Faculty of Computer Science

Data Science

Moscow
2024

Neural FCA – Homework Report

Presenter: Vladislav Pendishchuk

Outline

1. Dataset & problem statement
2. Building neural networks based on concept lattices
3. Concept networks performance
4. Comparing FCA-based & classic algorithms
5. Conclusions

Dataset

Kaggle's Titanic survival prediction dataset was used for this task. The principal source for data about Titanic passengers is the Encyclopedia Titanica. The dataset contains data on some of the passengers on the Titanic's last voyage.

Details:

- 891 rows
- 12 features, including numerical and categorical data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, (…)	male	22	1	0	A/5 21171	7.25	NaN	S
2	1	1	Cumings, (…)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, (…)	female	26	0	0	STON/O2. 3101282	7.9250	NaN	S

Dataset

- **PassengerId** – the ID of the passenger from the original dataset
- **Survived** – 1 if survived, 0 if didn't survive
- **Pclass** – passenger class
- **Name**, **Sex**, **Age** – self-explanatory
- **SibSp** – the number of siblings and spouses aboard
- **Parch** – the number of parents and children aboard
- **Fare** – the passenger fare
- **Cabin** – the cabin number
- **Embarked** – the port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, (…)	male	22	1	0	A/5 21171	7.25	NaN	S
2	1	1	Cumings, (…)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, (…)	female	26	0	0	STON/O2. 3101282	7.9250	NaN	S

Problem statement

The goal is to predict the survival of passengers based on the given features. This is a binary classification problem with two classes:

- passengers who survived (Survived = 1);
- passengers who did not survive (Survived = 0).

FCA-based approach

Feature preprocessing

- Features *Name*, *PassengerId*, *Ticket*, *Fare* and *Cabin* were discarded due to a high number of missing values and high cardinalities.
- Numerical features *Age*, *Parch* and *SibSp* were binarized using ordinal greater-than-or-equal encoding.

Age	SibSp	Parch
22	1	0
38	1	0
26	0	0

→

Age >= 18	Age >= 40	Age >= 60	SibSp >= 1	SibSp >= 2	Parch >= 1	Parch >= 2
1	0	0	1	0	0	0
1	0	0	1	0	0	0
1	0	0	0	0	0	0

- Categorical features were binarized via one-hot encoding.

The resulting number of features after binarization: 15.

FCA-based approach

Building networks

1. Build a formal context based on the given list of features.
2. Build a monotone concept lattice.
   CbO algorithm has proven to be too slow for the given data. Its polynomial time alternative, Sofia, was used instead.
3. Search for formal contexts with the highest classification F1 score.
   Concepts with the full feature set intent is intentionally avoided, never mind its actual F1 score ranking.
4. Build the concept network based on the best concepts.
   N best concepts are selected, where N is the minimum number of concepts sufficient to cover all train objects by their extents.

FCA-based approach

Building networks

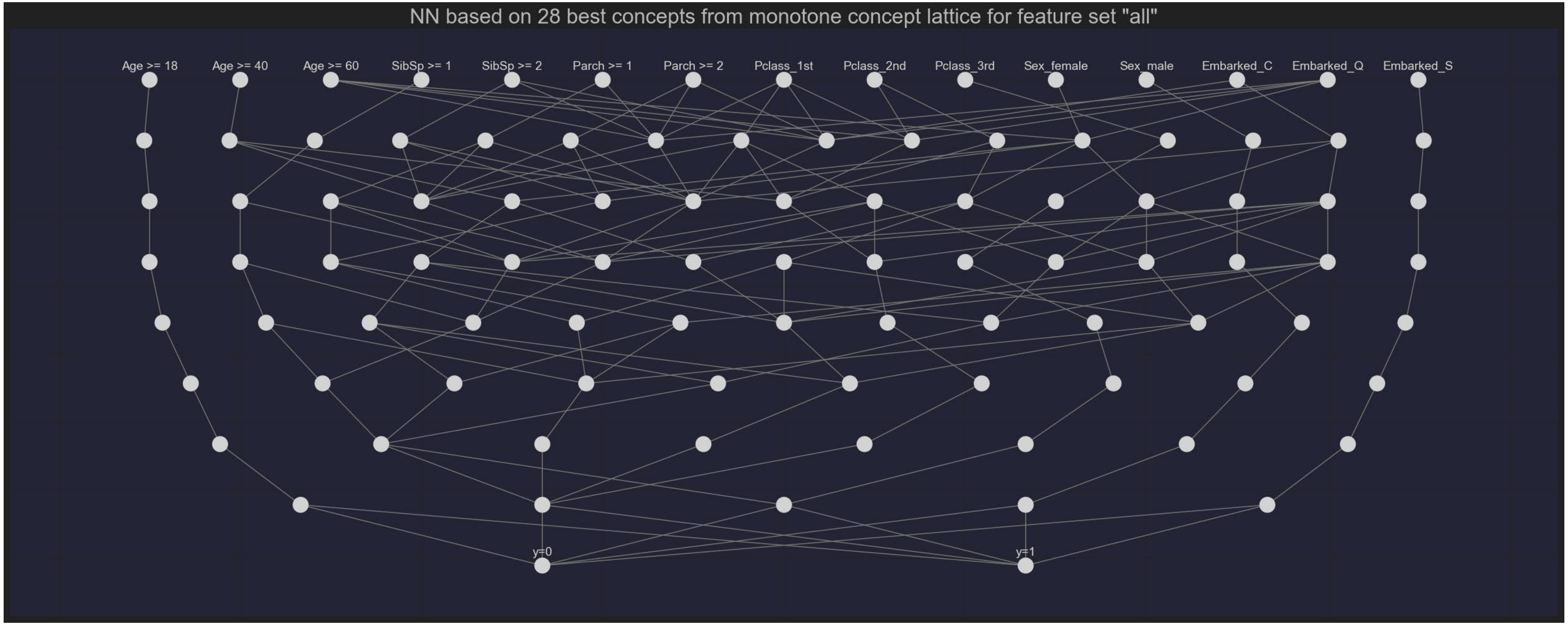
4 different sets of features were tested for building concept networks:

- all features (15 features);
- personal features – age & sex (5 features);
- passenger data – passenger class & port of embarkation (6 features);
- family-related data – SibSp and Parch (4 features).

It was discovered that it is impossible to build a concept network using only the family-related features.

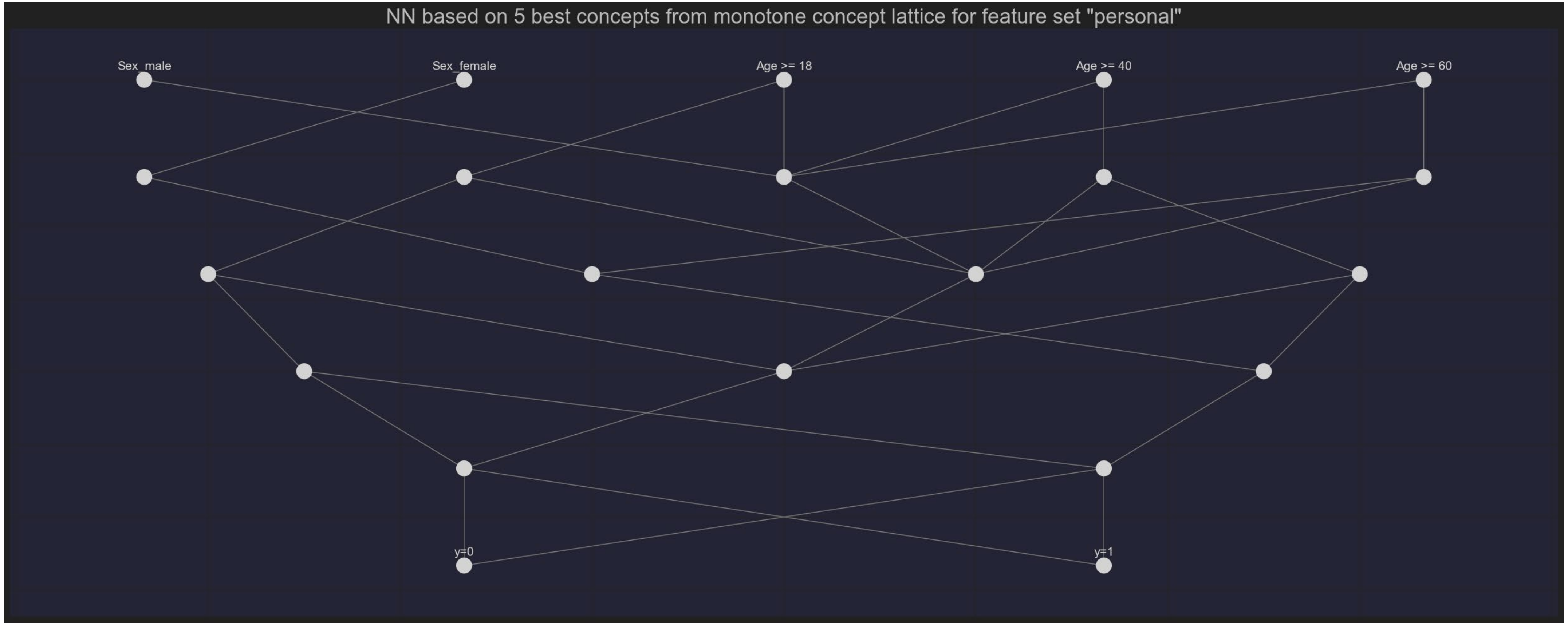
FCA-based approach

Building networks



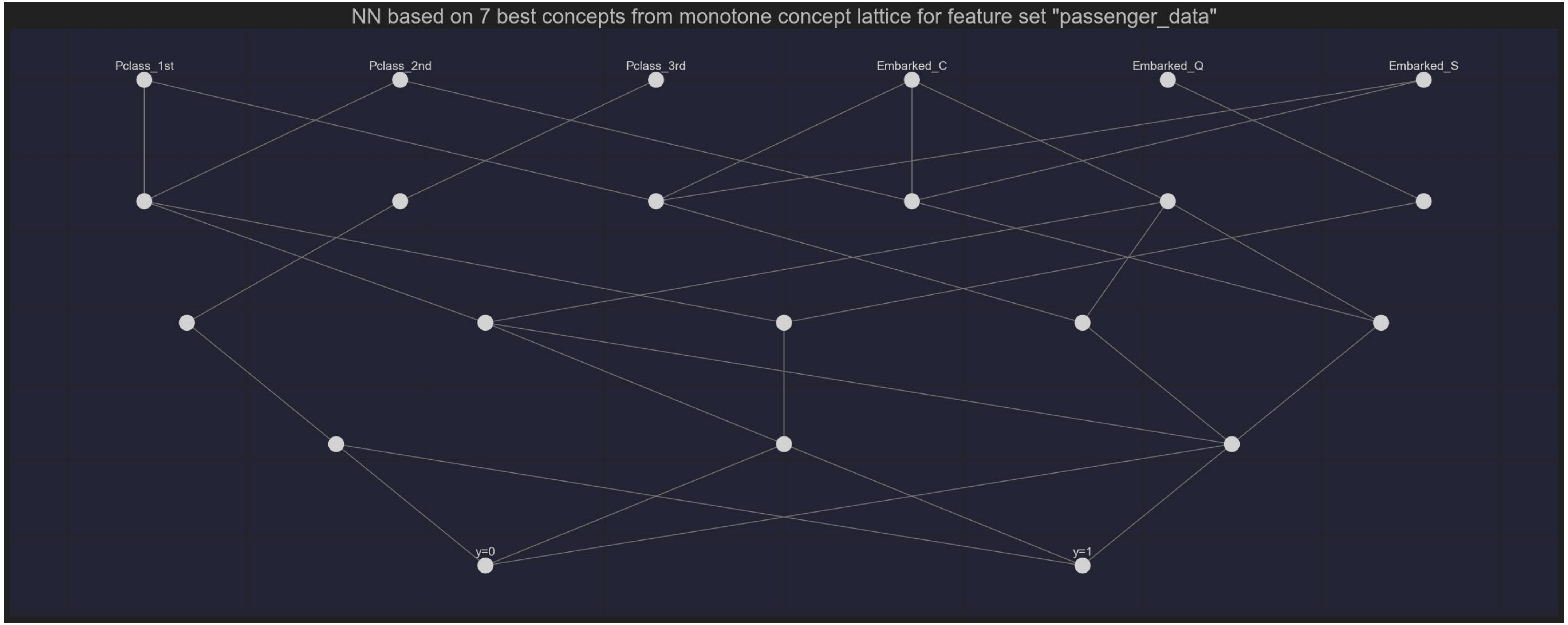
FCA-based approach

Building networks



FCA-based approach

Building networks



FCA-based approach

Training networks

For each architecture, networks were trained using different non-linearities:

- GELU (Gaussian Error Linear Unit):

$$GELU(x) = x \times \Phi(x), \text{ where } \Phi \text{ is the Gaussian distribution CDF}$$

- Leaky ReLU (Rectified Linear Unit):

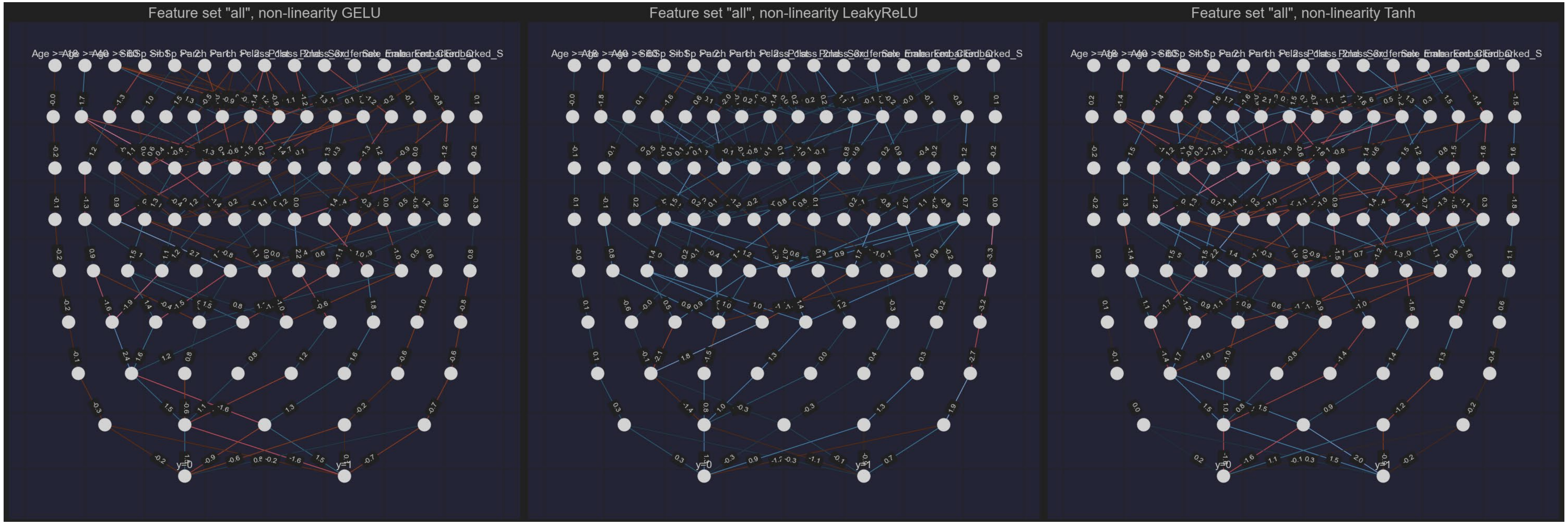
$$LeakyReLU(x) = \max(0, x) + \eta \times \min(0, x)$$

- Hyperbolic tangent:

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$$

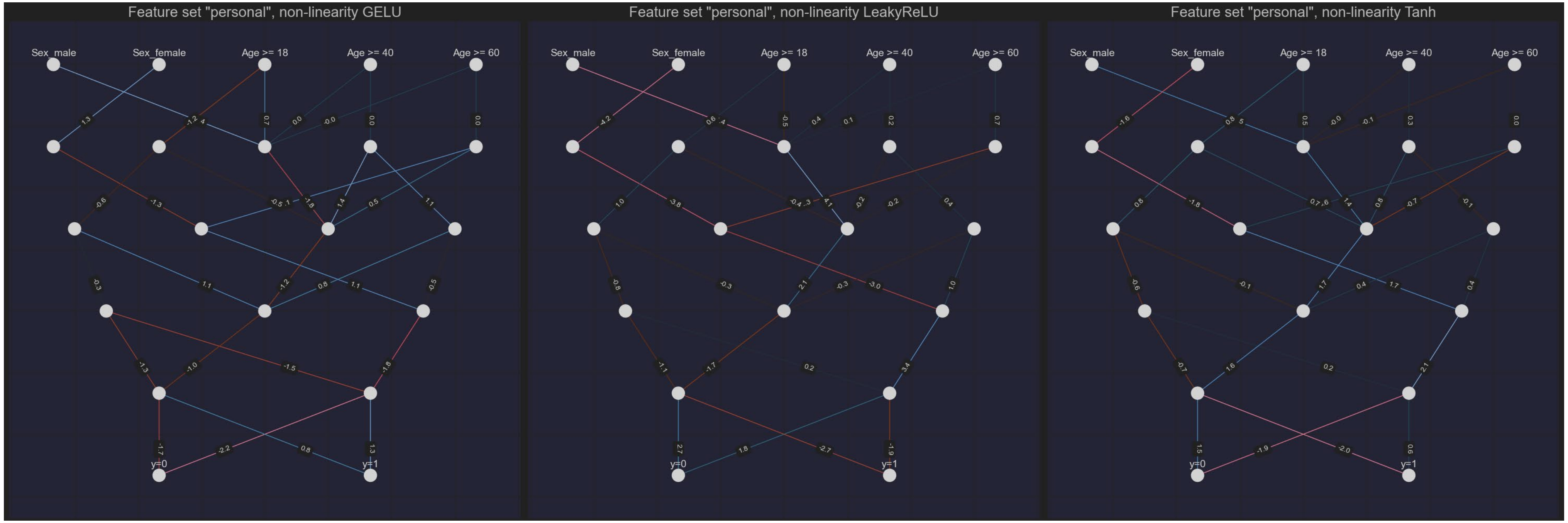
FCA-based approach

Training networks



FCA-based approach

Training networks



FCA-based approach

Performance

	Macro F1		
Feature set	GELU	Leaky ReLU	Tanh
All	0.721	0.739	0.728
Personal	0.761	0.749	0.761
Passenger data	0.589	0.589	0.589

Classic ML algorithms

The following algorithms were tested:

- Gaussian naïve Bayes;
- K nearest neighbors;
- Logistic regression;
- Decision tree classifier;
- Random tree classifier;
- XGBoost.

Feature preprocessing:

- Features *Name*, *PassengerId*, *Ticket*, *Fare* and *Cabin* were discarded for the same reasons.
- Categorical features were one-hot encoded.
- All of the resulting features were standardized.

Performance comparison

Model	Macro F1	Accuracy
Random Forest	0.807	0.836
KNN	0.774	0.804
Decision Tree	0.773	0.804
Concept Network	*0.761*	*0.785*
Logistic Regression	0.76	0.79
XGBoost	0.755	0.781
Gaussian NB	0.654	0.683

Conclusions

- While concept-based neural networks have the potential to provide insights into how features influence model predictions, in this specific case all of the architectures seem to be quite difficult to visually interpret for a human.
- Feature binarization drastically increases the dimensionality of data, making difficult to apply this approach to datasets with a high number of categorical features and\or features of high cardinality.
- Performance-wise the resulting network managed to outperform 3 out of 6 classic ML algorithms, placing it right in the middle of the performance ranking.

Thank you for your attention.

HSE UNIVERSITY • EST 1992 •