POLITECHNIKA GDAŃSKA

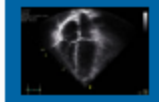
AI TECH

Wprowadzenie do Sztucznej Inteligencji

Marta Arendt
Maciej Mechliński
Michał Gąsecki
Stanisław Rachwał

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Baza danych: *"Echocardiogram"*



Echocardiogram

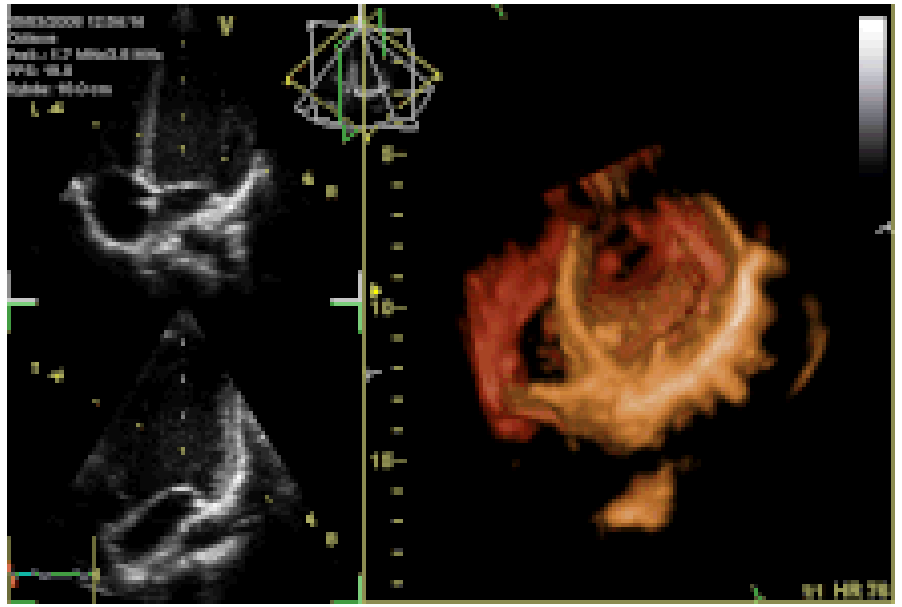
Donated on 2/28/1989

Data for classifying if patients will survive for at least one year after a heart attack

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Life	Classification
Attribute Type	**# Instances**	**# Attributes**
Categorical, Integer, Real	132	12

SOTA - Echokardiografia

- Badanie kardiologiczne znane pod innymi nazwami jako USG serca, echo serca oraz UKG. Jest jednym z najpopularniejszych metod badania serca (obok EKG).
- Umożliwia w sposób nieinwazyjny ocenę struktur anatomicznych serca, a także ocenę ruchu mięśnia sercowego i zastawek wewnątrzsercowych oraz przepływ krwi w obrębie przedsionków i komór serca, dużych naczyń sercowych (aorta, żyły główne, tętnica i żyły płucne) i naczyń wieńcowych.



SOTA – Baza danych

- Wybrana baza jest obecnie nieużywana - niepełna i archaiczna (tylko zastosowania edukacyjne) (1989 r.).
- Ostatnie odnalezione użycie bazy w artykułach - rok 2005 r.
- Przeznaczona dla zagadnień klasyfikacji binarnej.
- Wykorzystane klasyfikatory: kNN (słabe wyniki: dokładność ~60%), samodzielnie stworzone algorytmy.

Dane zawarte w bazie danych

1. survival -- the number of months patient survived (has survived, if patient is still alive). Because all the patients had their heart attacks at different times, it is possible that some patients have survived less than one year but they are still alive. Check the second variable to confirm this. Such patients cannot be used for the prediction task mentioned above.
2. still-alive -- a binary variable. 0=dead at end of survival period, 1 means still alive
3. age-at-heart-attack -- age in years when heart attack occurred
4. pericardial-effusion -- binary. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid
5. fractional-shortening -- a measure of contracility around the heart lower numbers are increasingly abnormal
6. epss -- E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.
7. lvdd -- left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
8. wall-motion-score -- a measure of how the segments of the left ventricle are moving
9. wall-motion-index -- equals wall-motion-score divided by number of segments seen. Usually 12-13 segments are seen in an echocardiogram. Use this variable INSTEAD of the wall motion score.
10. mult -- a derivate var which can be ignored
11. name -- the name of the patient (I have replaced them with "name")
12. group -- meaningless, ignore it
13. alive-at-1 -- Boolean-valued. Derived from the first two attributes. 0 means patient was either dead after 1 year or had been followed for less than 1 year. 1 means patient was alive at 1 year.

Prezentacja danych przed oczyszczeniem

```
1 df = pd.read_csv('echocardiogram.csv', on_bad_lines='skip', names=['survival', 'still_alive', 'age_at_heart_attack', 'pericardial_effusion', 'fractional_shortening', 'epss', 'lvdd', 'wall_motion_score',
  'wall_motion_index', 'mult', 'name', 'group', 'alive_at_1'])
2 df.head(10)
```

Executed in 31ms, 10 May at 22:00:46

10 rows | 10 rows × 13 columns pd.Dataframe | CSV

	survival	still_alive	age_at_heart_attack	pericardial_effusion	fractional_shortening	epss	lvdd	wall_motion_score	wall_motion_index	mult	name	group	alive_at_1
0	11	0	71	0	0.260	9	4.600	14	1	1	name	1	0
1	19	0	72	0	0.380	6	4.100	14	1.700	0.588	name	1	0
2	16	0	55	0	0.260	4	3.420	14	1	1	name	1	0
3	57	0	60	0	0.253	12.062	4.603	16	1.450	0.788	name	1	0
4	19	1	57	0	0.160	22	5.750	18	2.250	0.571	name	1	0
5	26	0	68	0	0.260	5	4.310	12	1	0.857	name	1	0
6	13	0	62	0	0.230	31	5.430	22.5	1.875	0.857	name	1	0
7	50	0	60	0	0.330	8	5.250	14	1	1	name	1	0
8	19	0	46	0	0.340	0	5.090	16	1.140	1.003	name	1	0
9	25	0	54	0	0.140	13	4.490	15.5	1.190	0.930	name	1	0

Dane po wstępnym przetworzeniu

```
df_values_drop_row_reindex = pd.concat([df_values_drop1, df_values_drop2], ignore_index=True)
df_values_drop_row_reindex.head(10)
```

	survival	still_alive	age_at_heart_attack	pericardial_effusion	fractional_shortening	epss	lvdd	wall_motion_index	alive_at_1
0	11.0	0	71	0	0.260	9	4.600	1	0
1	10.0	0	57	0	0.240	14.800	5.260	1.380	NaN
2	10.0	0	66	0	0.290	15.600	6.150	1	0
3	9.0	0	73	0	0.12	NaN	6.78	1.39	NaN
4	19.0	0	72	0	0.380	6	4.100	1.700	0
5	16.0	0	55	0	0.260	4	3.420	1	0
6	57.0	0	60	0	0.253	12.062	4.603	1.450	0
7	19.0	1	57	0	0.160	22	5.750	2.250	0
8	26.0	0	68	0	0.260	5	4.310	1	0
9	13.0	0	62	0	0.230	31	5.430	1.875	0

```
[61] df_values_drop_row_reindex.shape
```

```
(92, 9)
```

Podstawowe statystyki

	survival	still_alive	age_at_heart_attack	pericardial_effusion	fractional_shortening	epss	lvdd	wall_motion_index	alive_at_1
count	92.00000	92.000000	89	92.000000	89	83	88	92	48
unique	NaN	NaN	31	NaN	59	65	80	47	1
top	NaN	NaN	62	NaN	0.20	0	4.48	1	0
freq	NaN	NaN	8	NaN	4	6	3	33	48
mean	30.12500	0.086957	NaN	0.130435	NaN	NaN	NaN	NaN	NaN
std	11.50003	0.283315	NaN	0.338627	NaN	NaN	NaN	NaN	NaN
min	9.00000	0.000000	NaN	0.000000	NaN	NaN	NaN	NaN	NaN
25%	22.00000	0.000000	NaN	0.000000	NaN	NaN	NaN	NaN	NaN
50%	29.00000	0.000000	NaN	0.000000	NaN	NaN	NaN	NaN	NaN
75%	36.25000	0.000000	NaN	0.000000	NaN	NaN	NaN	NaN	NaN
max	57.00000	1.000000	NaN	1.000000	NaN	NaN	NaN	NaN	NaN

Źródła korzystające z bazy

- Sebban, Marc & Nock, Richard & Dept, Grimaag. (2003). Stopping Criterion for Boosting-Based Data Reduction Techniques: From Binary to Multiclass Problems. https://www.researchgate.net/publication/2559870_Stopping_Criterion_for_Boosting-Based_Data_Reduction_Techniques_From_Binary_to_Multiclass_Problems

Źródła korzystające z bazy

- Gabor Melli. A Lazy Model-Based Approach to On-Line Classification. University of British Columbia. 1989

https://www.researchgate.net/publication/2511678_A_Lazy_Model-Based_Approach_to_On-Line_Classification

- Federico Divina and Elena Marchiori. Handling Continuous Attributes in an Evolutionary Inductive Learner. Department of Computer Science Vrije Universiteit.

https://www.researchgate.net/publication/3418815_Handling_Continuous_Attributes_in_an_Evolutionary_Inductive_Learner

Źródła korzystające z bazy

- D. Randall Wilson and Roel Martinez. Improved Center Point Selection for Probabilistic Neural Networks. Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms.

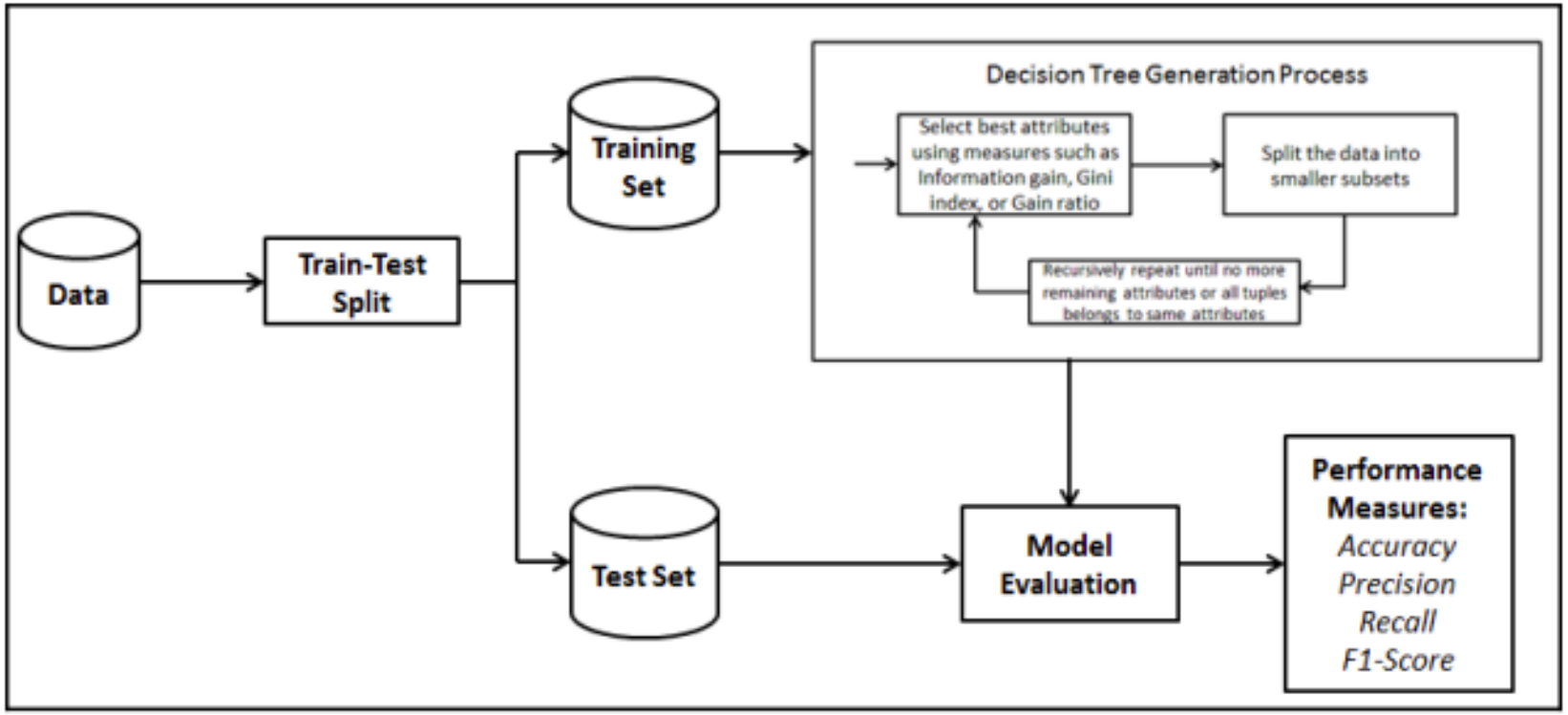
https://axon.cs.byu.edu/papers/wilson.icannga97.rpnn.pdf

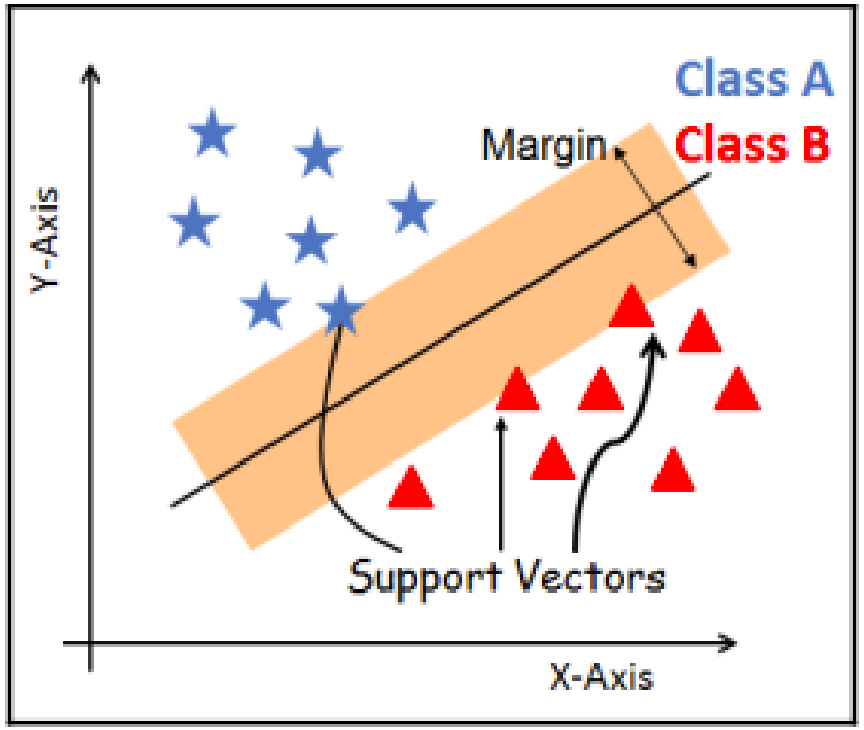
- Zhi-Hua Zhou and Xu-Ying Liu. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem

https://sci2s.ugr.es/keel/pdf/algorithm/articulo/2006%20-%20IEEE_TKDE%20-%20Zhou_Liu.pdf

Wybrane klasyfikatory [1]

- SVM (Support Vector Machine)
- Drzewa decyzyjne





- kNN (k-nearest neighbours)

Bibliografia

[1] Navlani, Avinash, Armando Fandango, and Ivan Idris. Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python. Packt Publishing Ltd, 2021. (data dostępu: 11.05.2023r.)

POLITECHNIKA GDAŃSKA

AI TECH

Dziękujemy

Maciej Mechliński, Marta Arendt

Stanisław Rachwał, Michał Gąsecki

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.