Knowledge on Heart Condition of Children based on Demographic and Physiological Features

CBMS 2013 – June 21st 2013 – Porto, Portugal

Pedro Ferreira
Tiago T. V. Vinhoza
Ana Castro
Felipe Mourato
Thiago Tavares
Sandra Mattos
Inês Dutra
Miguel Coimbra

DigiScope

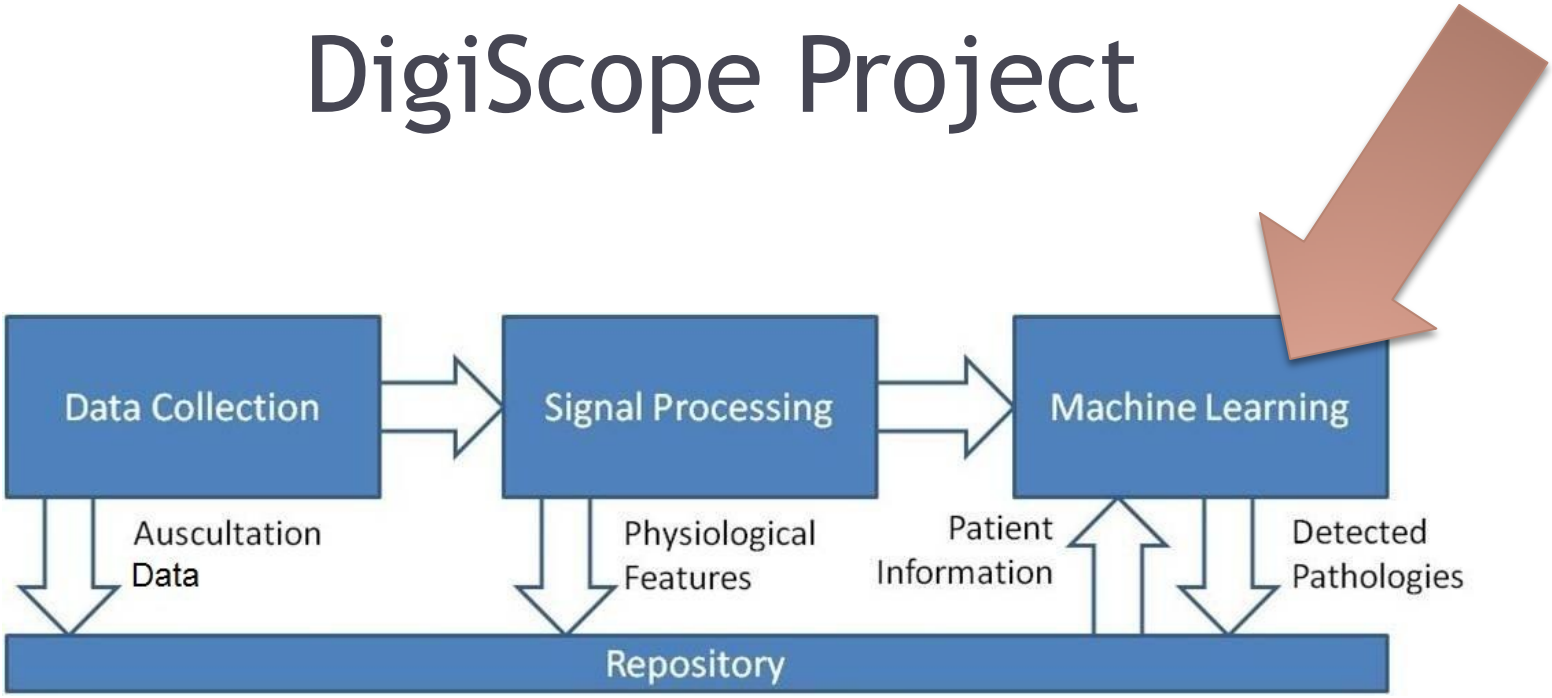
DigiScope Project

DigiScope

- **Help General Practitioners (GPs)** in their daily medical routine
- Capable of automatically **extract clinical features** from collected data
- May provide **clinical second opinion** on specific heart pathologies

DigiScope

DigiScope Project



DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

Heart Diseases in Children

- **6 million**
  - children **worldwide** suffer from heart disease [1]
- **500**
  - cardiac surgeries in children per year in **Portugal** [2]
- **8-10 out of 1000**
  - babies are born with a congenital heart disease in **Portugal**, **Brazil** and **USA** [2,3,4]

Sources:
1) *European Society of Cardiology* – June 2013
2) *Apifarma, Portuguese Association of the Pharmaceutical Industry* – June 2013
3) *Revista Brasileira de Cirurgia Cardiovascular*– June 2013
4) *Lucile Packard Children's Hospital at Stanford*– June 2013

DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

Objectives

- Study **relations between** demographic and physiological **features** in the occurrence of a **pathological/non-pathological** heart condition in children

- **Build classifiers** that, in a automatic way, **distinguish between normal** and **pathological cases**

DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

State of the Art

UCIRVINE

- Cleveland database
- **<u>Goal:</u>** distinguish **presence/absence** of a **cardiac disease**
  - Presence {1,2,3,4}
  - Absence {0}

DigiScope

State of the Art

- [1] D. Aha and D. Kibler, "**Instance-based prediction of heart-disease presence with the Cleveland database"**, tech. rep., University of California, Mar. 1988.
  - **Accuracy: 75.7%**
- [2] S. M. Kamruzzaman, A. R. Hasan, A. B. Siddiquee, and M. E. H. Mazumder, **"Medical diagnosis using neural network"**, in 3rd International Conference on Electrical & Computer Engineering (ICECE), pp. 28–30, Dec. 2004.
  - **Accuracy: 87.5%**
- [3] B. O'Hora, J. Perera, and A. Brabazon, **"Designing radial basis function networks for classification using differential evolution"**, inProc. International Joint Conference on Neural Networks (IJCNN), pp. 2932 –2937, 2006.
  - **Accuracy: 84%**

DigiScope

State of the Art

- [4] J. Wu, J. Roy, and W. F. Stewart, **"Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches"**, Medical Care, vol. 48, pp. 106–113, Jun. 2010.

  - **<u>Result:</u> detection of heart failure** more than **6 months before** the actual date of clinical **diagnosis**
    - **AUC: 0.77**

DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

Methodology

Dataset

- Recife, Pernambuco – Brazil
- Collected between **October 2003** to **September 2009**
- [**2-19**] year old **children**
  - Average age: 8.60

DigiScope

Methodology

17k

1st phase

7603

2nd phase

7199 instances

404 instances removed from phase 1 to phase 2

Dataset

Preprocessing tasks

data cleaning

data transformation

data normalization

2507 (34.8%) pathological (+)

4692 (65.2%) normal (-)

DigiScope

Methodology

7199 instances

Dataset

Preprocessing tasks

33 attributes

removal of irrelevant features*

17 attributes

* patient ID, name of the physician, health insurance information, etc.

DigiScope

Methodology

Attribute
Height (cm)
Weight (kg)
Sex
Age Range
Body Mass Index Percentile
Systolic Blood Pressure (SBP)
Diastolic Blood Pressure (DBP)
Result-SBP-DBP
Murmur
Second Heart Sound (S2)
Pulses
Heart Rate (bpm)
Current Disease History 1 (CDH 1)
Current Disease History 2 (CDH 2)
Primary Reason
Secondary Reason
Pathology (class)

Dataset

17 attributes

Note:

Some of the attributes are in fact **annotations provided by a cardiologist**, not features extracted from the raw sound data itself

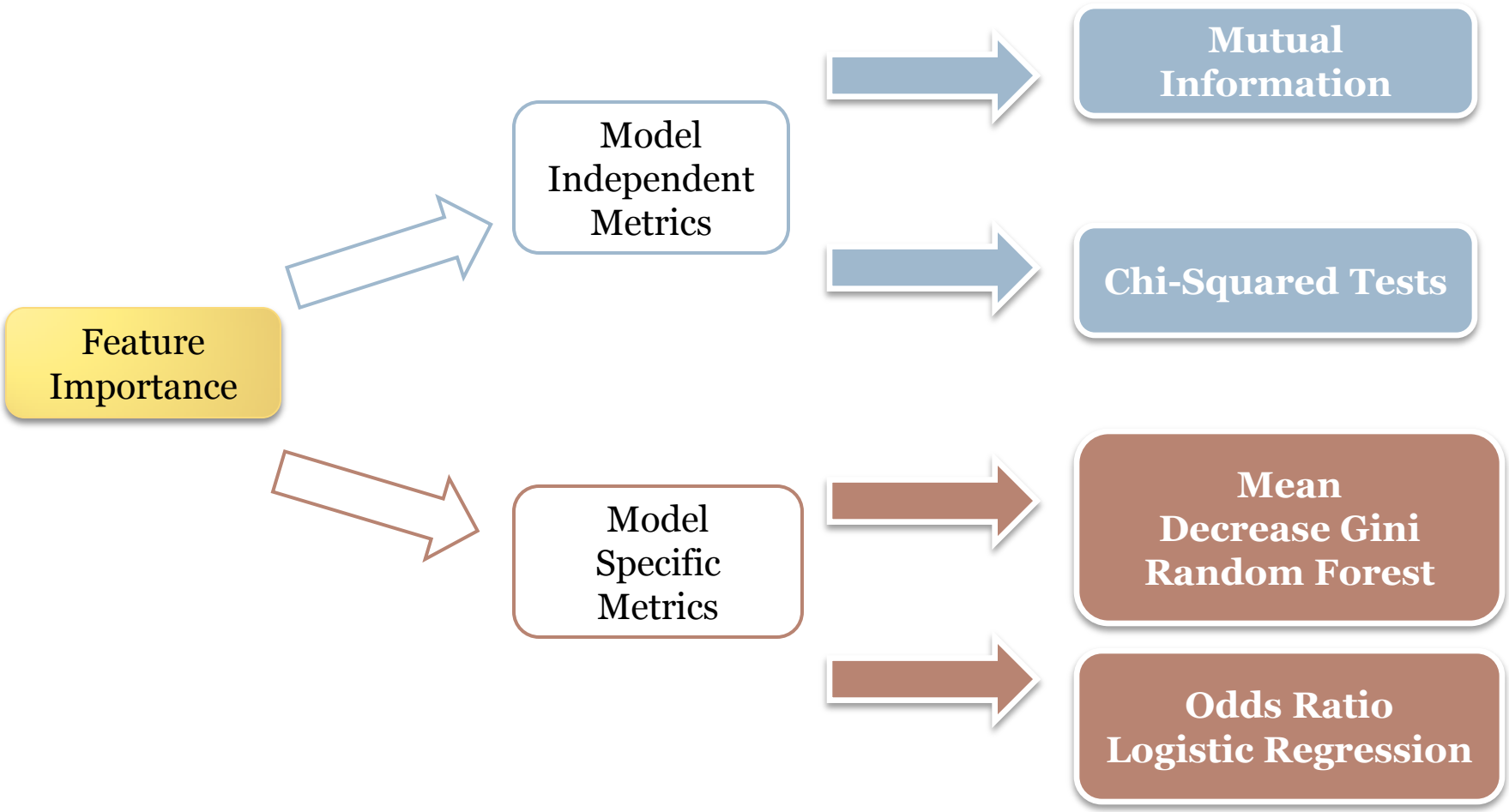
DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

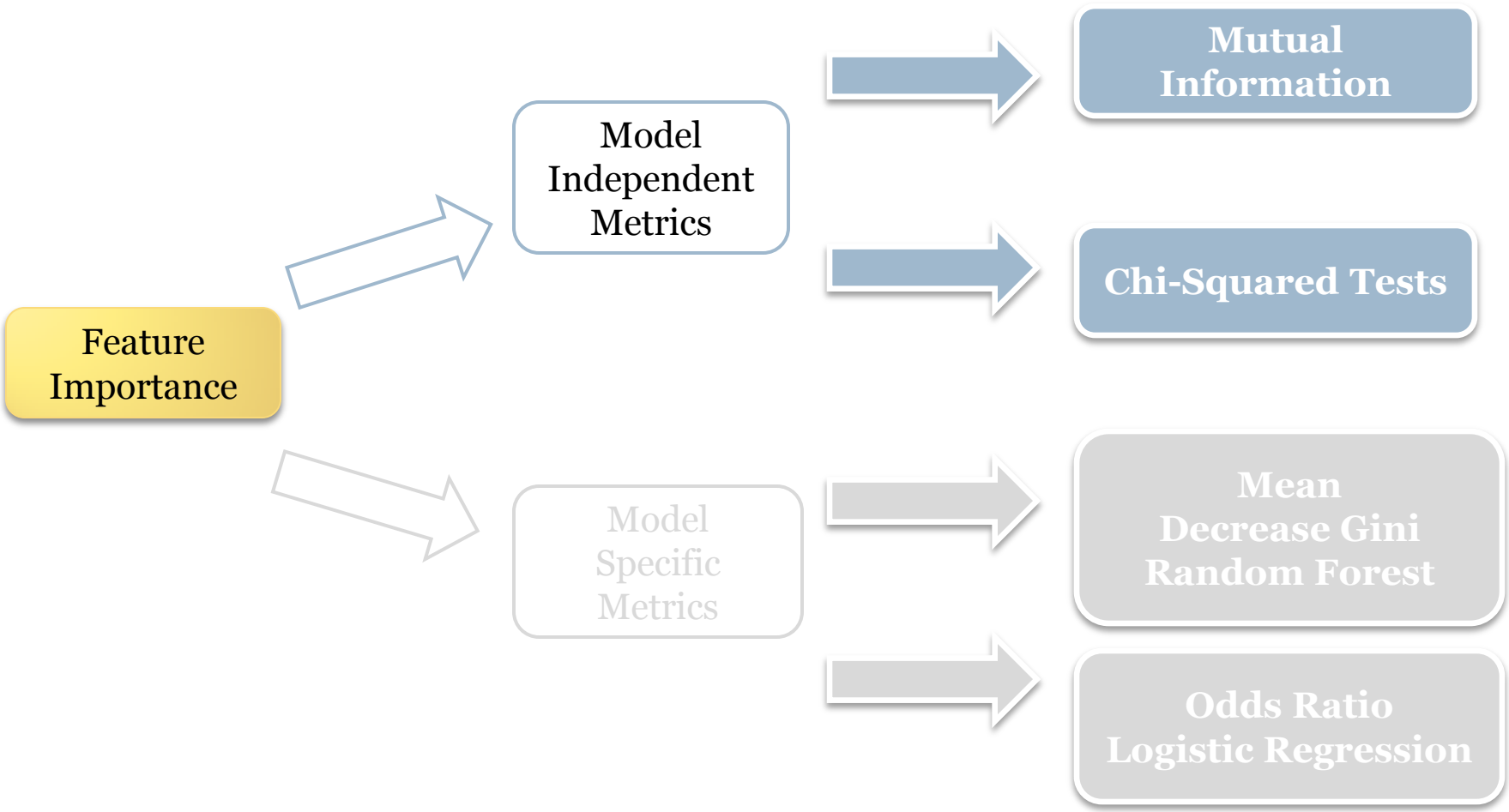
DigiScope

Methodology



DigiScope

Methodology



DigiScope

Methodology

Feature Importance

Model Independent Metrics → **Mutual Information**

- The **mutual information** tells how the knowledge of a variable Y **reduces the uncertainty** about a variable X:

$$I(X;Y) = H(X) - H(X|Y)$$

- We use a **normalized version** (bounded between 0 and 1):

$$I_{\text{norm}}(X;Y) = I(X;Y)/H(X) = 1 - H(X|Y)/H(X)$$

DigiScope

Methodology

Feature Importance

Model Independent Metrics

Mutual Information

Results

All 7199 cases

Feature	Score
Murmur	0.61
Secondary Reason	0.10
Weight	0.09
Primary Reason	0.05
HR	0.05

5000 cases where Murmur = Absent

Murmur Absent

Absent – 5000

Feature	Score
Weight	(60%)
HR	
Height	
S2	0.04
Secondary Reason	

Continuous –

Diastolic – 6 (

Systolic 218

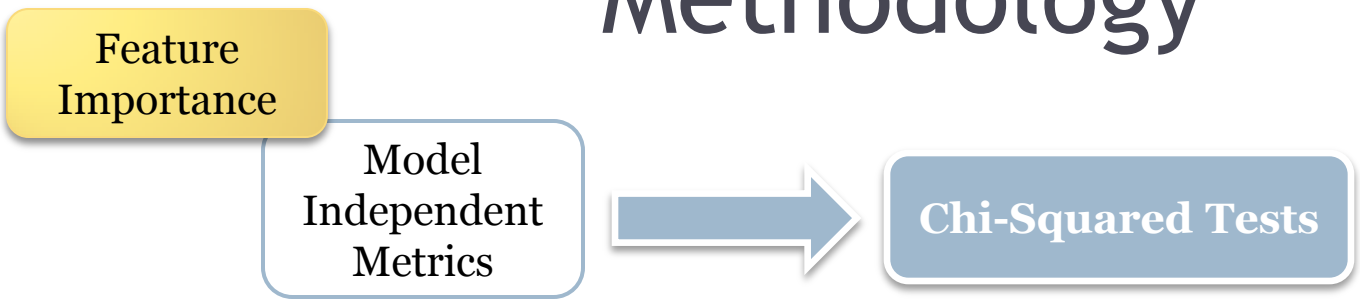
5000

404 (8.1%) pathological (+)

4596 (91.9%) normal (-)

DigiScope

Methodology



- The **chi-squared test** is used to test **two** different **hypothesis**:
  - The **variables** are **dependent**;
  - The **variables** are **independent**.

DigiScope

Methodology

Feature Importance

Model Independent Metrics

Chi-Squared Tests

Results

All 7199 cases

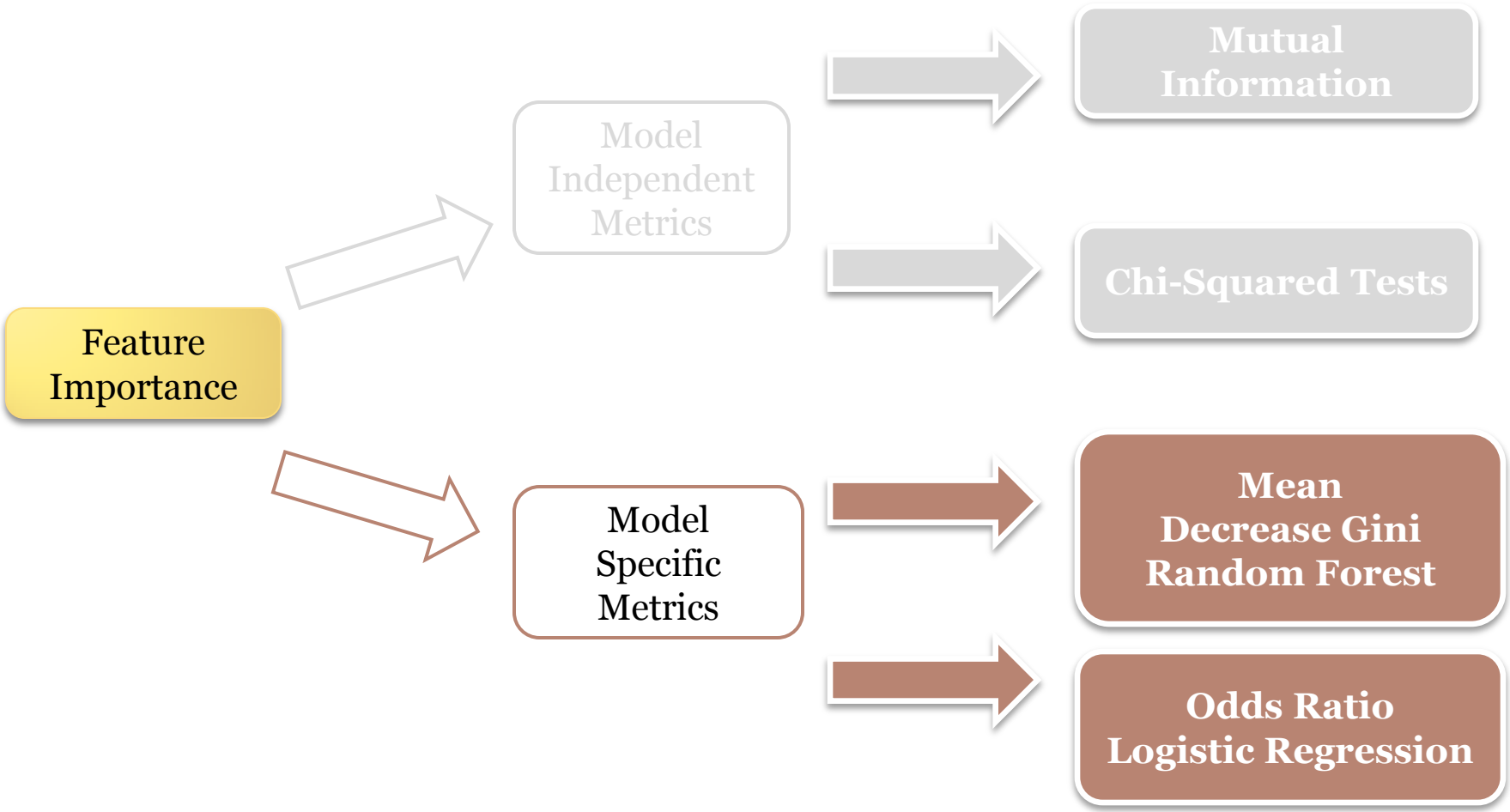
Feature	Score
Murmur	5160.57
Weight	730.00
Secondary Reason	702.92
Primary Reason	501.63
HR	445.33

5000 cases where Murmur = "Absent"

Feature	Score
Weight	577.86
S2	255.56
HR	149.01
Height	125.22
Secondary Reason	62.89

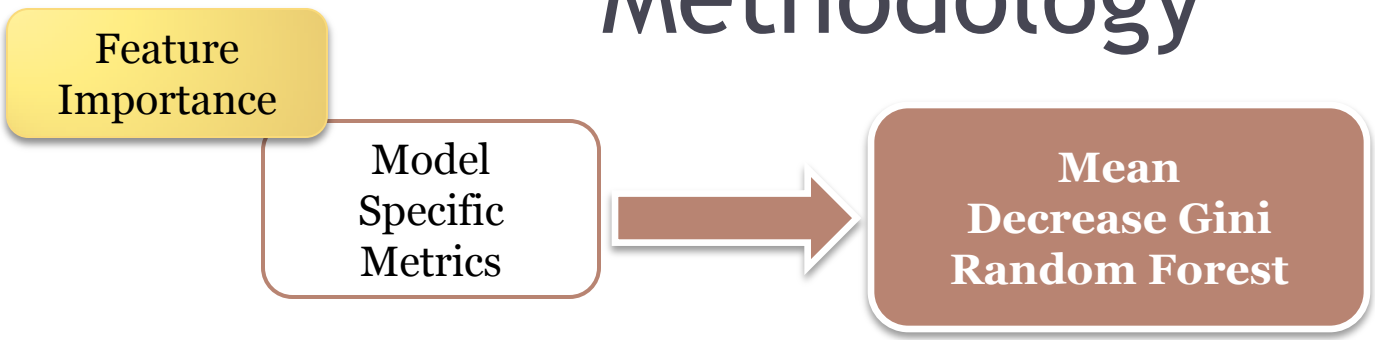
DigiScope

Methodology



DigiScope

Methodology



- We calculate the **variable importance** as **measured** by a **random forest** classifier
- Variable importance is related to the degree of **node purity**
- **Mean Decrease Gini:** related to the Gini Index which shows how **unequal** is the **frequency of occurences** in a distribution

DigiScope

Methodology

Feature Importance

Model Specific Metrics

Mean Decrease Gini Random Forest

Results

All 7199 cases

Feature	Mean Decrease Gini
Murmur	1975.98
Secondary Reason	216.44
Weight	189.82
Height	172.65
HR	149.61

5000 cases where Murmur = "Absent"

Feature	Mean Decrease Gini
Weight	131.59
Height	119.36
HR	84.64
CDH 1	48.05
Secondary Reason	46.45

DigiScope

Methodology

Feature Importance

Model Specific Metrics

Odds Ratio Logistic Regression

- In a **logistic regression**, we can think of the **class variable *x*** as having a **Bernoulli distribution** with parameter *p* given by:

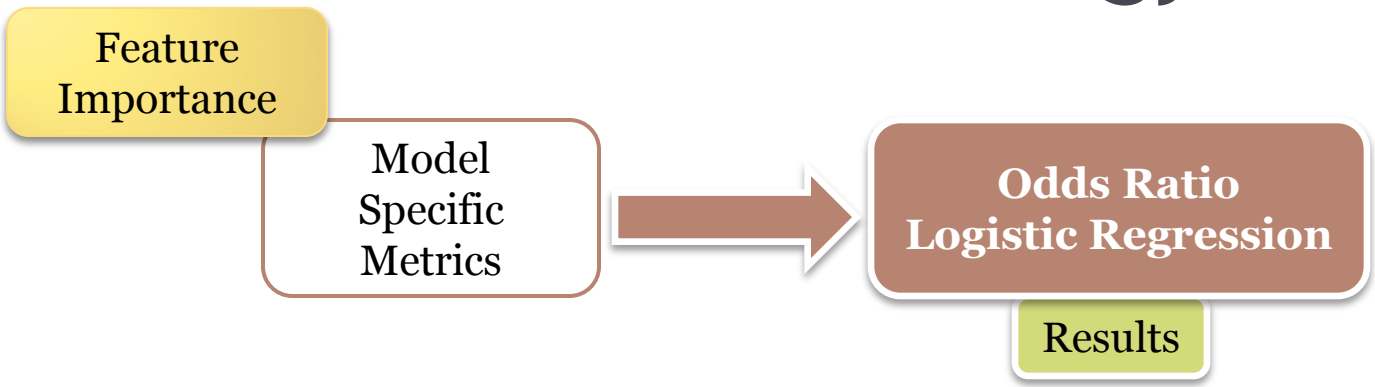
$$p = P(x = 1|\mathbf{\Theta}^T\mathbf{y}) = h\left(\mathbf{\Theta}^T\mathbf{y}\right)$$

- **y** is the feature vector and **Θ** are the regression coefficient vector
- Categorical features are converted into binary features
  - **E.g. Murmur ∈ {Absent, Systolic, Diastolic, Continuous}**

Murmur_Absent ∈ {0,1}
Murmur_Systolic ∈ {0,1}
Murmur_Diastolic ∈ {0,1}
Murmur_Continuous ∈ {0,1}

DigiScope

Methodology



- **Odds Ratio:** how an **increase (presence)** of a numerical (categorical) **feature influence** the **probability of ocurrence** of the **class** variable

pathology

  - **Murmur_Systolic: 320**
  - **S2_Hyperphonetic: 6**

DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

Classification Procedure

- **Training** set:

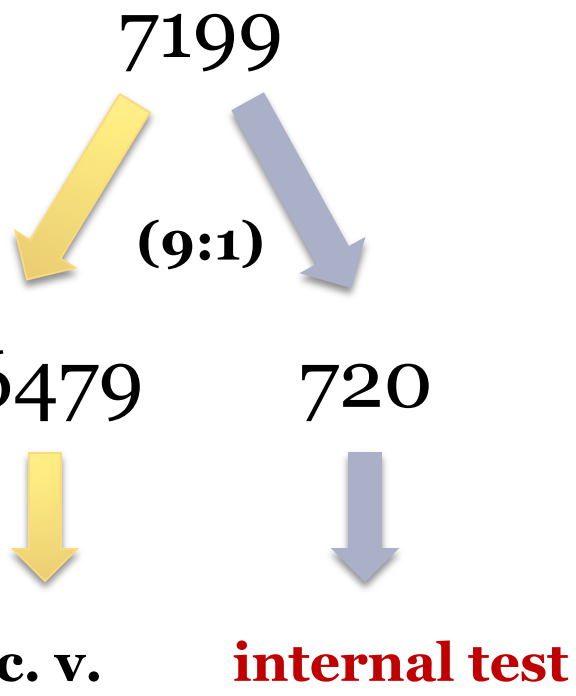
  7199 cases

- **External Test** set:

  169 cases

  (from previous work [5])

- **Nested Cross-Validation**



- [5] P. Ferreira et al., "**Detecting cardiac pathologies from annotated auscultations**", in Proc. International Symposium on Computer-Based Medical Systems (CBMS), 2012.

DigiScope

Classification – Algorithms

- ZeroR (baseline classifier)
- OneR
- DTNB
- PART

rules

- J48
- DecisionStump
- RandomForest
- SimpleCart
- NBTree

trees

- NaiveBayes
- BayesNet (TAN)

bayes

- SMO

functions

- AdaBoostM1
- Bagging
- Dagging
- Grading
- Stacking
- Vote

meta-learning

DigiScope

Classification Procedure

- **Training** set:

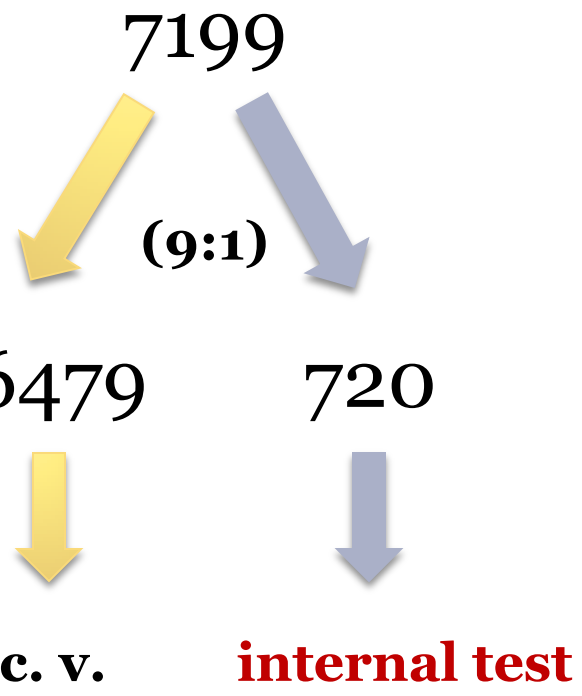
  7199 cases

- **External Test** set:

  169 cases

  (from previous work [5])

- **Nested Cross-Validation**



- [5] P. Ferreira et al., "**Detecting cardiac pathologies from annotated auscultations**", in Proc. International Symposium on Computer-Based Medical Systems (CBMS), 2012.

DigiScope

Classification – Results

7199

Metrics	Nested c.v.	internal test
CCI (%)	93.31	**93.32**
Sensitivity	0.85	**0.85**
Specificity	0.98	**0.98**
AUC	0.93	**0.93**

Best algorithm in all folds: **NaiveBayes**

169[5]

Metrics	Nested c.v.	internal test
CCI (%)	91.56	90.53
Sensitivity	0.72	0.70
Specificity	0.98	0.97
AUC	0.85	0.83

- [5] P. Ferreira et al., "**Detecting cardiac pathologies from annotated auscultations**", in Proc. International Symposium on Computer-Based Medical Systems (CBMS), 2012.

DigiScope

Classification Procedure

- **Training** set:

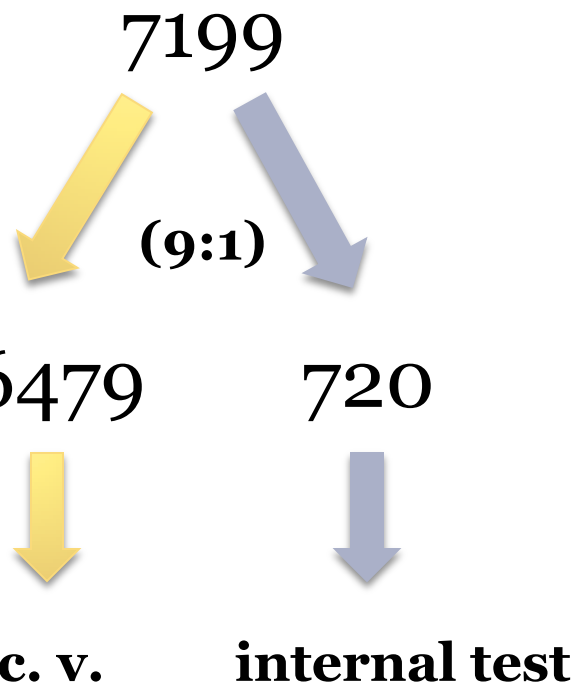
  7199 cases

- **External Test** set:

  169 cases

  (from previous work [5])

- **Nested Cross-Validation**



- [5] P. Ferreira et al., "**Detecting cardiac pathologies from annotated auscultations**", in Proc. International Symposium on Computer-Based Medical Systems (CBMS), 2012.

DigiScope

Classification – Results

7199

169[5]

Best algorithm in all folds: **NaiveBayes**

NaiveBayes model applied

Metrics	Nested c.v.	internal test	external test (169)
CCI (%)	93.31	**93.32**	**91.12**
Sensitivity	0.85	**0.85**	**0.73**
Specificity	0.98	**0.98**	**0.97**
AUC	0.93	**0.93**	**0.85**

- [5] P. Ferreira et al., "**Detecting cardiac pathologies from annotated auscultations**", in Proc. International Symposium on Computer-Based Medical Systems (CBMS), 2012.

DigiScope

Outline

- Heart Diseases in Children
- Objectives
- State of the Art
- Methodology
  - Dataset
  - Feature Importance
    - Model Independent Metrics
    - Model Specific Metrics
- Classification Tasks
- Conclusions and Future Work

DigiScope

Conclusions and Future Work

a) It is **crucial** to have accurate **information on murmur presence**, according to the feature importance metrics

b) **Nested Cross-Validation** produced a **model** that can achieve a **performance of 91.1%**, **sensitivity of 0.73** and **specificity of 0.97** on predicting cardiac pathologies on an **external dataset**

DigiScope

Conclusions and Future Work

a) **Build** classifiers when **murmur = absent**

b) Try to correctly **distinguish innocent murmurs from pathological** ones

   i. **Detailed murmur description**

c) **Incorporate models** in the **DigiScope** Prototype, for cardiac pathology assessment

Thank you!

pedroferreira@dcc.fc.up.pt

www.dcc.fc.up.pt/~pedroferreira

Appendices

DigiScope

Methodology

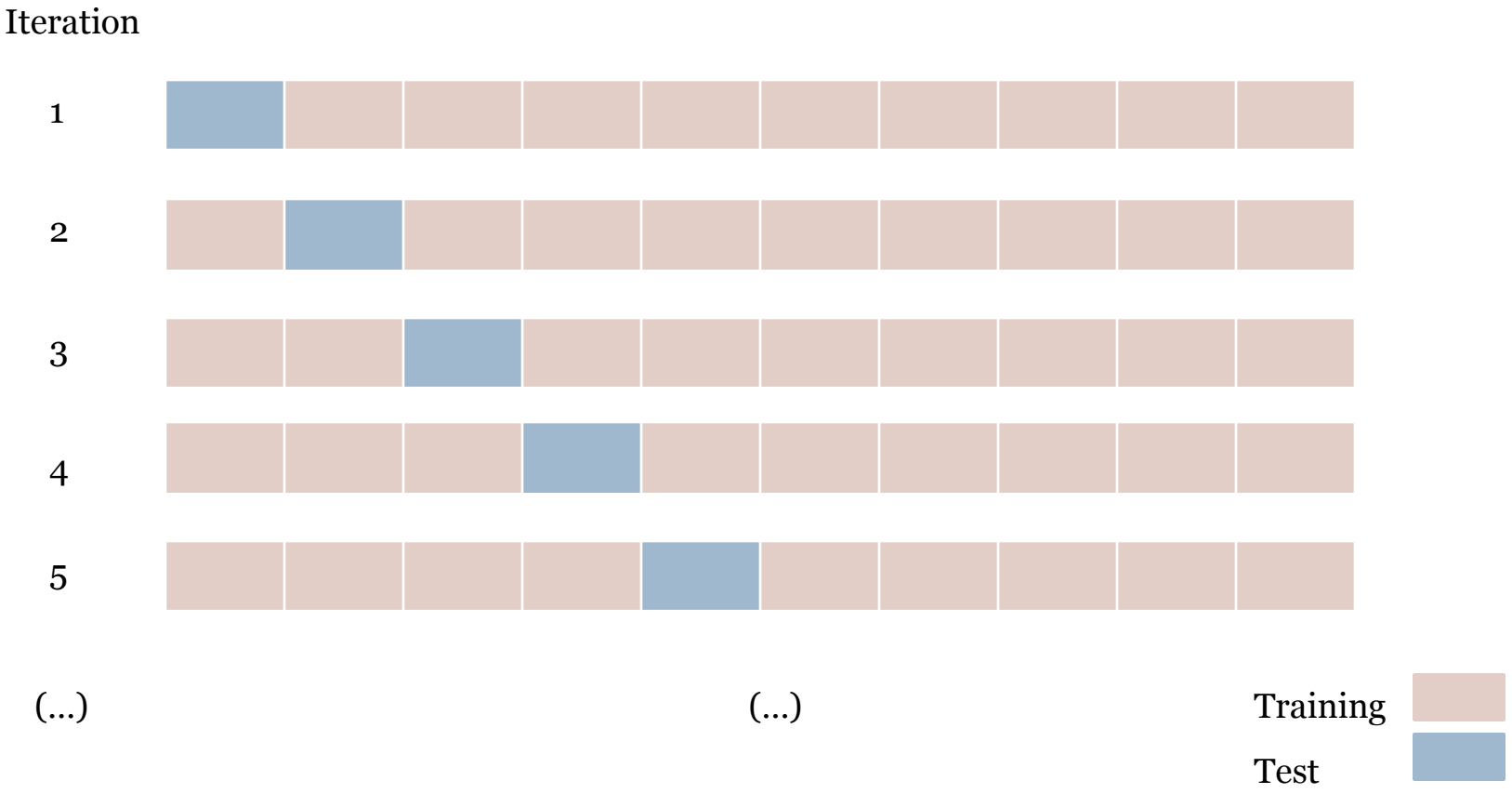
Dataset

Attribute	Value
Height (cm)	Numeric
Weight (kg)	Numeric
Sex	{Female, Male}
Age Range	{Pre-School, School, Pre-Teen, Teenager}
Body Mass Index Percentile	{Low Weight, Normal, Overweight, Obese}
Systolic Blood Pressure (SBP)	{Normal, Limit, Hypertense}
Diastolic Blood Pressure (DBP)	{Normal, Limit, Hypertense}
Result-SBP-DBP	{Normal, Limit, Hypertense}
Murmur	{Absent, Systolic, Diastolic, Continuous}
Second Heart Sound (S2)	{Normal, Fixed Split, Unique, Hyperphonetic}
Pulses	{Normal, Diminished Femoral}
Heart Rate (bpm)	Numeric
Current Disease History 1 (CDH 1)	{Asymptomatic, Cyanosis, Precordial pain, Dyspnea, Palpitation, Faint/Dizziness, Weight Gain}
Current Disease History 2 (CDH 2)	{Cyanosis, Precordial pain, Dyspnea, Palpitation, Faint/Dizziness, Weight Gain}
Primary Reason	{Cardiopathy, Routine check-up, Cardiology Screening, Possible Cardiopathy, Others}
Secondary Reason	{Physical Activity, Congenital Cardiopathy, Surgery, Risk factors, Presence of Murmurs, Others}
Pathology (class)	{Yes, No}

DigiScope

Classification

10 x 10 fold stratified cross-validation



DigiScope

Classification – Metrics

CCI
K
MAE
Sensitivity
Specificity
Precision
F-Measure
AUC