Multiple Diseases Prediction

Lecturer:

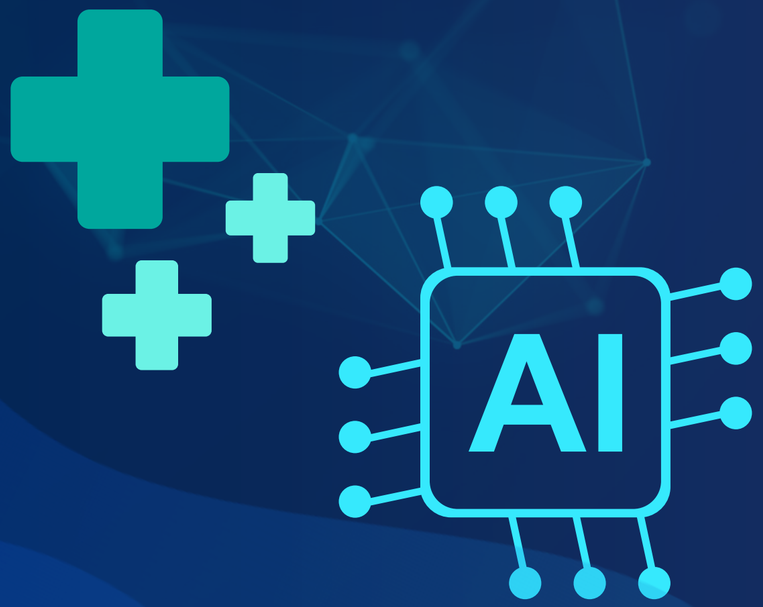
Ms. Te Sonita
Ms. Lay Puthineath
Ms. Monirath Sokchovy
Mr. Pov Phannet

Members:

Prak Pichey
Heng Nenghak
Chhim Kakada
Vuth Watnakpiseth

Agenda

1. Introduction
2. Problems
3. Machine Learning Pipeline
4. Conclusion



Introduction

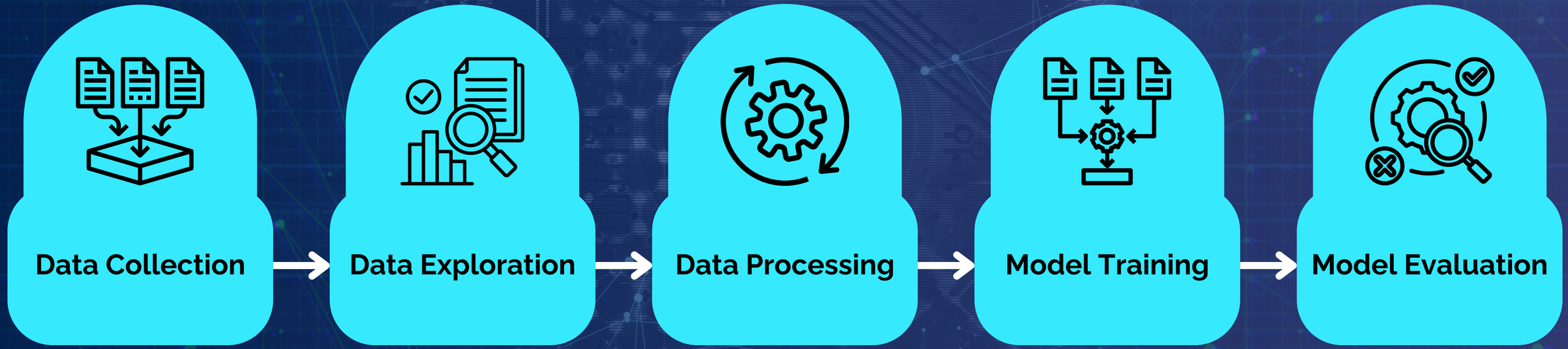
This project focuses on developing a machine learning model to predict common diseases in Cambodia, such as **stroke, hypertension, heart disease, and diabetes**. By analyzing health and lifestyle factors, the model aims to support early detection and intervention, ultimately improving public health outcomes and reducing the risk of severe health complications.

Problems

LIMITED ACCESS TO PREDICTIVE HEALTHCARE TOOLS

LACK OF EARLY DISEASE DETECTION

Machine Learning Pipeline

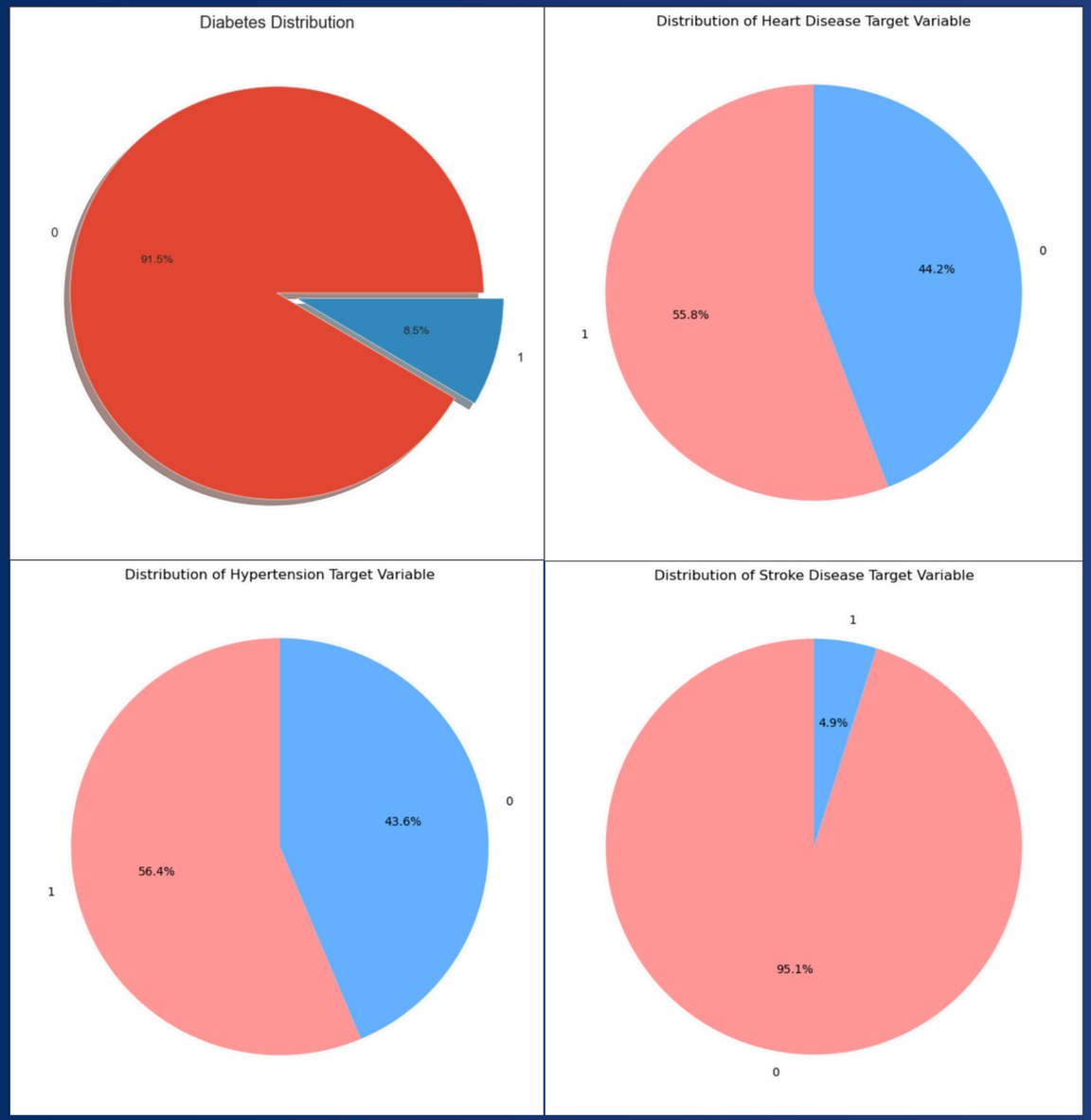


Data Collection

All the datasets were collected from a popular website called Kaggle.

- **Diabetes disease**: 10000 rows, 7 features
- **Heart disease**: 1027 rows, 14 features
- **Stroke disease**: 5110 rows, 11 features
- **Hypertension disease**: 24422 rows, 14 features

Data Exploration



All datasets collected are mostly, balanced except Diabetes and Stroke datasets.

- To handle this we apply SMOTE to balance the data of target variable.

Data Exploration

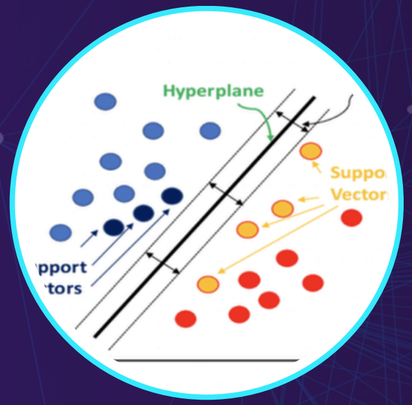
06

Datasets	Relevant Features	Irrelevant Features
Diabetes	sex, age, hypertension, heart disease, smoking history, bmi, HbA1c level, blood glucose level	None
Heart Diesease	cp, thalach, ca, oldpeak, thal, exang	age, sex, trestbps, chol, fbs, rectecg, slope
Hypertension	cp, thalach, oldpeak, exang, slope, thal, target	age, sex, chol, fbs
Stroke	age, hypertension, heart disease, bmi, avg_glucose_level, gender, smoking status	work type, ever married, residence type

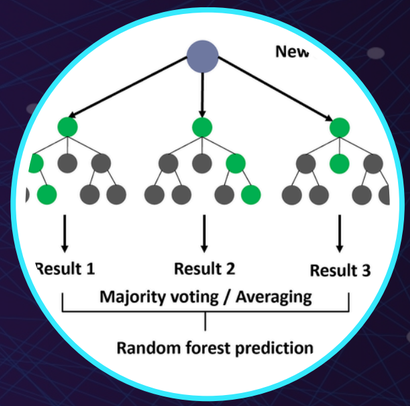
Data Processing

- Check and handle missing and duplicated values
- Check and remove outliers using IQR method
- Clean and pre-processing data across all 4 datasets (Stroke, Diabetes, Hypertension, Heart Disease)
- Before training, features across all the datasets to merged
- Implement 4 different methods to train:
  - Using Raw Data
  - Undersampling Data
  - Oversampling Data
  - Implement K-Cross Validation

Models Selection

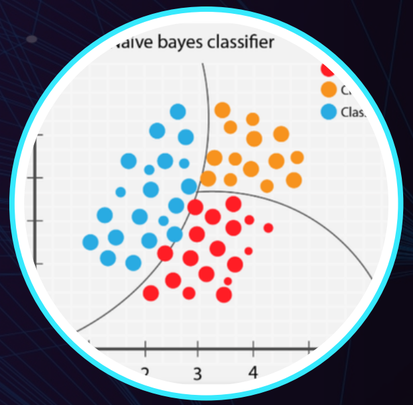


Support Vector Machine



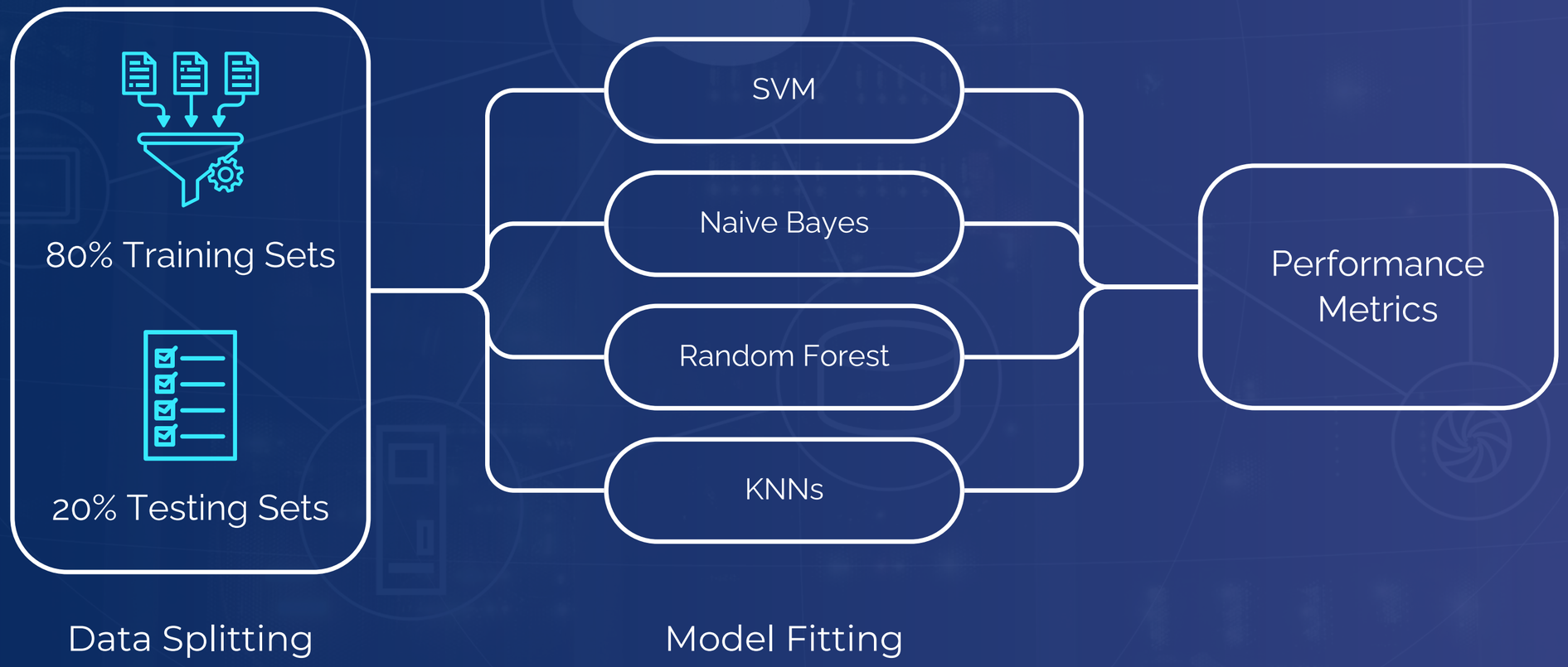
Random Forest

K-Nearest Neighbors



Naive Bayes

Training


80% Training Sets
20% Testing Sets
SVM
Naive Bayes
Random Forest
KNNs
Performance Metrics
Data Splitting
Model Fitting

Evaluation (Do nothing)

Target	Metrics			
	Accuracy	Precision	Recall	F1-Score
Model:	K-Nearest Neighbors			
has_stroke	0.98	0.05	0.19	0.08
has_hypertension	0.85	0.57	0.78	0.66
has_heart_disease	0.87	0.13	0.52	0.20
has_diabetes	0.91	0.30	0.74	0.43
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.86	0.78	0.36	0.50
has_heart_disease	0.97	0.00	0.00	0.00
has_diabetes	0.96	0.00	0.00	0.00
Model:	Naive Bayes			
has_stroke	0.82	0.01	0.99	0.03
has_hypertension	0.82	0.50	0.70	0.58
has_heart_disease	0.60	0.07	0.92	0.13
has_diabetes	0.30	0.06	1.00	0.11
Model:	Random Forest			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.93	0.92	0.69	0.79
has_heart_disease	0.96	0.33	0.08	0.13
has_diabetes	0.97	0.83	0.51	0.63

Evaluation (OverSampling)

Target	Metrics			
	Accuracy	Precision	Recall	F1-Score
Model:	K-Nearest Neighbors			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.93	0.92	0.69	0.79
has_heart_disease	0.96	0.31	0.06	0.11
has_diabetes	0.97	0.84	0.46	0.60
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.86	0.78	0.36	0.50
has_heart_disease	0.97	0.00	0.00	0.00
has_diabetes	0.96	0.00	0.00	0.00
Model:	Naive Bayes			
has_stroke	0.82	0.01	0.98	0.03
has_hypertension	0.82	0.52	0.71	0.60
has_heart_disease	0.58	0.06	0.91	0.12
has_diabetes	0.30	0.05	1.00	0.10
Model:	Random Forest			
has_stroke	0.99	0.50	0.06	0.10
has_hypertension	0.93	0.93	0.69	0.79
has_heart_disease	0.96	0.35	0.10	0.16
has_diabetes	0.97	0.81	0.49	0.61

Evaluation (UnderSampling)

Target	Metrics			
	Accuracy	Precision	Recall	F1-Score
Model:	K-Nearest Neighbors			
has_stroke	0.95	0.52	0.23	0.32
has_hypertension	0.89	0.67	0.44	0.53
has_heart_disease	0.99	0.98	0.87	0.92
has_diabetes	1.00	0.72	0.36	0.48
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.95	0.00	0.00	0.00
has_hypertension	0.86	0.00	0.00	0.00
has_heart_disease	0.95	0.00	0.00	0.00
has_diabetes	1.00	0.00	0.00	0.00
Model:	Naive Bayes			
has_stroke	0.20	0.05	0.97	0.10
has_hypertension	0.81	0.29	0.24	0.26
has_heart_disease	0.92	0.33	0.74	0.46
has_diabetes	0.96	0.05	1.00	0.09
Model:	Random Forest			
has_stroke	1.00	1.00	0.94	0.97
has_hypertension	0.99	1.00	0.94	0.97
has_heart_disease	1.00	1.00	0.96	0.98
has_diabetes	1.00	0.90	0.56	0.69

Evaluation (K-Fold Cross Validation)

Target	Metrics (Check if they have the disease so look for 1)			
	Accuracy	Precision	Recall	F1-Score
Model:	**K-Nearest Neighbors**			
has_stroke	0.010	0.010	0.010	0.020
has_hypertension	0.685	0.685	0.685	0.790
has_heart_disease	0.037	0.037	0.037	0.066
has_diabetes	0.314	0.314	0.314	0.438
Model:	**Support Vector Machine (Base Parameters)**			
has_stroke	0.99	0.00	0.00	0.123
has_hypertension	0.86	0.78	0.36	0.022
has_heart_disease	0.97	0.00	0.00	0.50
has_diabetes	0.96	0.00	0.00	0.026
Model:	**Naive Bayes**			
has_stroke	0.823	0.013	0.980	0.026
has_hypertension	0.820	0.501	0.697	0.583
has_heart_disease	0.597	0.066	0.911	0.123
has_diabetes	0.301	0.055	1.000	0.104
Model:	**Random Forest**			
has_stroke	0.997	0.450	0.011	0.022
has_hypertension	0.933	0.923	0.686	0.787
has_heart_disease	0.965	0.318	0.087	0.136
has_diabetes	0.975	0.815	0.504	0.624

Conclusion

Learning Experience

Four models trained

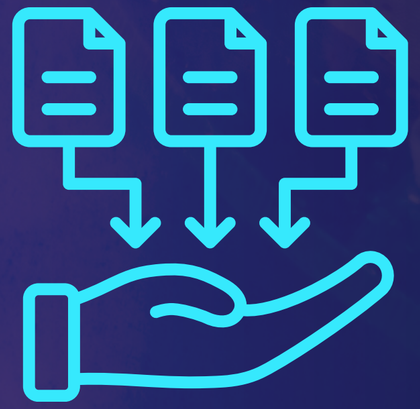
Challenges

New Concepts

Imbalanced Data

Combining Datasets

Future Work



Refined Data Collection

Explore more Techniques

Thank You!