Friedrich-Alexander-Universität
DYNAMICS, CONTROL
MACHINE LEARNING
AND NUMERICS

FAU

CML: Control, Machine Learning, and Numerics
SoSe 2023

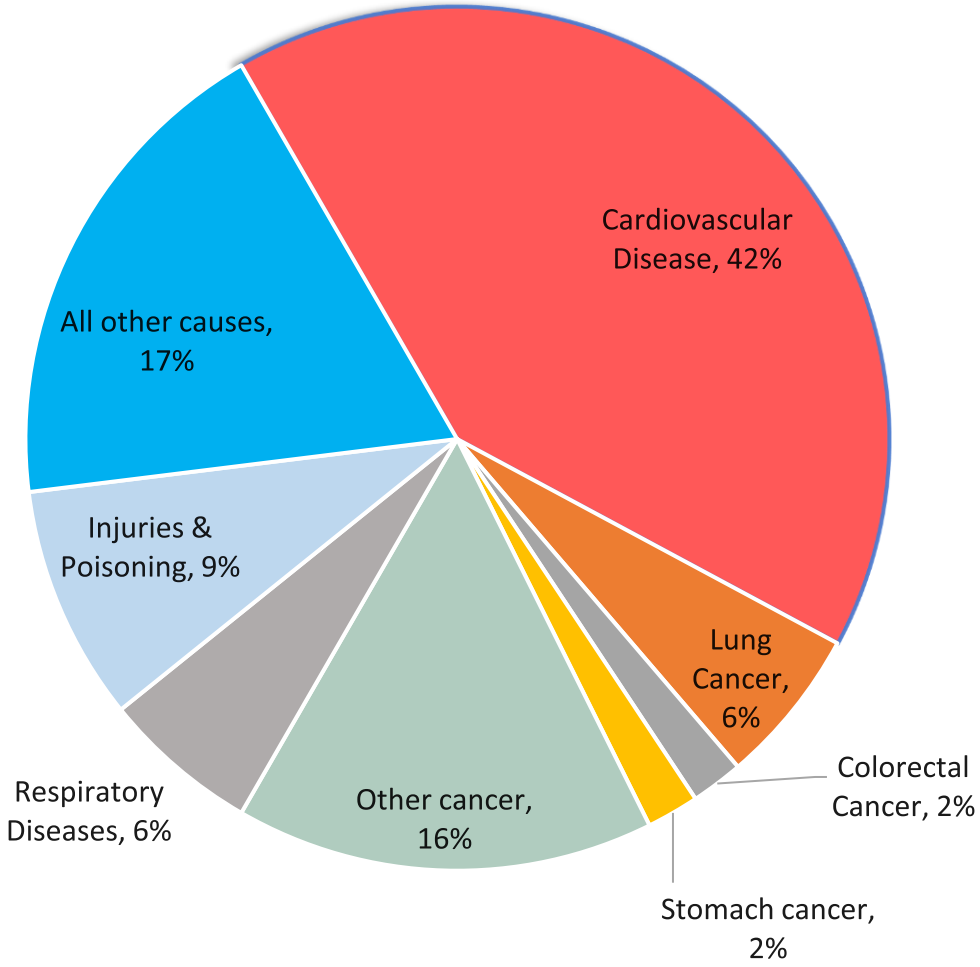
Leveraging Random Forest For Heart Disease Prediction

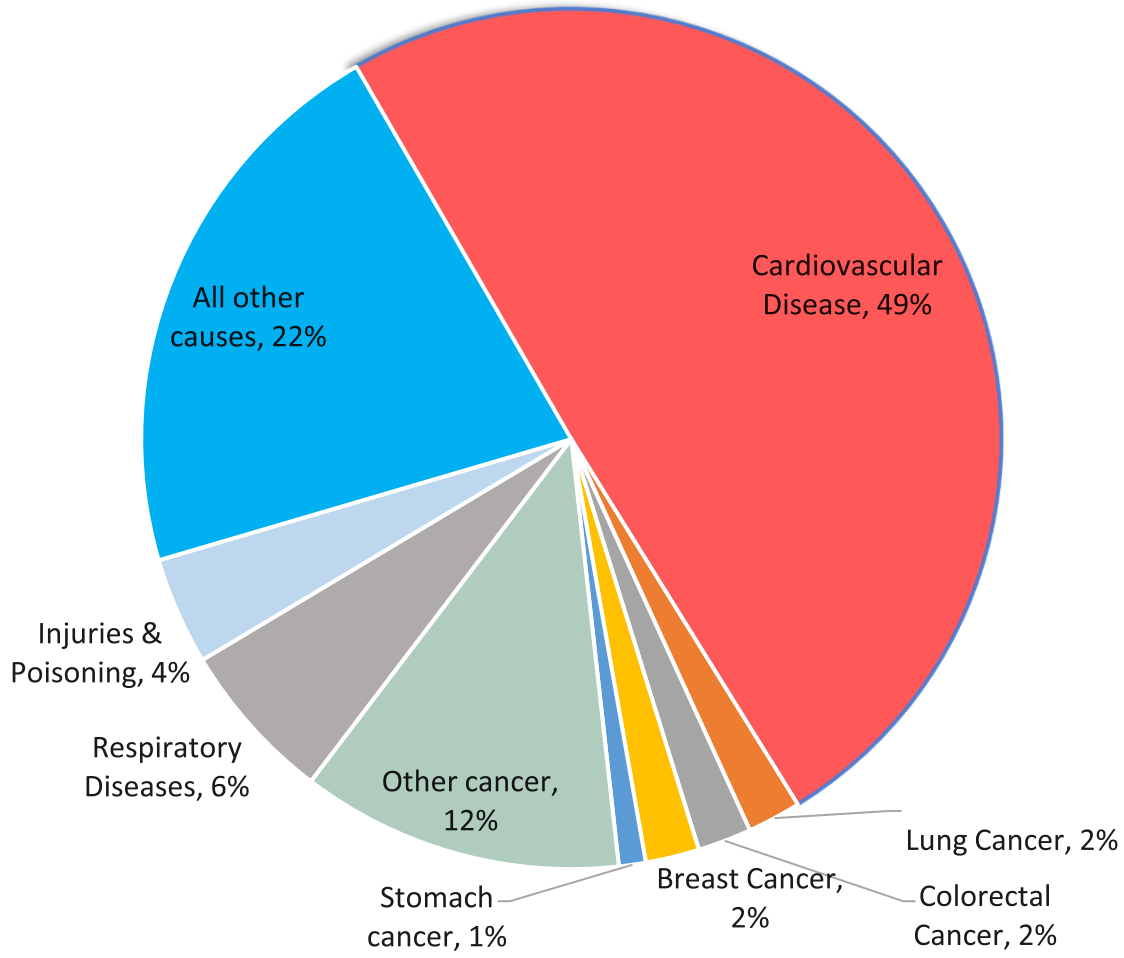
Prof. Dr. DhC. Enrique Zuazua
Dr. Yongcun Song

Group 14:
Alisha Mund
Anuraag Mishra
Harini Sankaran
Malavika Radhakrishnan
Shriya Mittapalli

Content

Introduction – Why Heart Disease?

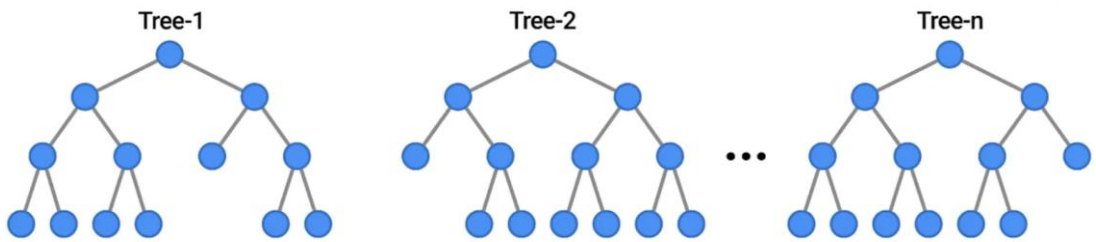




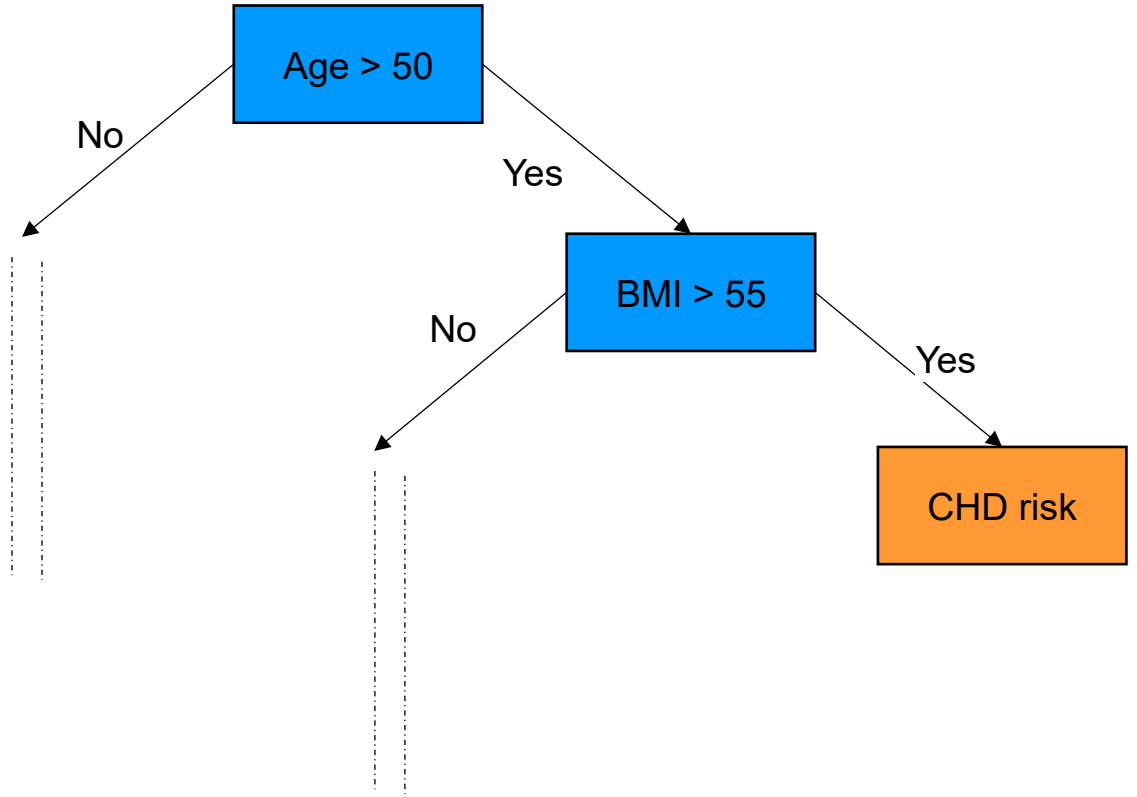
Proportion of all deaths due to major causes in Europe among men (on the left) & women (on the right) (www.who.int).

Introduction – Why Random Forest?

- Robustness to Overfitting
- Interpretability

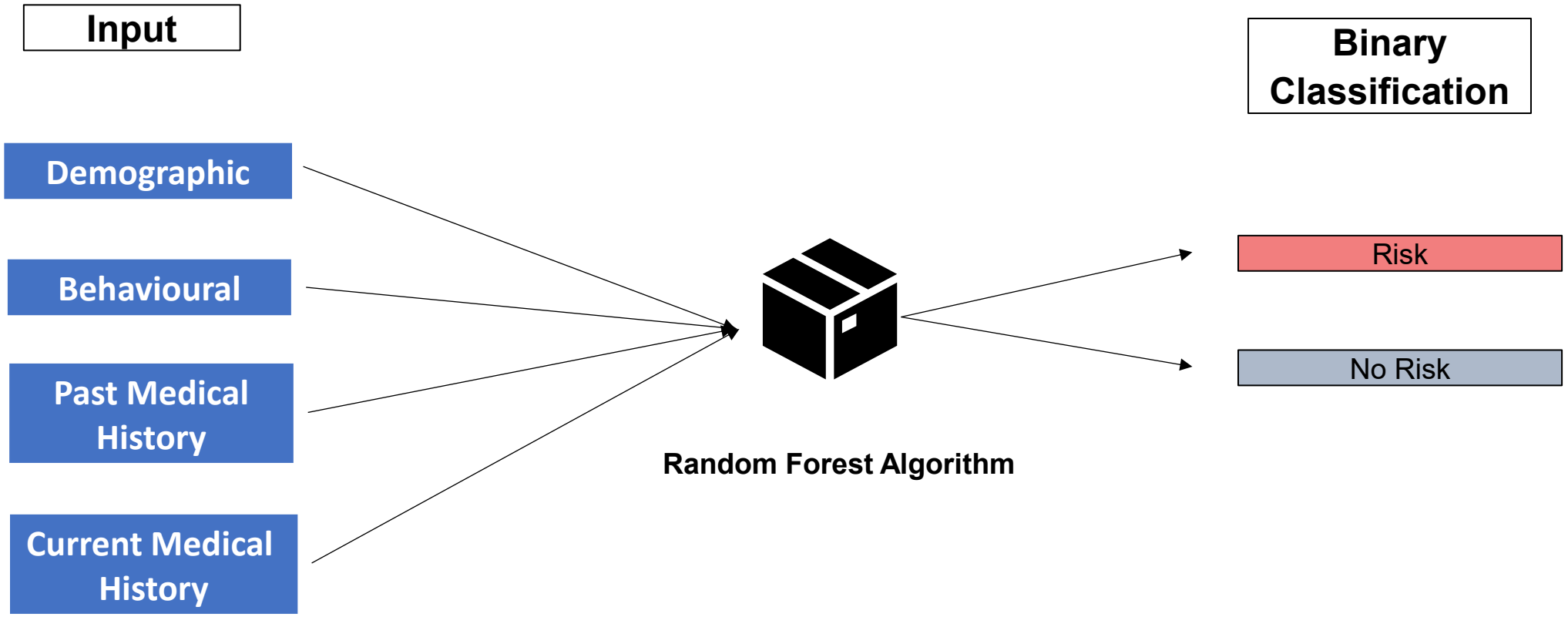


Random Forest with n decision trees

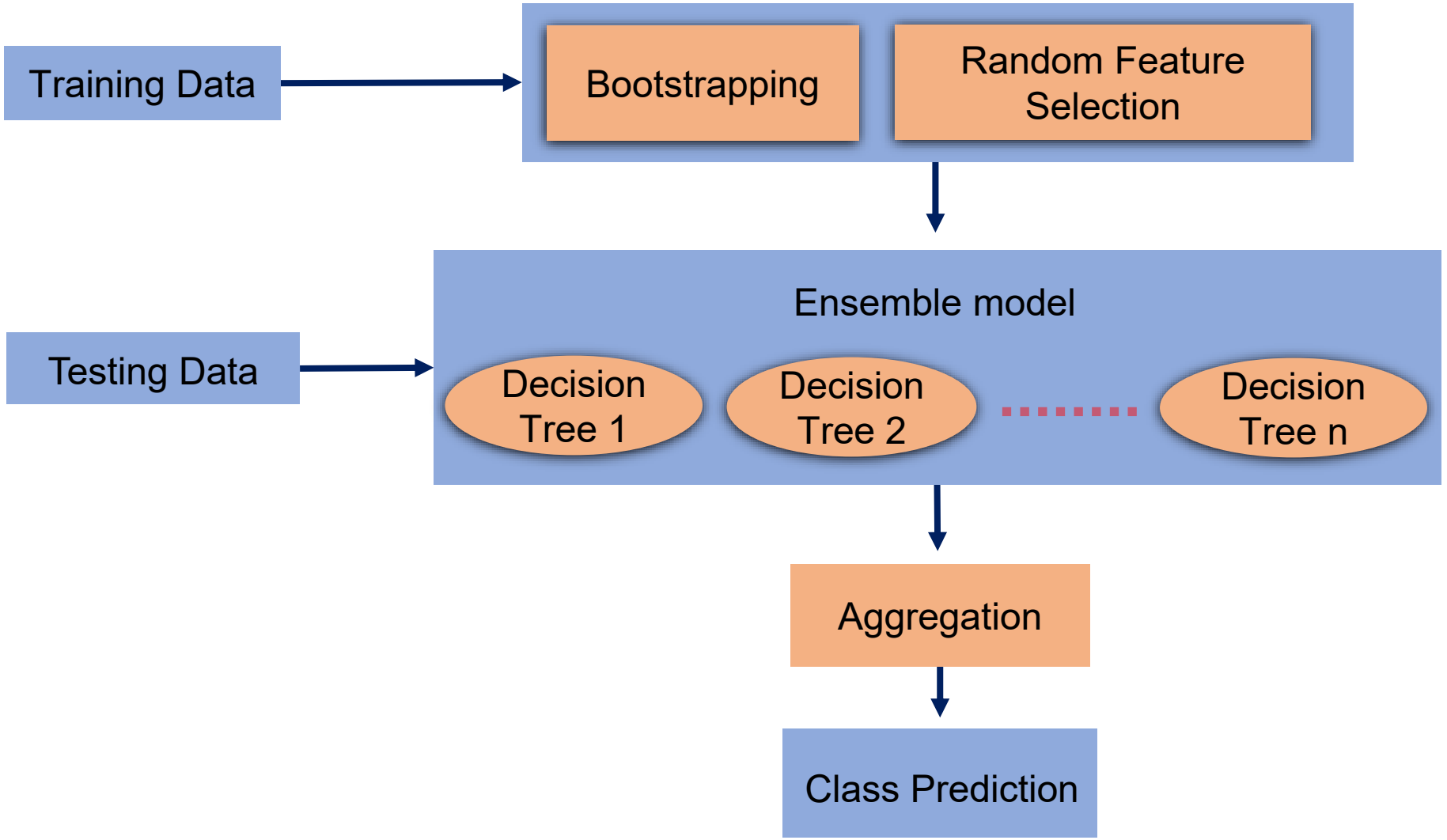


Partial Visualization of a Decision Tree to show Interpretability

Goals and Expectations



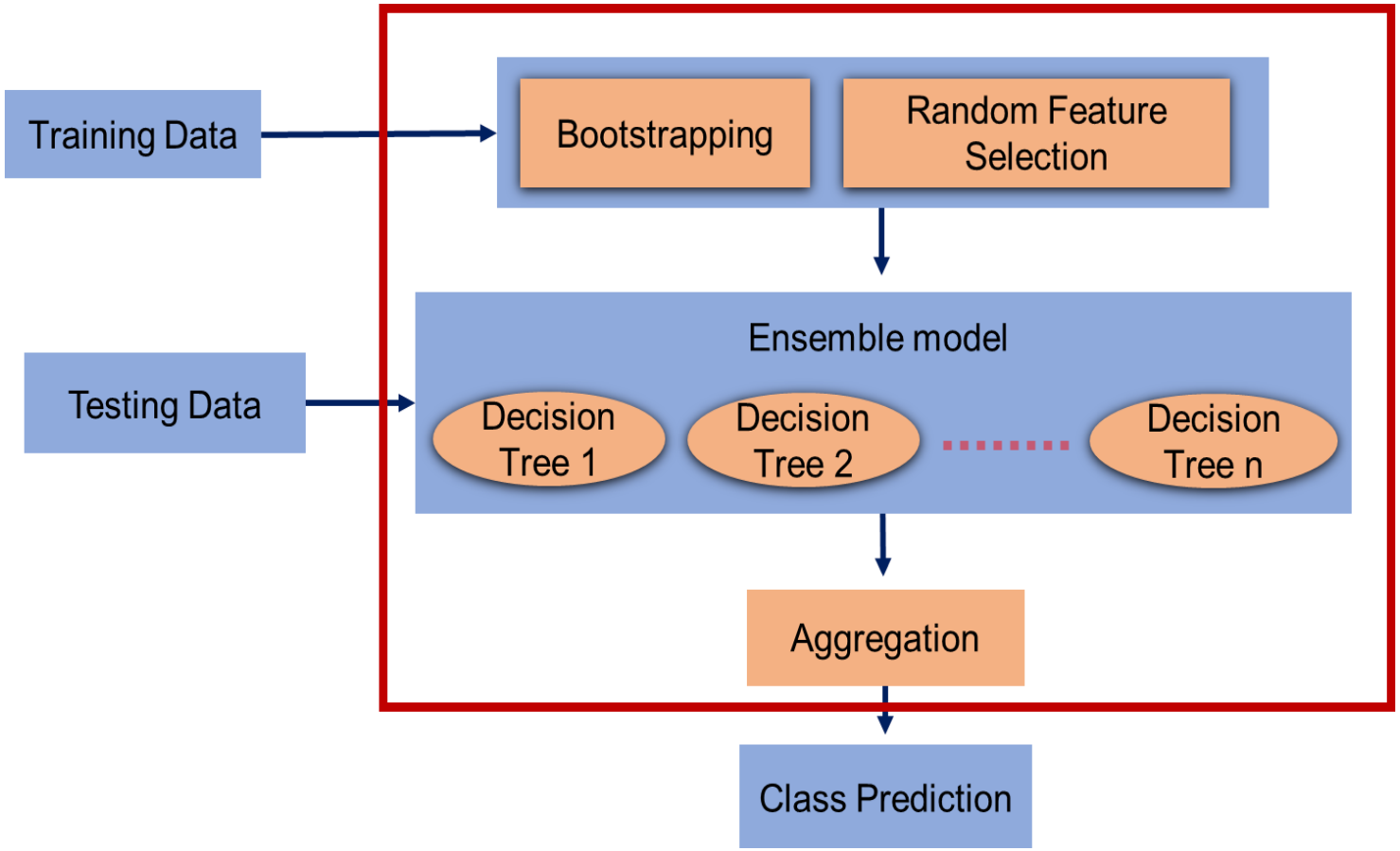
Architecture



Bagging

1. Powerful method based on 2 steps
    - Bootstrapping
    - Aggregation

2. Parallel process



Bagging – Rationale behind Randomization

- For N statistically independent and identically distributed (i.i.d) trees,

variance: $$\sigma^2_{i.i.d\ ensemble} = \frac{1}{N}\sigma^2 \qquad (1)$$

- Bagging introduces only identically distributed variables with correlation ρ,

variance: $$\sigma^2_{i.d\ ensemble} = \rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 \qquad (2)$$

- Randomization decorrelates trees.

Bagging – *Bootstrapping & Random Feature Selection*

Dataset:

ID	Male	Age	Current Smoker	Diabetes	totChol	sysBP	diaBP	HeartRate	Glucose	CHD
0	1	39	0	0	195	106	70	80	77	0
1	0	55	1	0	250	121	81	95	76	0
2	1	48	1	0	245	127.5	80	75	70	0
3	0	61	1	0	225	150	95	65	103	1
4	0	46	1	0	285	130	84	85	85	0

Bootstrapped Samples:

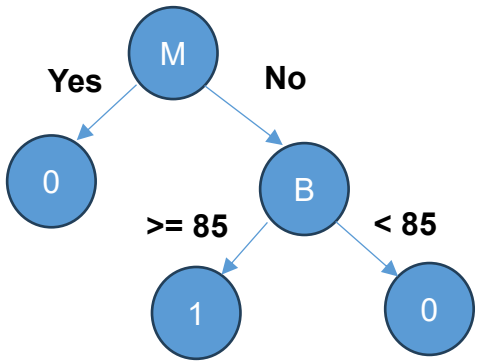
ID: 0, 4, 3, 0, 4 | **ID: 3, 2, 3, 2, 3** | **ID: 2, 3, 2, 0, 3**

Out of Bag Samples:

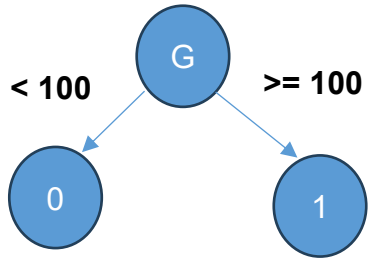
ID: 1, 2 | **ID: 0, 1, 4** | **ID: 1, 4**

Feature selection:

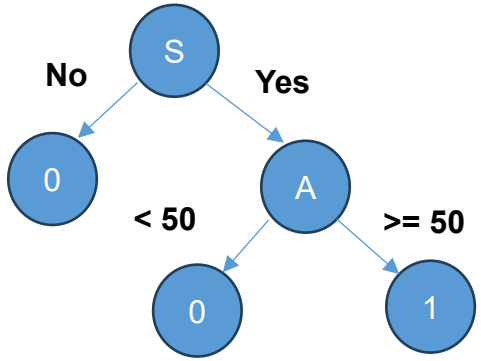
Male(M) , diaBP(B)



Glucose(G), totChol(T)



CurrentSmoker(S), age(A)



Bagging – *Aggregation*

Ensemble Model : Collection of trained trees, designed to capture different aspects of the data for a better prediction.

- *Averaging* operation is done to produce distributions which are heavily influenced by the most confident trees, which are eventually selected for the ensemble.

$$\boldsymbol{P(c|v) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{P_n}(c|v)}$$

where $\mathrm{P_n}(c|v)$ denotes the posterior distribution obtained by the n-th tree.

Aggregation : *Majority Voting*

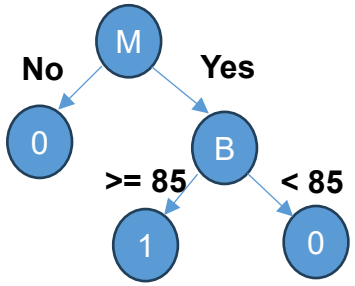
- Each decision tree independently predicts the target class
- Most voted class becomes the final prediction.

Bagging – *Aggregation*

Test Data

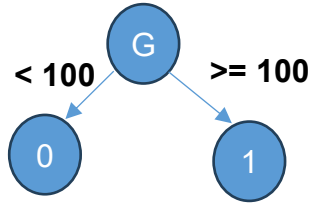
ID	M	A	S	Diabetes	T	sysBP	B	HeartRate	G	CHD
1	0	55	1	0	250	121	81	95	76	0

Classifier 1



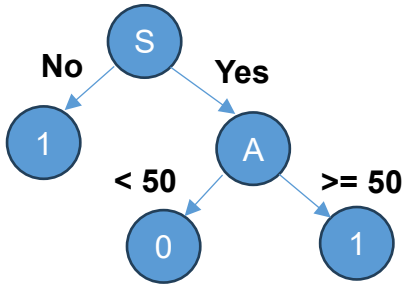
Prediction: 0
No Risk

Classifier 2



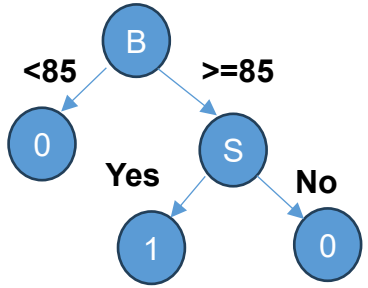
Prediction: 0
No Risk

Classifier 3



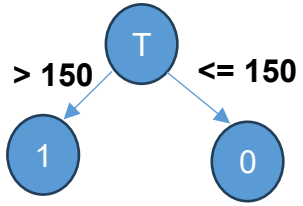
Prediction: 1
Risk

Classifier 4



Prediction: 0
No Risk

Classifier 5



Prediction: 1
Risk

Final Prediction : No Risk

Bagging – *Out of Bag Samples*

If there are d samples in a data set,

- probability of selecting each sample = $\frac{1}{d}$
- probability of sample not being chosen = $(1 - \frac{1}{d})$
- probability that a sample will not be chosen during whole d time = $(1 - \frac{1}{d})^d$

When d is large, we have the probabilities as below,

$$\lim_{d \to \infty} \left(1 - \frac{1}{d}\right)^d = \mathrm{e}^{-1} = 0.368$$

- **36.8%** of samples are not selected and form the validation set
- **63.2%** will form the training set

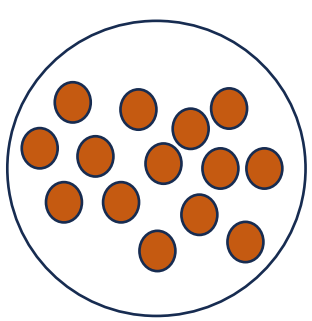
Splitting Criteria – *Entropy*

Entropy: $Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$

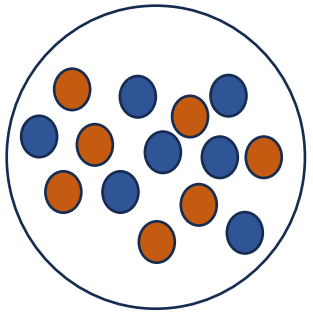
$$p_i = \frac{|C_i|}{|D|}$$

D = samples in the splitting node
m = number of classes
C_i *= Class from i = 1 to m*
p_i *= proportion of samples in D belonging to* C_i

- Measures the disorder of the feature with the target
- Ranges from 0 to 1 (usually), 0 being the optimum



Info(D) = 0



Info(D) = 1

Splitting Criteria – *Information gain*

Information Gain: $Gain(A) = Info(D) - Info_A(\mathrm{D})$

$$Info_A(\mathrm{D}) = \sum_{j=1}^{v} \frac{|D_j|}{|\mathrm{D}|} Info(D_j)$$

A = Splitting Attribute
j = distinct values of A
D_j = j^{th} *child node*

It is the difference between apriori Shannon entropy $Info(D)$ and the conditional entropy $Info_A(\mathrm{D})$

- Helps to determine the order of attributes in the nodes
- Choose attribute with highest information gain

Splitting Criteria – Gain Ratio

$$\textbf{Gain Ratio} = \frac{Gain(A)}{SplitInfo_A(\mathrm{D})}$$

$$SplitInfo_A(\mathrm{D}) = -\sum_{j=1}^{v} \frac{|D_j|}{|\mathrm{D}|} \log_2\left(\frac{|D_j|}{|\mathrm{D}|}\right)$$

D = samples in the splitting node
A = Splitting Attribute
j = distinct values of A
D_j *= j^{th} child node*

- Normalization to information gain using a “split information”
- Removes bias toward tests with many outcomes
- Maximum gain ratio is selected as the splitting attribute

Splitting Criteria – *Gini Index*

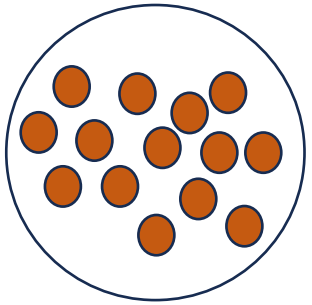
Gini Index: $Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$

m = number of classes
p_i *= proportion of samples in D belonging to* d_i
A = Splitting Attribute
D_j = j^{th} *child node*

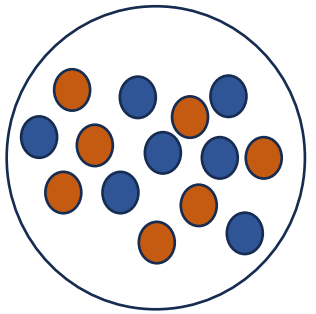
$$Gini_A(\text{D}) = \frac{|D_1|}{|\text{D}|} Gini(D_1) + \frac{|D_2|}{|\text{D}|} Gini(D_2)$$

$$\Delta\text{Gini}(\text{A}) = \text{Gini}(\text{D}) - Gini_A(\text{D})$$

- *Considers a binary split for each attribute*
- *Measures* the frequency at which any element of the dataset will be mislabelled when it is randomly labelled
- Ranges from 0 to 0.5, <u>0 being the optimum value</u>



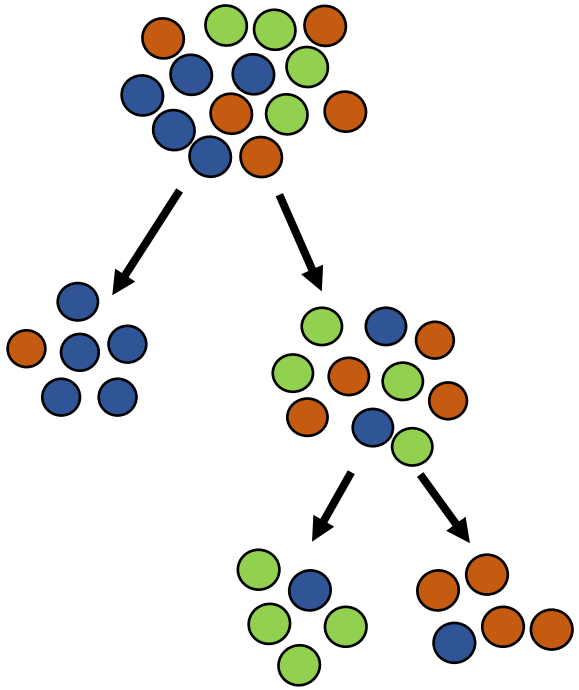
Gini(D)= 0



Gini(D)= 0.5

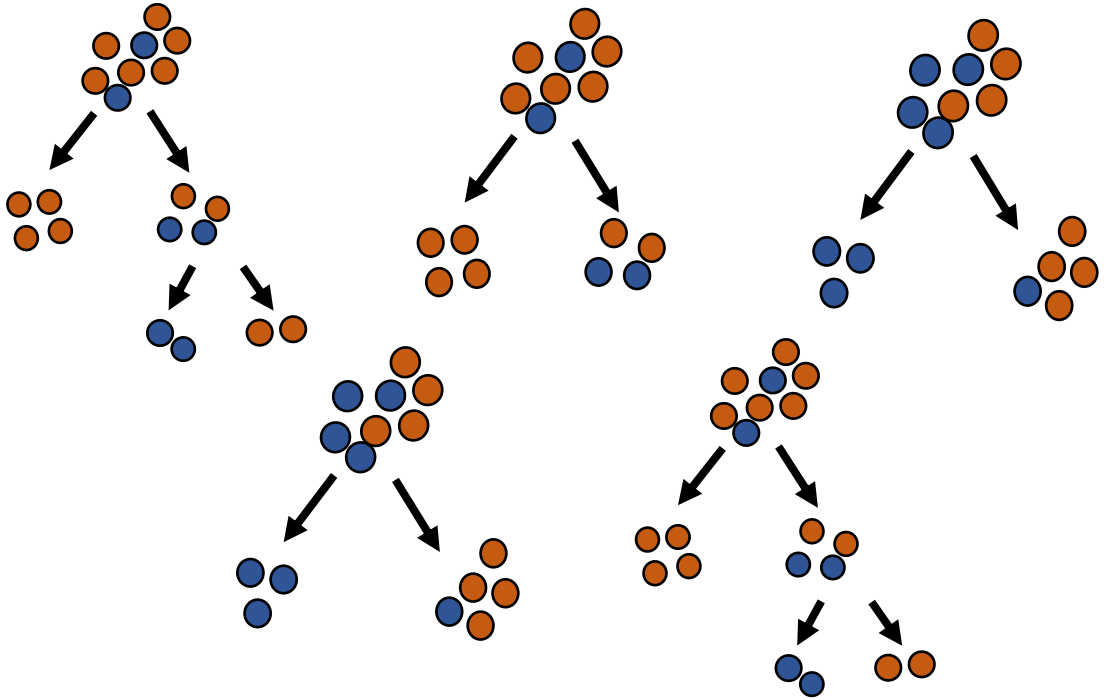
Hyperparameter Tuning

1. ***Max depth -*** longest path between the root node and the leaf node



Depth = 2

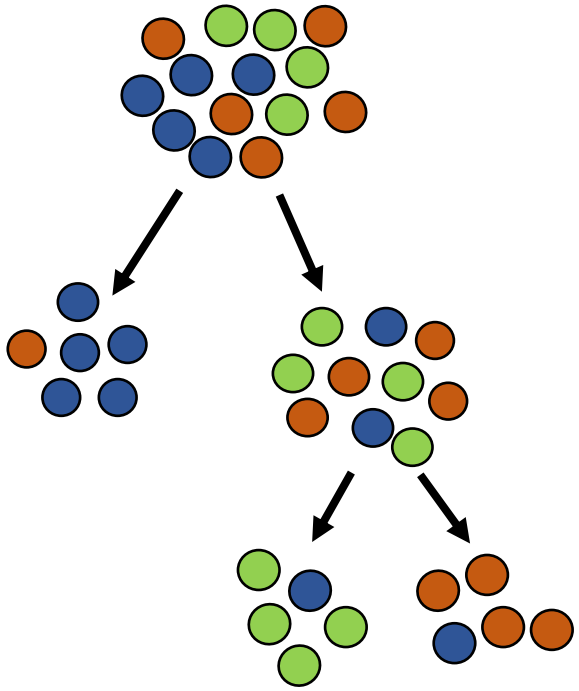
2. ***N estimators -*** number of decision trees



N estimators = 5

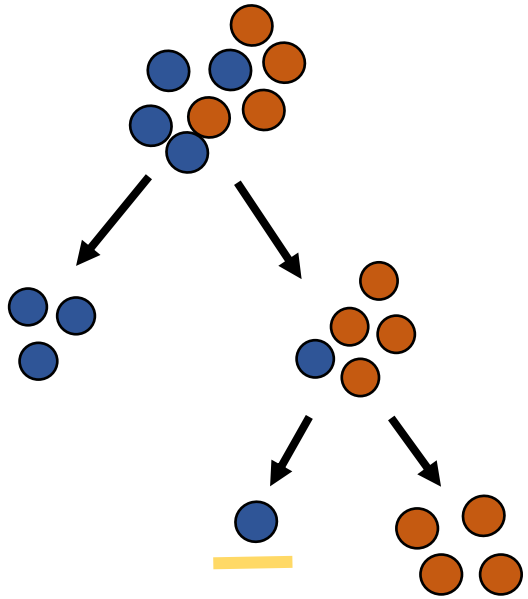
Hyperparameter Tuning

3. ***Max leaf nodes*** **-** condition on the splitting of the nodes restricts the growth of the tree.



Max leaf nodes = 3

4. ***Min sample leaf*** **-** minimum number of samples that should be present in the leaf node after splitting a node



Min sample leaf = 3

Generalization Error

If h1(x), h2(x), ... hk(x) is an ensemble of classifiers with the training set {**X, Y**}, the margin function is:

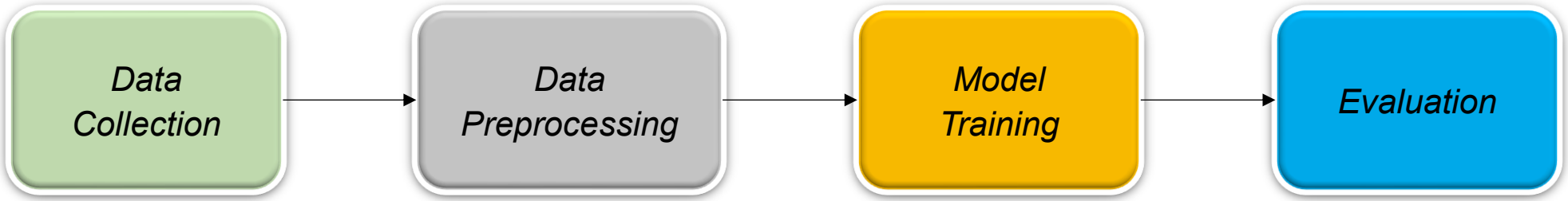
$$mg(x,y) = av_k I(h_k(x) = y) - \max_{j \neq Y} av_k I(h_k(x) = j)$$

and the generalisation error is given by:

$$PE^* = P_{X,Y}(mg(X,Y) < 0)$$

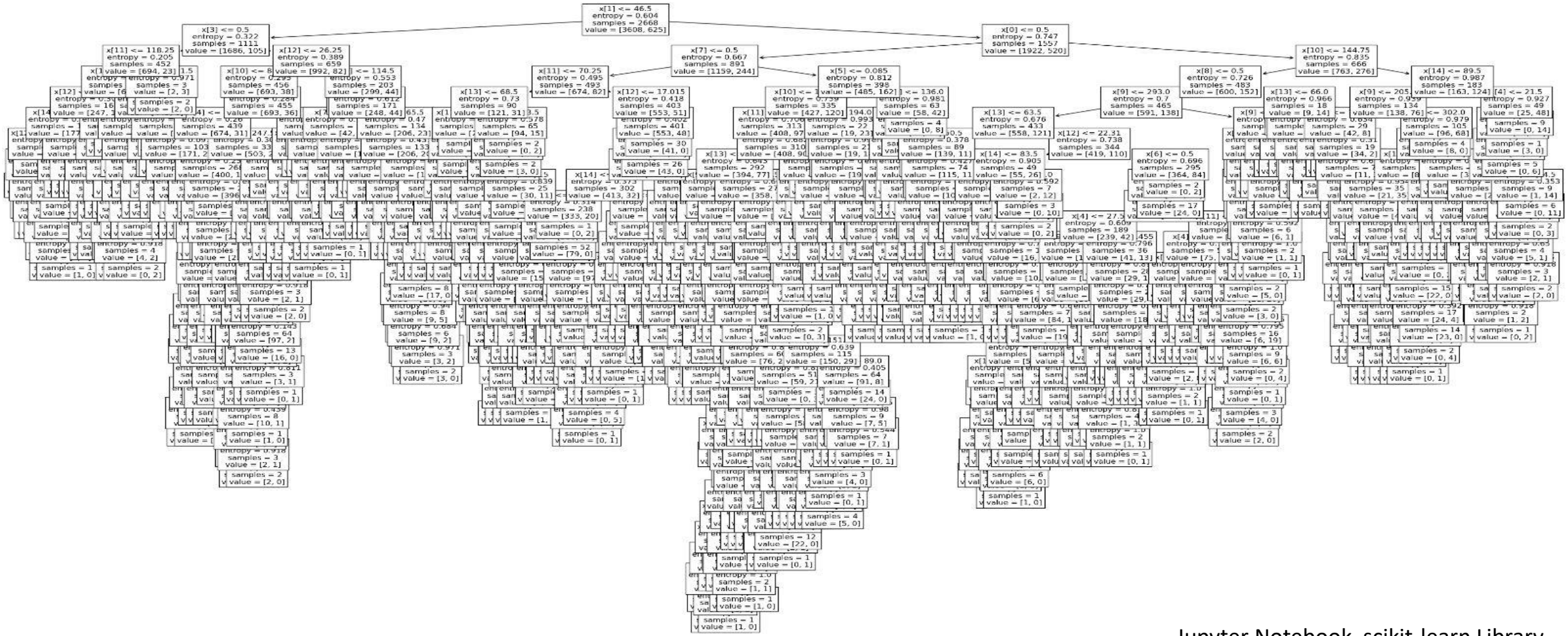
Our goal is to reduce the generalization error and have a high margin as it indicates a higher confidence in correct classification.

Simulation – Model Training



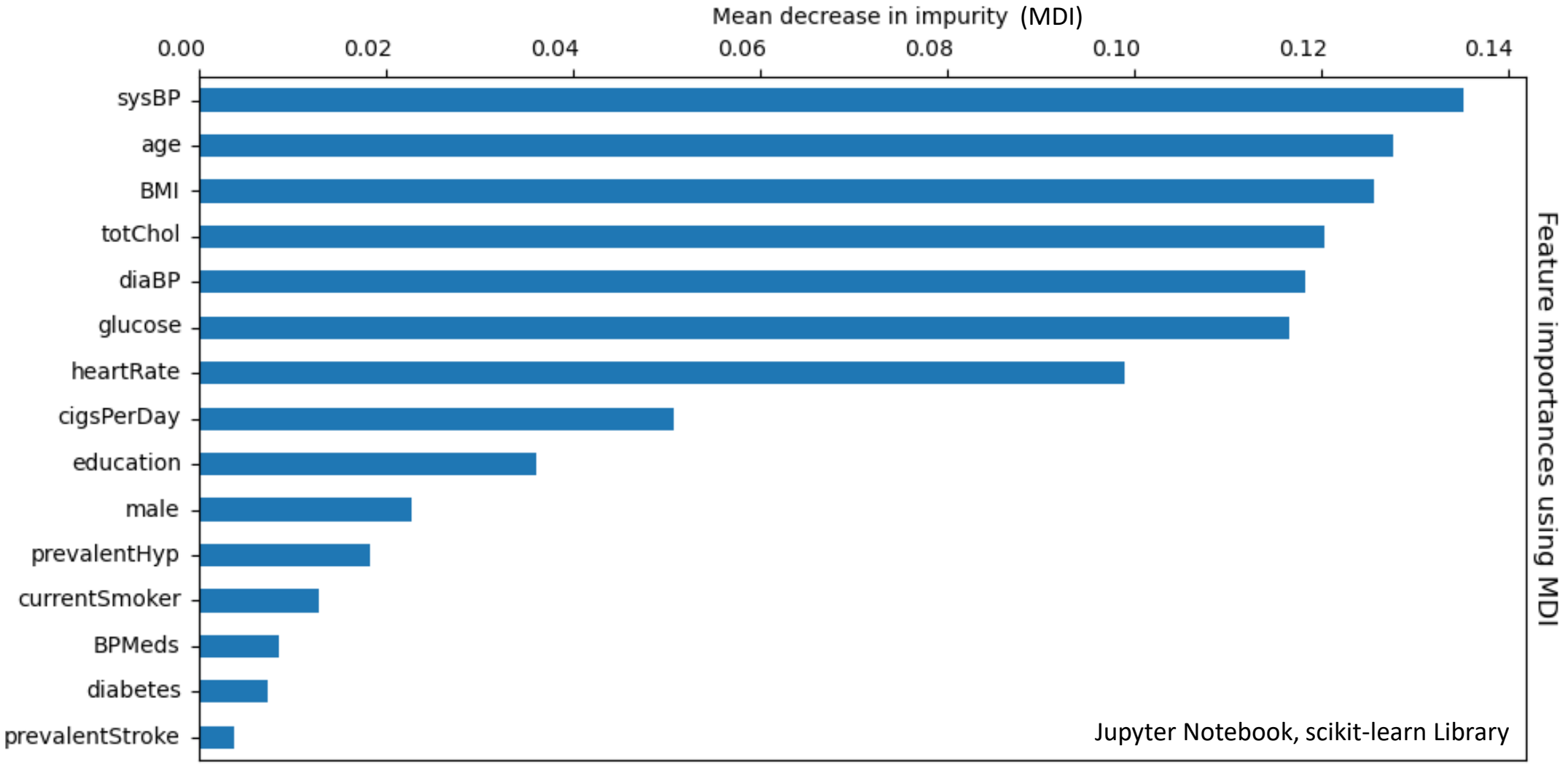
Workflow Visualization for Random-Forest-Classifier model training.

Simulation – Decision Tree



Jupyter Notebook, scikit-learn Library

Simulation – Feature Importance



Simulation – MDI

$$MDI(A) = \frac{1}{n(A)} \sum_{t \in N} DI(t, A)$$

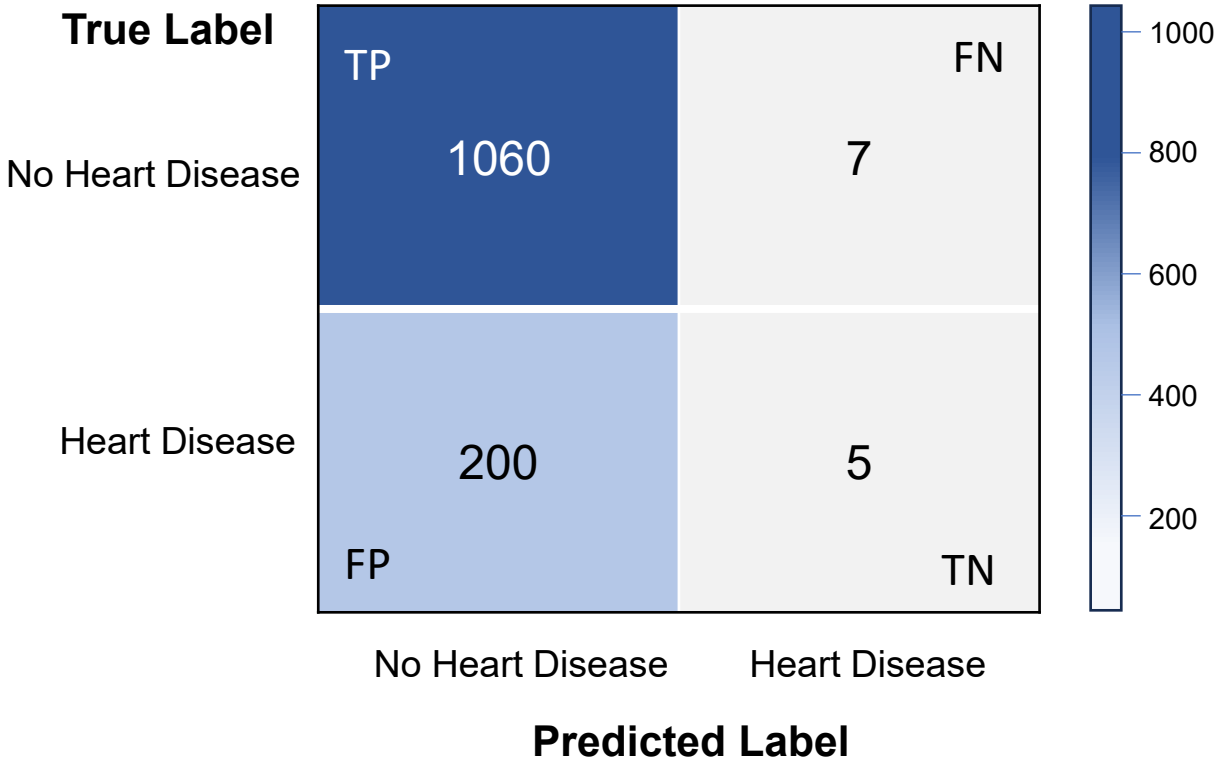
Where,
- MDI(A): Mean Decrease Impurity for feature A.
- t: Individual tree in the set of all trees N.
- DI(t, A): The decrease in impurity in tree t due to feature A.
- n(A): Total number of nodes where A splits.

For each decision tree, Accuracy/Out-of-bag Error is given by,

$$OOB\ Error = \frac{1}{M_0} \sum_{i=1}^{M_0} I(y_i \neq y)$$

Where,
- M_0: the total number of Out of Bag samples
- I: indicator function

Simulation – Confusion Matrix



Accuracy = (TP + TN) / (P + N) = 83.3%

Precision = TP / (TP + FP) = 99.3%

Recall = TP / (TP + FN) = 84.1%

$$\textbf{\textit{F1}} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} = 91.3\%$$

Conclusion

- Accuracy, Robustness & Interpretability.
- Early and accurate detection of heart disease can lead to timely interventions, improved patient outcomes, and reduced healthcare costs.
- In regard to our simulation, the features didn't have high mean decrease impurity and imbalanced dataset, which didn't help in good classification.

Discussion

Some Disadvantages:

- Memory-intensive
- Sensitive to noise

The future of heart disease prediction using Random Forest holds promising developments. Here are some future predictions:

- **Hybrid Models Ensemble:** Leveraging other ML models with RF to improve CHD risk detection.

- **Features selection importance:** Train RF after removing all correlated features as well as the remove features with low feature importance. This may improve Generalization at the cost of higher variance.

References

- Lecture Pattern Analysis Part 04: Random Forests and their Variants by Prof. Christian Riess
- Townsend, N., Nichols, M. S., Scarborough, P., & Rayner, M. (2015, October). Proportion of all deaths due to major causes in Europe, latest.
- Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning.
- Breiman, L. (2011, January). Random Forests. Statistics Department, University of California, Berkeley, CA 94720.
- Death Causes : www.who.int
- Dataset: www.kaggle.com
- Data Mining Concepts and Techniques – Jiawei Han, Micheline Kamber, Jian Pei

Thank you!