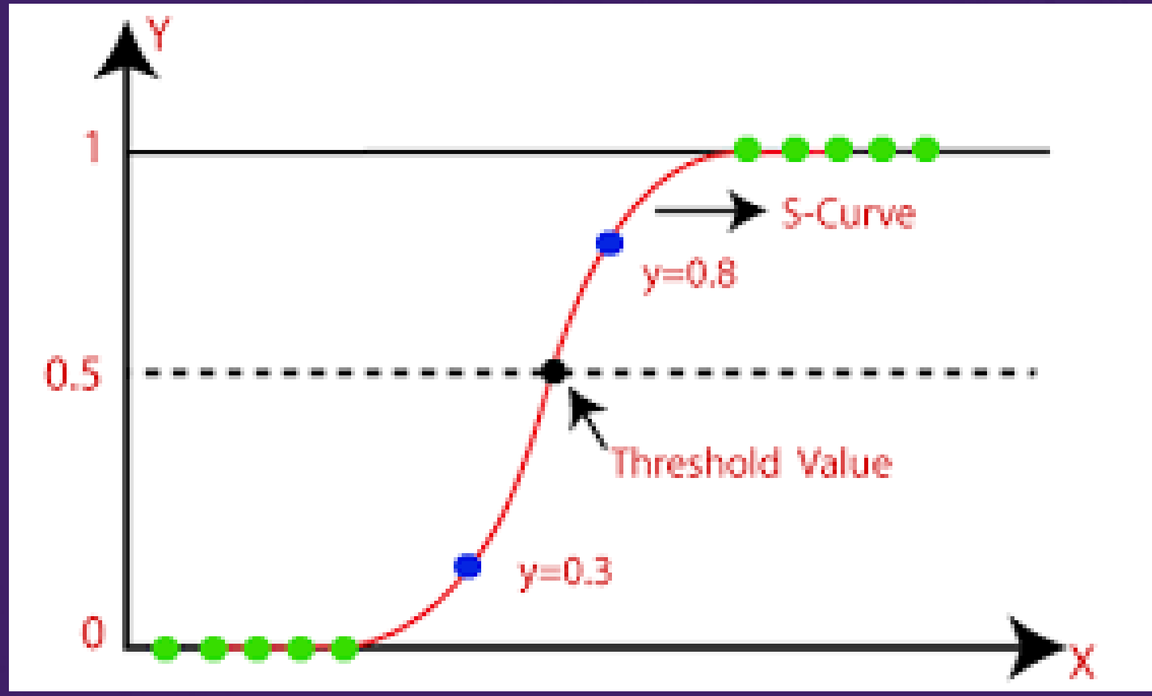
- Objective:
  - Develop a predictive model for coronary heart disease (CHD) detection
  - Compare the performance of different algorithms
  - Dataset: Medical records with demographic information, clinical measurements, and CHD outcomes
- Data Preprocessing:
  - Handle missing values
  - Balance dataset using SMOTE
  - Normalize numerical features
  - Encode categorical variables
  - Apply feature scaling and standardization techniques

2. Methodology

Logistic Regression:

- Models the probability of an event based on independent variables
- Used for classification and predictive analytics
- Utilize *LogisticRegression* class from scikit-learn

  - Time complexity: Depends on the number of iterations required to converge, typically linear or slightly higher.
  - Space complexity: Low as it does not require significant memory to store parameters.

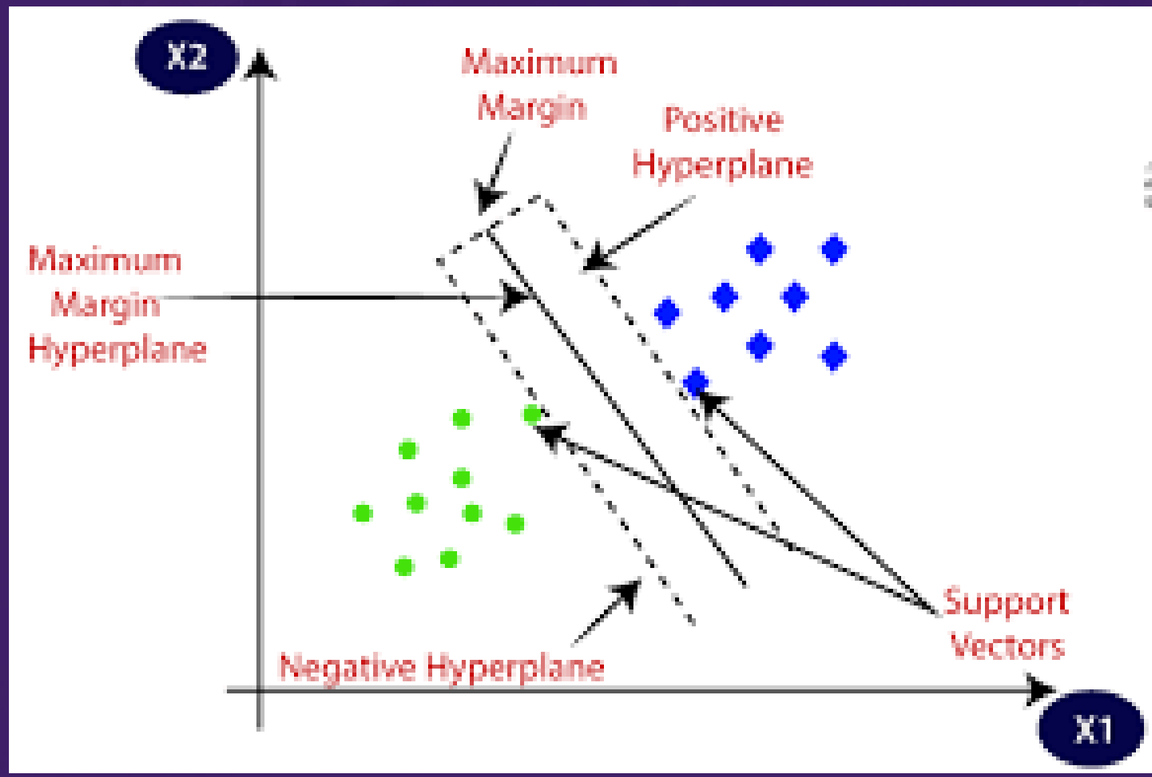


2. Methodology

Support Vector Machines (SVM):

- Finds a hyperplane with the largest margin to separate data
- Used for classification and regression
- Use *SVC* class from scikit-learn

  - Time complexity: Depends on the number of support vectors, typically between O(N^2) and O(N^3), where N is the number of training samples.
  - Space complexity: Moderate as it requires storing the support vectors.

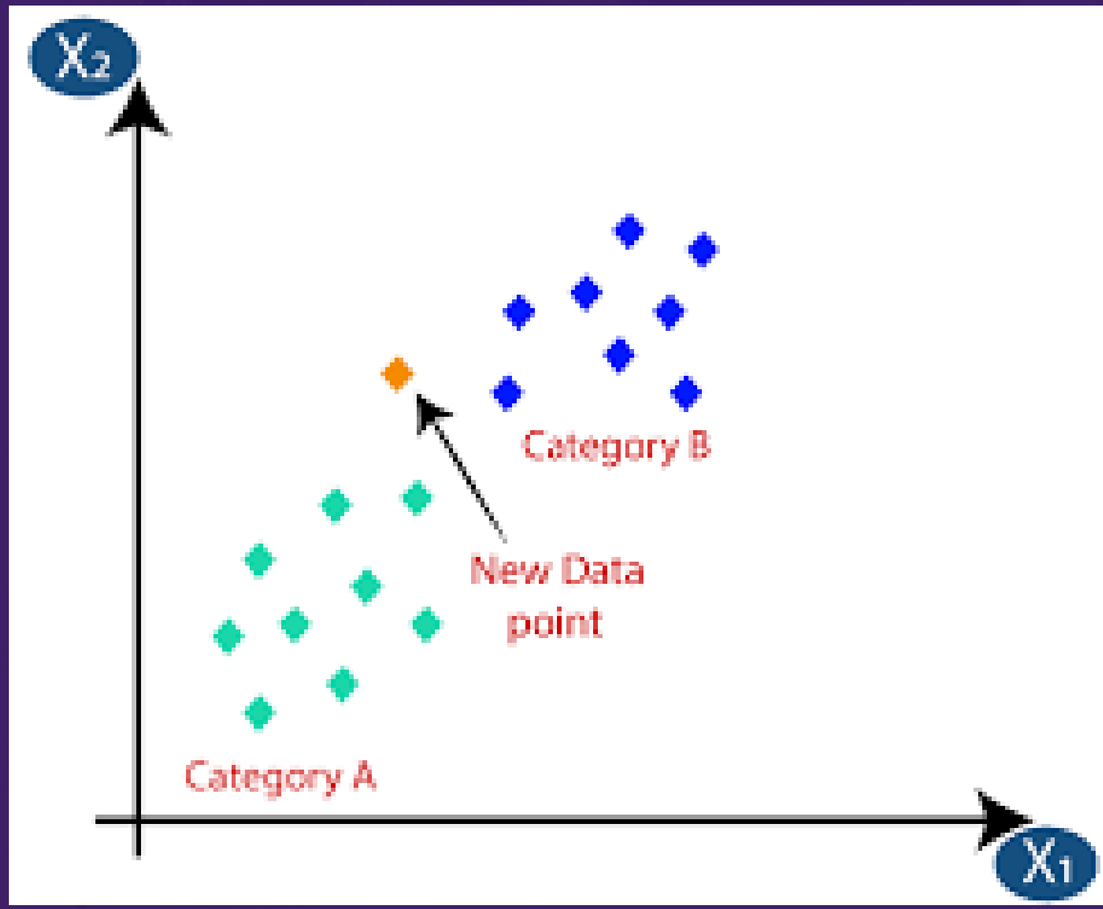


2. Methodology

K Neighbors Classifier

- Finding the k nearest neighbors of a new data point and assigning it the majority vote of their labels (for classification)
- A non-parametric and instance-based method
- It does not make any assumptions about the data distribution, and it stores all the training data in memory.

- Time complexity: The time complexity of KNN for classification is O(N × M)
- Space complexity: The space complexity of KNN is high as it requires storing the entire training dataset.

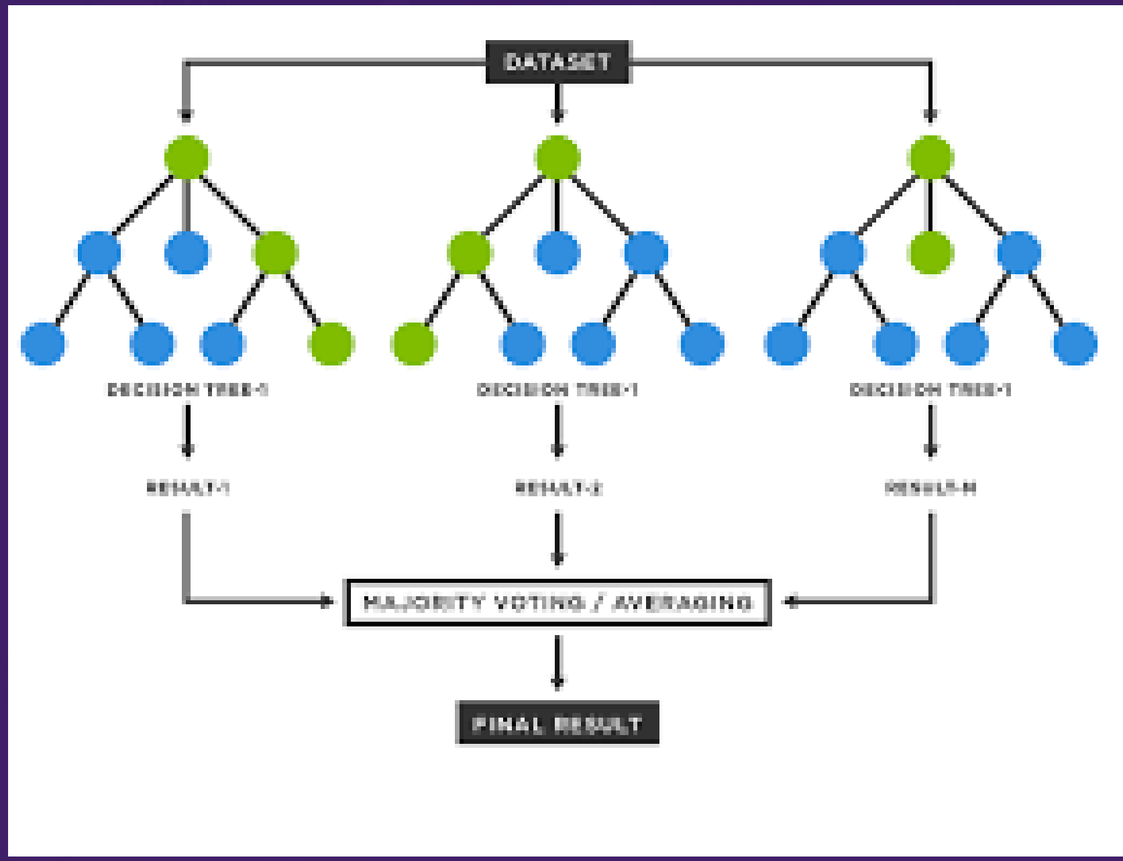


2. Methodology

Random Forest Classifier

- A model that uses multiple decision trees to make predictions.
- In classification tasks, each tree within the random forest contributes a classification or "vote", and the forest ultimately selects the classification with the highest number of "votes".

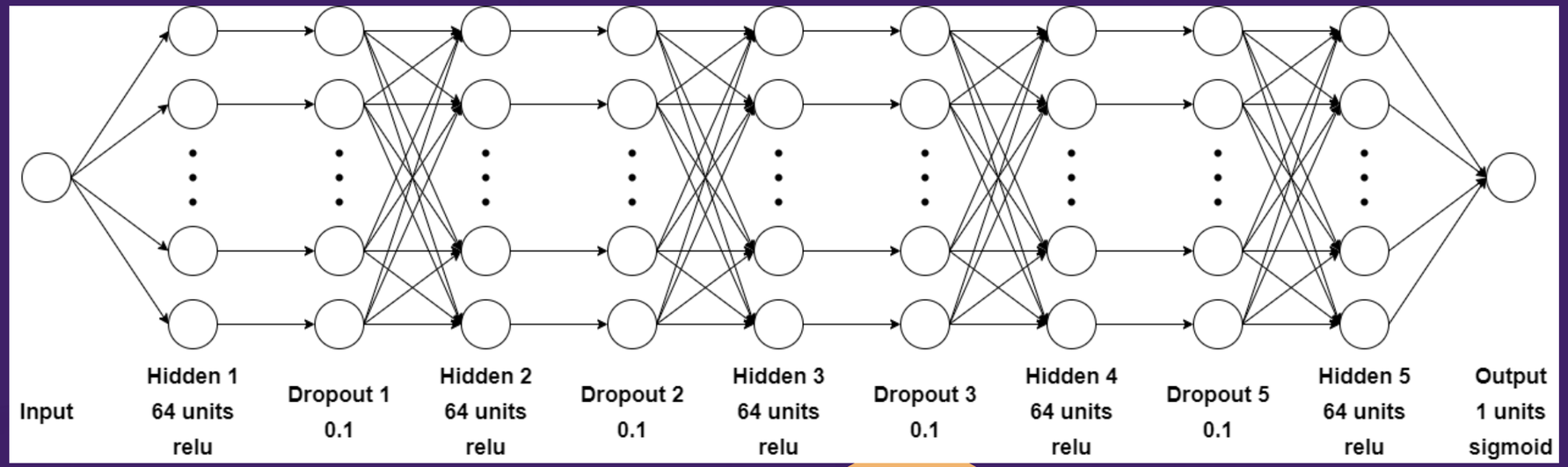
- Time complexity: The time complexity of training a random forest is O(N × M × log(N))
- Space complexity: The space complexity of random forest is high as it requires storing multiple decision trees.



2. Methodology

Deep Learning with Artificial Neural Networks (ANN):

- Sequential model with dense and dropout layers
- Sigmoid activation function
- 64 units in each hidden layer
- Dropout layers with 0.1 probability
- Output calculated using the sigmoid method



4. Experiment

4.1 Data

Table 3. Features description

Gender	0: Female, 1: Male
Age	Age at the time of examination
Education	Education level ranging from 1 (some high school) to 4 (college degree)
CurrentSmoker	0: nonsmoker; 1: smoker
CigsPerDay	Number of cigarettes smoked per day (estimated average)
BPMeds	Whether the patient was on blood pressure medication
PrevalentStroke	Whether the patient had previously experienced a stroke
PrevalentHyp	Whether the patient was hypertensive
Diabetes	Whether the patient had diabetes
TotChol	Total cholesterol level
SysBP	Systolic blood pressure
DiaBP	Diastolic blood pressure
BMI	Body Mass Index
HeartRate	Heart rate
Glucose	Glucose level
TenYearCHD	Occurrence of heart disease in the next 10 years

- Cohort dataset from the Framingham Heart Study (FHS)
- Initial dataset: 4240 rows
- Cleaned dataset: 3658 rows
- Choose 10 features ("male", "age", "currentSmoker", "BMI", "diabetes", "totChol", "sysBP", "diaBP", "heartRate", "glucose") for models.

4. Experiment

4.2 Preprocessing data

- **Imbalance problem:** Huge difference between individuals with and without cardiovascular diseases
- **Solution:** SMOTE (Synthetic Minority Oversampling Technique)

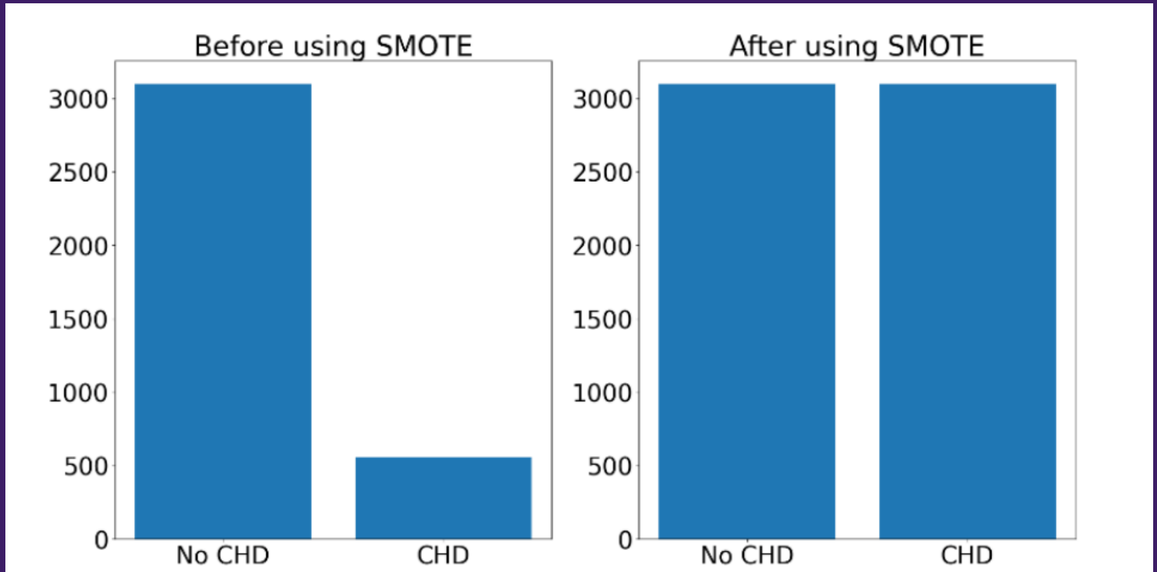
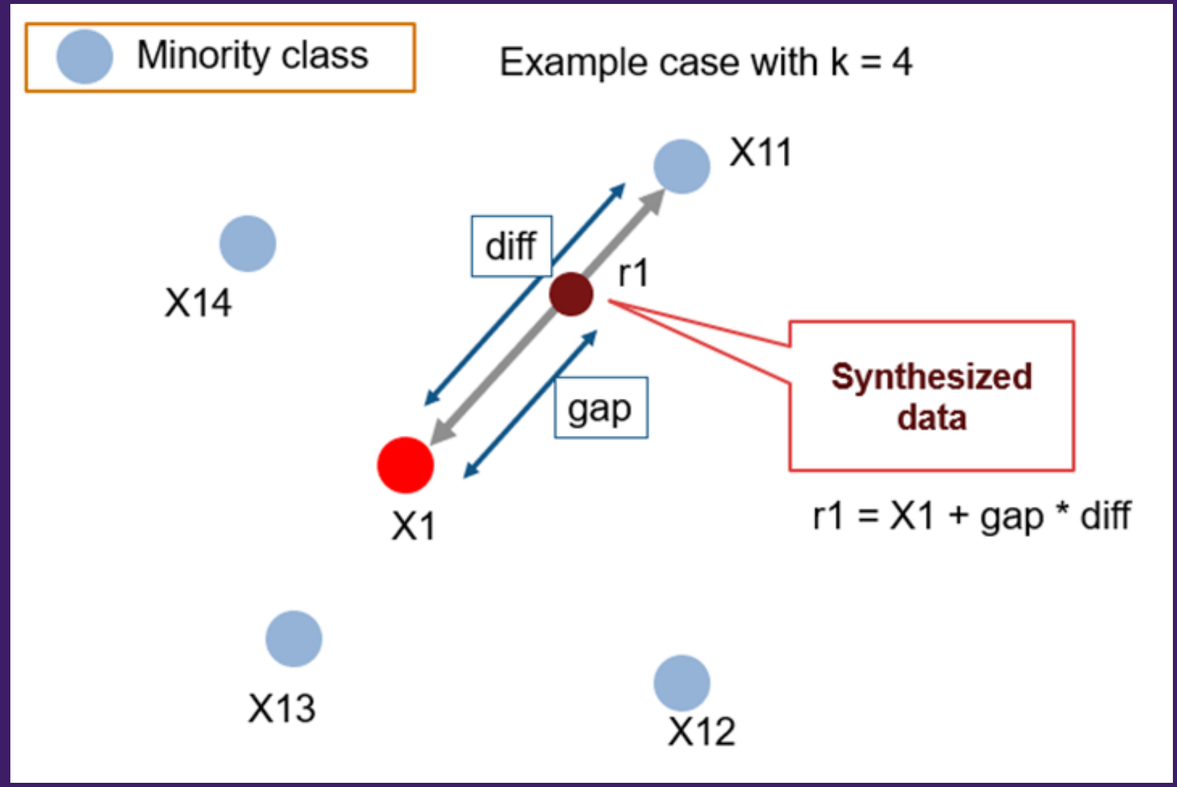


Fig. 3. Before and after using SMOTE.

How SMOTE work



4. Experiment

4.3 Hyperparameters

Logistic Regression	SVM	KNN	Random Forest
C = 0.1	C = 100	n_neighbors = 3	criterion = “gini”
max_iter = 1000	gamma = 0.001		max_depth = 15
penalty = “l1”	kernel = “rbf”		n_estimators = 300
solver = “liblinear”			

Deep Learning			
Nodes	Dropout Probability	Learning rate	Batch size
64	0.1	0.001	64

4. Experiment

4.4 Training process and evaluations

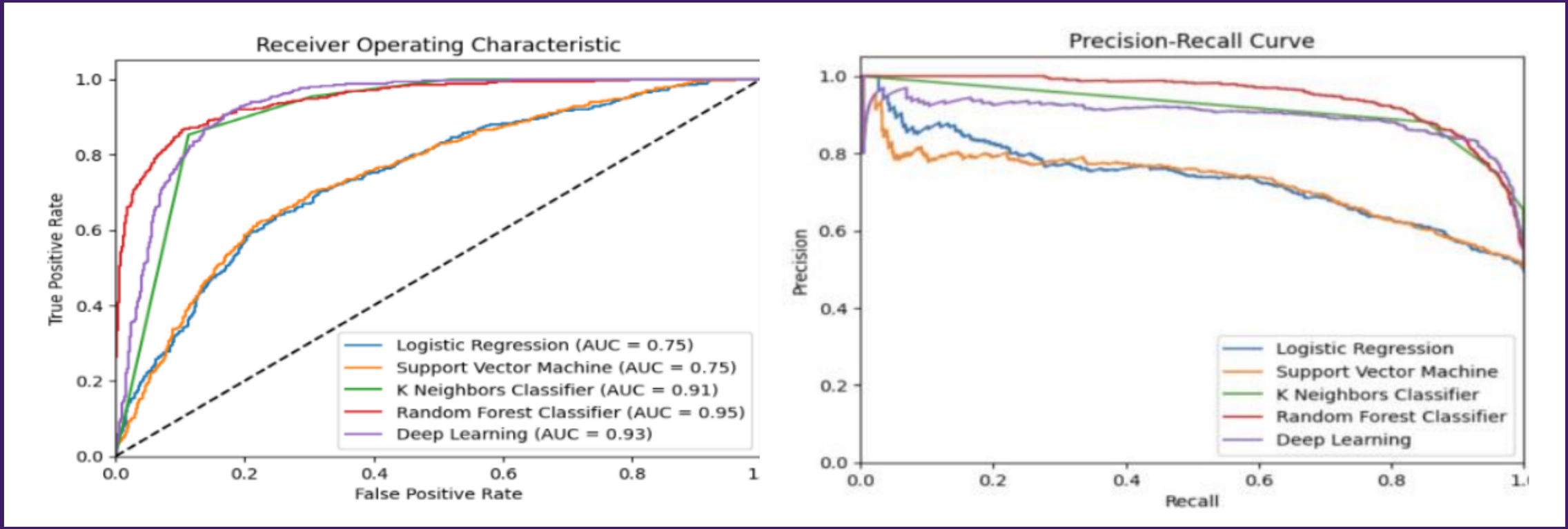
- Dataset split: Training (70%) and Test (30%)
- Performance evaluated using various metrics:
  - Accuracy, Precision, Recall, F1 score

Model Performance

	Logistic Regression		SVM		KNN		Random Forest		Neural Network	
Accuracy	0.69		0.688		0.824		0.871		0.866	
Precision (0|1)	0.7	0.69	0.71	0.67	0.94	0.75	0.88	0.86	0.91	0.83
Recall (0|1)	0.69	0.7	0.65	0.72	0.7	0.95	0.86	0.88	0.81	0.92
F1 score (0|1)	0.69	0.69	0.68	0.7	0.8	0.84	0.87	0.87	0.86	0.87

4. Experiment

4.4 Training process and evaluations

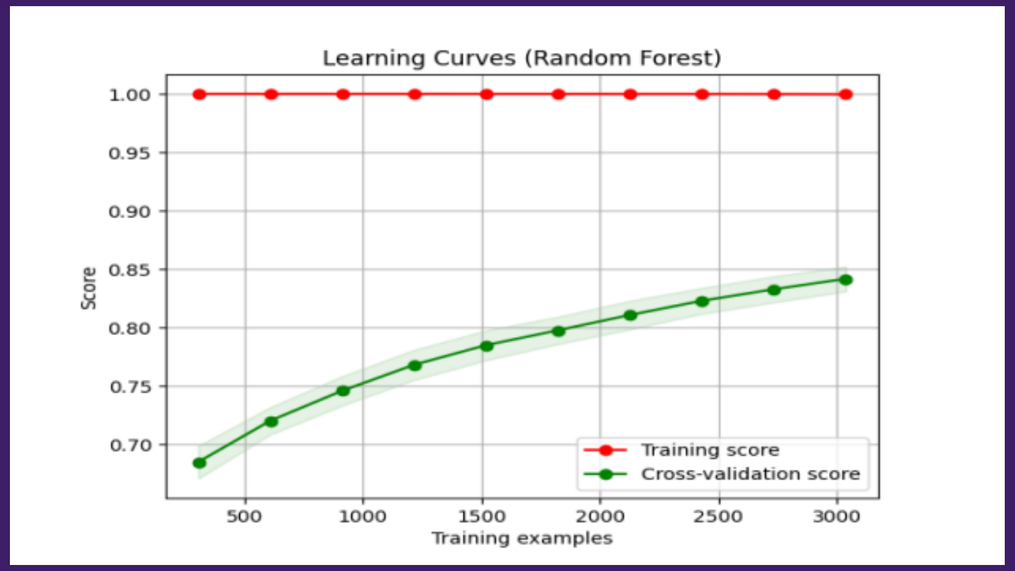


- Random Forest model exhibits highest accuracy
- Precision and recall show similar differences
- Random Forest model has highest AUC metric

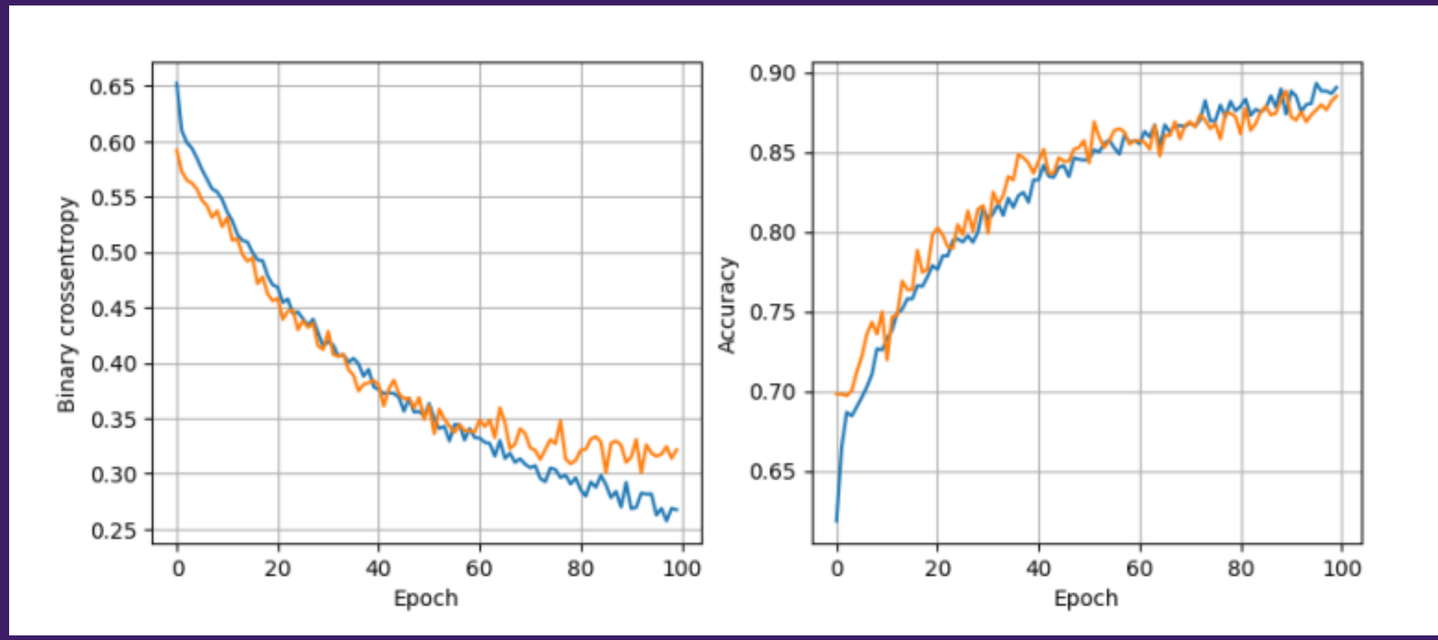
4. Experiment

4.4 Training process and evaluations

- Random Forest model learning curve not ideal

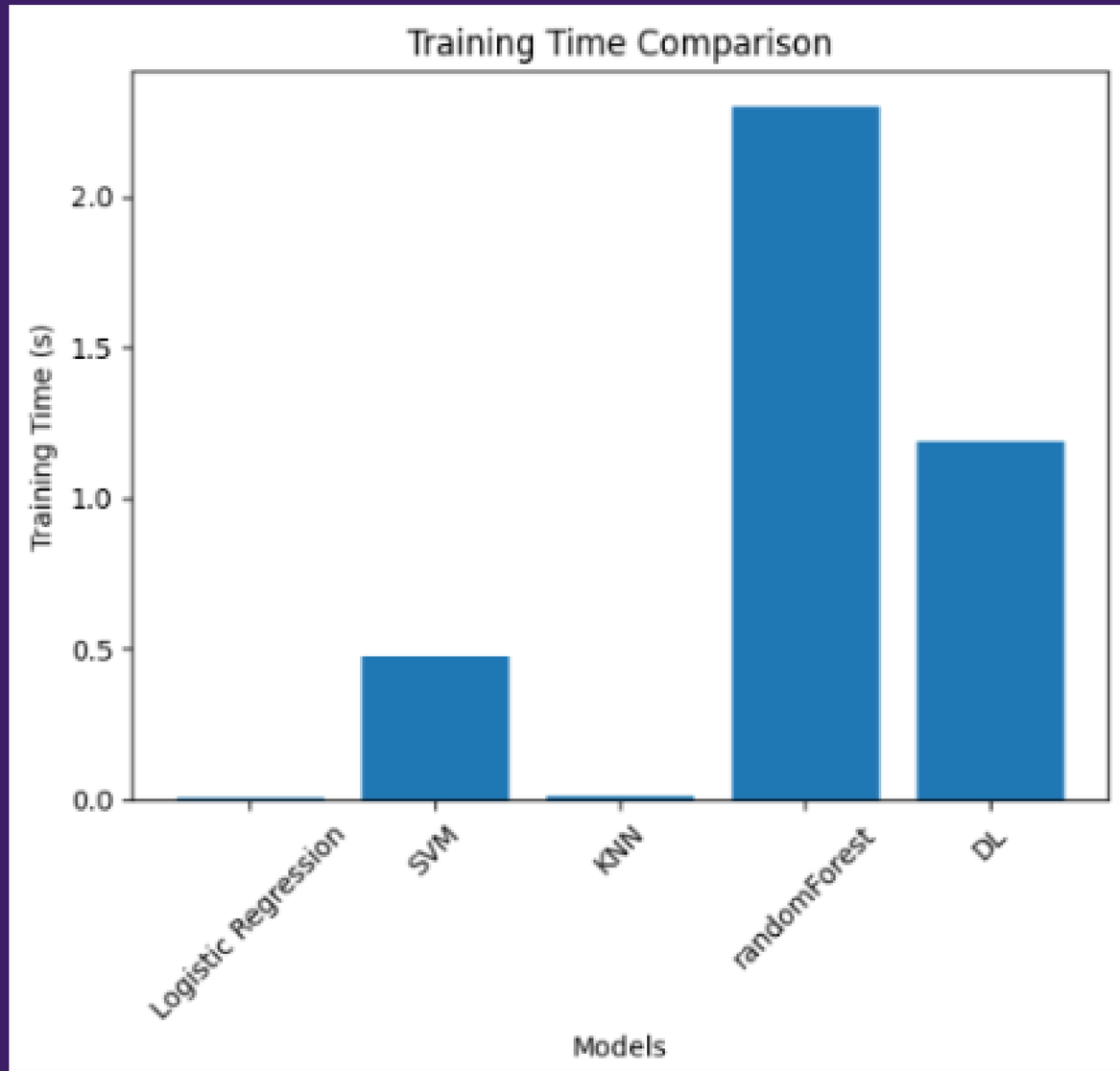


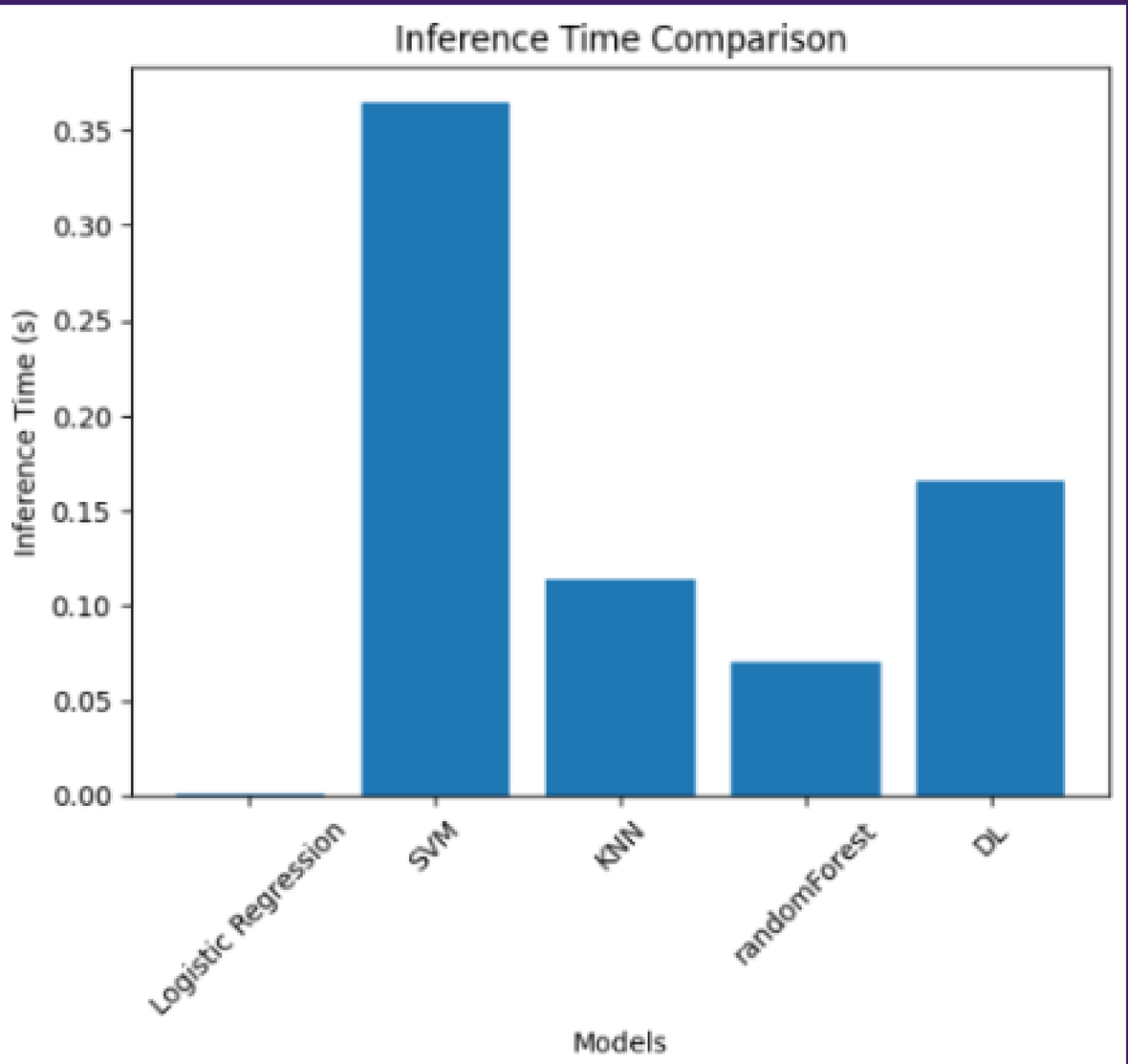
- Second-ranked model (DL) performs well during cross-validation
- Minimal discrepancy between training and validation curves
- Expectation of good performance



(*)The training dataset is represented by the blue line, while the orange line depicts the validation dataset.

5. Discussion





5. Discussion

Model and Parameter:

- Logistic Regression: **C= 0.1**, max_iter= 1000, penalty='L1', sovler= liblinear
- SVM: **C= 100**, Gamma= 0.001, kernel= 'rbf'
- KNN: **n_neighbors= 3**
- RandomForest: **criterion= 'gini'**, max_depth= 15 and n_estimators= 300
- DL: sixty-four neurons count, sigmoid output layer, optimized using Adam with lr= 0.001

	Algorithm	Accuracy
0	Random Forest	0.871037
1	DL	0.866201
2	KNN	0.823751
3	Logistic Regression	0.690489
4	SVM	0.688340

5. Discussion

In conclusion:

Two top-performing models:

- Neural network model and random forest achieve the highest accuracy (Approximately 87%).
- They also demonstrate the AUC-ROC score around 0.93 and 0.94.

5. Discussion

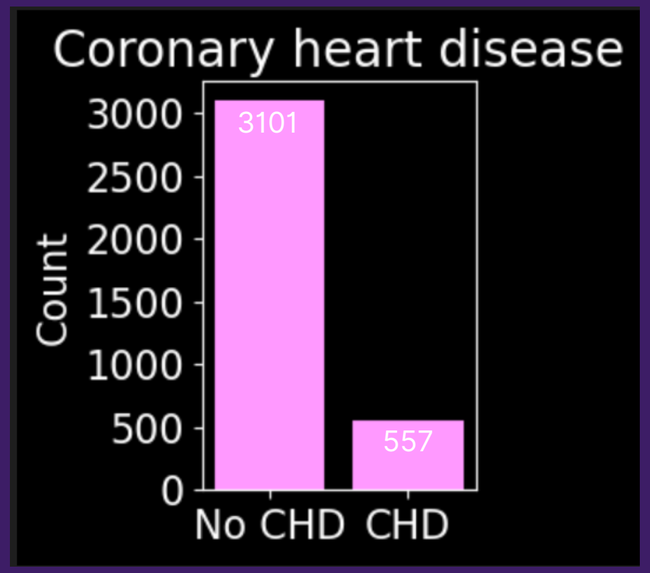
KNN, LR, SVM:

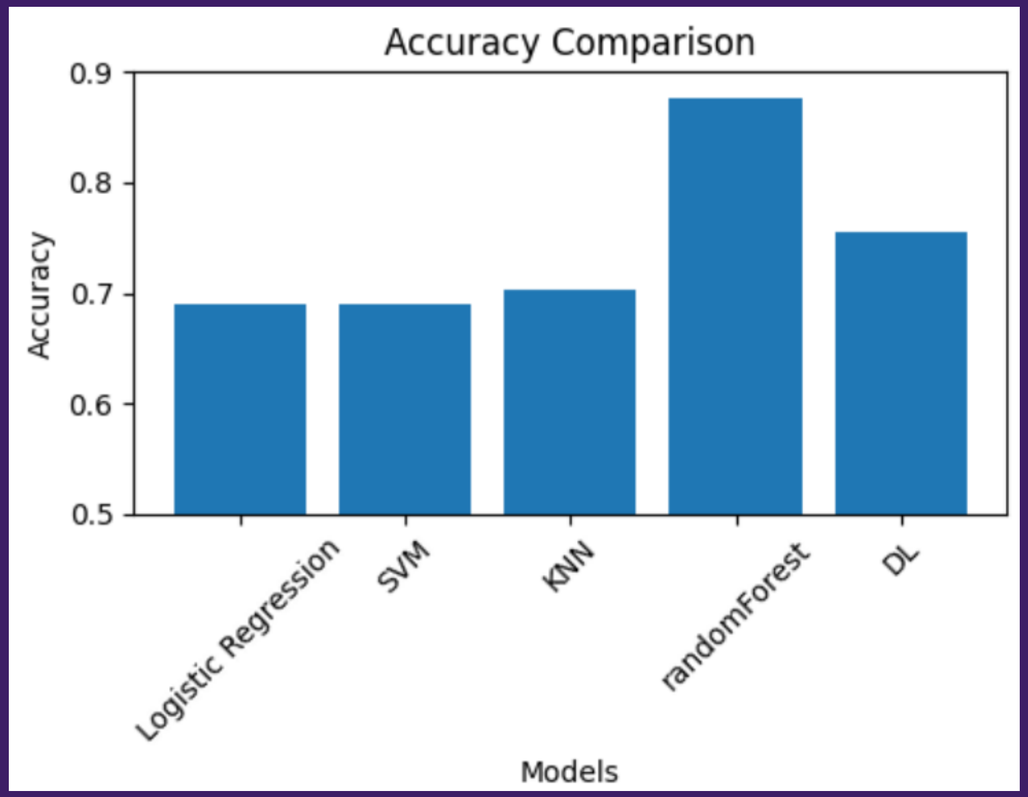
- **KNN**, **Logistic Regression** and **SVM** displayed respectable performance.
- Suitable for different situations.
- Contribute to early detection and preventive strategies.
- It is important to careful consideration of dataset characteristics, interpretability requirements, and performance objectives
- Further research on KNN, Logistic Regression and SVM:
  - Focus on refining these models
  - Incorporating additional features
  - Validating their performance on diverse, larger datasets.

6. Conclusion

6.1 Summarize research objectives

6.2 Main results





=> Random Forest Classifier is the model with the highest and best completeness. However, this model is a bit overfitting and tends to predict people without heart disease better because of the disparity between CHD and NoCHD

6. Conclusion

6.3 Meaning and importance of the results

- Potential to improve the efficiency of heart disease prediction work
- Significant impact on the medical field returns quickly and accurately using a variety of models.

6.4 Proposing directions for further research

- Identify the key mechanisms by which demographic and medical data impact various specific cardiovascular diseases (...)
- Be able to predict the incidence of a particular disease

BY GROUP 1

Thank You!

FOR YOUR SINCERE LISTENING

ANY QUESTIONS ?