Geologist

Authors:

- Kristina Zheltova, SibSU
- Sofia Burmistrova, ITMO University

Mentor:

- Petr Andriushchenko, ITMO University

Plan

- Task description
- Data understanding
- Anomaly detection
- Data generation
- Filling missing values

Tasks

- Filling in missing values:
  - **Porosity** is the percentage of voids and open spaces in a rock [1].
  - **Tectonic regime** is characteristic of processes that control the structure and properties of the Earth's crust and its evolution through time.
- Assess the reliability of parameter values (anomaly detection).
- Generate oil deposit synthetic data.

[1] Mohammad Ali Ahmadi, Zhangxing Chen (2019), “Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs”

Why these tasks are important?

Incomplete data

- Fill missing values

Not enough data

- Generate new data

Enough data

- Verify the data

Review on existing solutions

- Wei Chen, Liuqing Yang **“Deep learning reservoir porosity prediction based on multilayer long short-term memory network”**
- Mohammad Ali Ahmadi, Zhangxing Chen **“Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs”**
- Fatai Adesina Anifowose **“Prediction of Porosity and Permeability of Oil and Gas Reservoirs using Hybrid Computational Intelligence Models”**

Data understanding

Dataset

- The dataset consists of 20 different categorical and continuous parameters of 513 oil deposit.
- There are 72 blanks in one categorical and one continuous parameter.

Field_name	Reservoir_unit	Country	Region	Basin_name	Tectonic_regime	Latitude	Longitude	Operator_company	...	Hydrocarbon_type_(main)	Reservoir_status_(current)	Structural_setting	Depth_(top_reservoir_ft_TVD)	Reservoir_period	Lithology_(main)	Thickness_(gross_average_ft)	Thickness_(net_pay_average_ft)	P
ABQAIQ	ARAB D	SAUDI ARABIA	MIDDLE EAST	THE GULF	COMPRESSION	26.0800	49.8100	SAUDI ARAMCO	...	OIL	REJUVENATING	FORELAND	6050	JURASSIC	LIMESTONE	250.0	184.0	
ABU GHARADIG	BAHARIYA	EGYPT	AFRICA	ABU GHARADIG	EXTENSION	29.7422	28.4925	GUPCO	...	GAS-CONDENSATE	MATURE PRODUCTION	RIFT	10282	CRETACEOUS	SANDSTONE	745.0	144.0	
ABU MADI-EL QARA	ABU MADI (LEVEL II)	EGYPT	AFRICA	NILE DELTA	STRIKE-SLIP	31.4382	31.3616	IEOC	...	GAS	DECLINING PRODUCTION	WRENCH	9843	NEOGENE	THINLY-BEDDED SANDSTONE	115.0	66.0	
ABU MADI-EL QARA	ABU MADI (LEVEL III)	EGYPT	AFRICA	NILE DELTA	STRIKE-SLIP	31.4382	31.3616	IEOC	...	GAS	DECLINING PRODUCTION	WRENCH	10499	NEOGENE	SANDSTONE	509.0	410.0	
AL HUWAISAH	SHUAIBA	OMAN	MIDDLE EAST	FAHUD SALT	COMPRESSION	21.9607	56.0452	PDO	...	OIL	REJUVENATING	SALT	4955	CRETACEOUS	LIMESTONE	250.0	100.0	
ALABAMA FERRY	UPPER GLEN ROSE D ZONE	USA	NORTH AMERICA	GULF OF MEXICO NORTHERN ONSHORE	GRAVITY	31.2143	-95.7981	NUMEROUS	...	OIL	MATURE PRODUCTION	PASSIVE MARGIN	8700	CRETACEOUS	LIMESTONE	95.0	15.0	
ALBA	ALBA	UK	EUROPE	NORTH SEA CENTRAL	NaN	58.0692	1.0784	CHEVRON	...	OIL	DECLINING PRODUCTION	RIFT	5642	PALEOGENE	THINLY-BEDDED SANDSTONE	300.0	270.0	
ALBION-SCIPIO	TRENTON-BLACK RIVER	USA	NORTH AMERICA	MICHIGAN	NaN	41.9937	-84.6016	NUMEROUS	...	OIL	NEARLY DEPLETED	INTRACRATONIC	3800	ORDOVICIAN	DOLOMITE	600.0	400.0	
ALIBEKMOLA	KT I	KAZAKHSTAN	FORMER SOVIET UNION	CASPIAN NORTH	COMPRESSION	48.4740	57.6667	KAZAKHOIL AKTOBE	...	OIL	DEVELOPING	SUB-SALT	6000	CARBONIFEROUS	LIMESTONE	300.0	105.0	
ALIBEKMOLA	KT II	KAZAKHSTAN	FORMER SOVIET UNION	CASPIAN NORTH	COMPRESSION	48.4740	57.6667	KAZAKHOIL AKTOBE	...	OIL	DEVELOPING	SUB-SALT	9580	CARBONIFEROUS	LIMESTONE	607.0	108.0	
ALPINE	ALPINE	USA	NORTH AMERICA	NORTH SLOPE	COMPRESSION	70.3266	-150.8920	CONOCOPHILLIPS	...	OIL	PLATEAU PRODUCTION	FORELAND	6600	JURASSIC	SANDSTONE	50.0	49.0	
ALTAMONT-BLUEBELL	GREEN RIVER AND COLTON/WASATCH	USA	NORTH AMERICA	UINTA	COMPRESSION	40.3000	-110.2100	NUMEROUS	...	OIL	MATURE PRODUCTION	FORELAND	15250	PALEOGENE	SANDSTONE	8000.0	575.0	
ALWYN NORTH	BRENT (BRENT EAST)	UK	EUROPE	NORTH SEA NORTHERN	INVERSION	60.7833	1.7333	TOTAL	...	OIL	NEARLY DEPLETED	RIFT	9790	JURASSIC	SANDSTONE	886.0	344.0	
ALWYN NORTH	STATFJORD	UK	EUROPE	NORTH SEA NORTHERN	EXTENSION	60.7833	1.7333	TOTAL	...	GAS-CONDENSATE	MATURE PRODUCTION	RIFT	10545	TRIASSIC	SANDSTONE	869.0	512.0	
ANASAZI	PARADOX (DESERT CREEK)	USA	NORTH AMERICA	PARADOX	COMPRESSION	37.0789	-109.2333	RIM ENERGY	...	OIL	DECLINING PRODUCTION	INTRACRATONIC	5575	CARBONIFEROUS	DOLOMITE	80.0	46.0	

1 columns

Fig. 1 – source dataset.

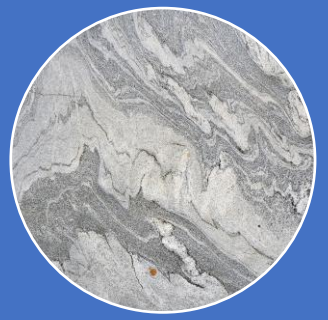
Groups within data

Reservoirs characteristics

- Depth
- Thickness (gross/net pay)
- Reservoir status
- Reservoir period
- Operator company
- Reservoir unit
- Hydrocarbon type
- Field name

Geographic characteristics

- Country, region
- Latitude, Longitude
- Onshore/Offshore
- Basin name



Rock's characteristics

- Lithology
- Porosity
- Permeability
- Tectonic regime
- Structural setting

Unique features

- N

Type of features

Categorical	Numerical
• Field name • Reservoir unit • Country • Region • Basin name • Tectonic regime • Operator company • Onshore or offshore • Hydrocarbon type (main) • Reservoir status (current) • Structural setting • Reservoir period • Lithology (main)	• N • Depth (top reservoir ft TVD) • Thickness (gross average ft) • Thickness (net pay average ft) • Porosity (matrix average %) • Permeability (air average mD) • Latitude • Longitude

Geospatial visualization of the dataset

Fig. 2 – visualization of geospatial data.

Distributions by target features

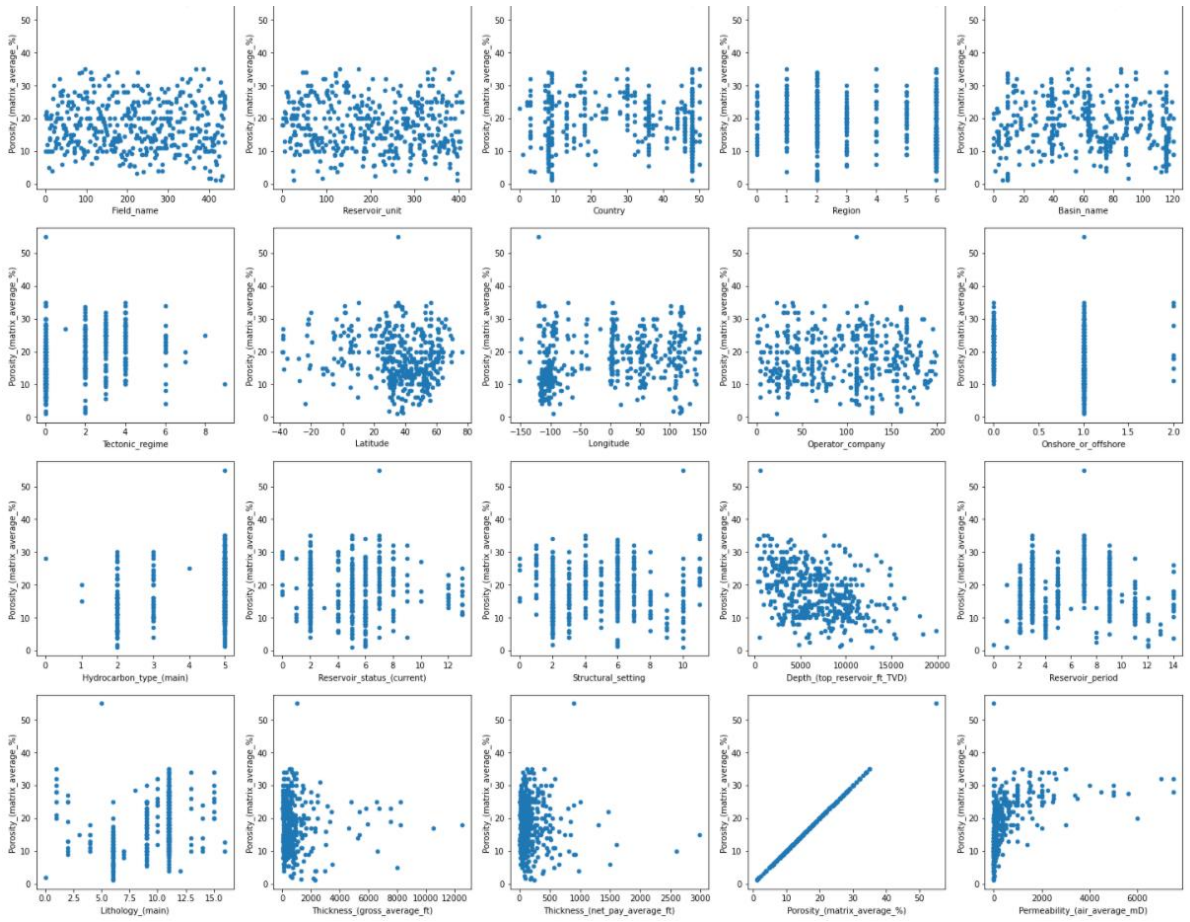


Fig. 3 – pairwise scatterplots with porosity feature.

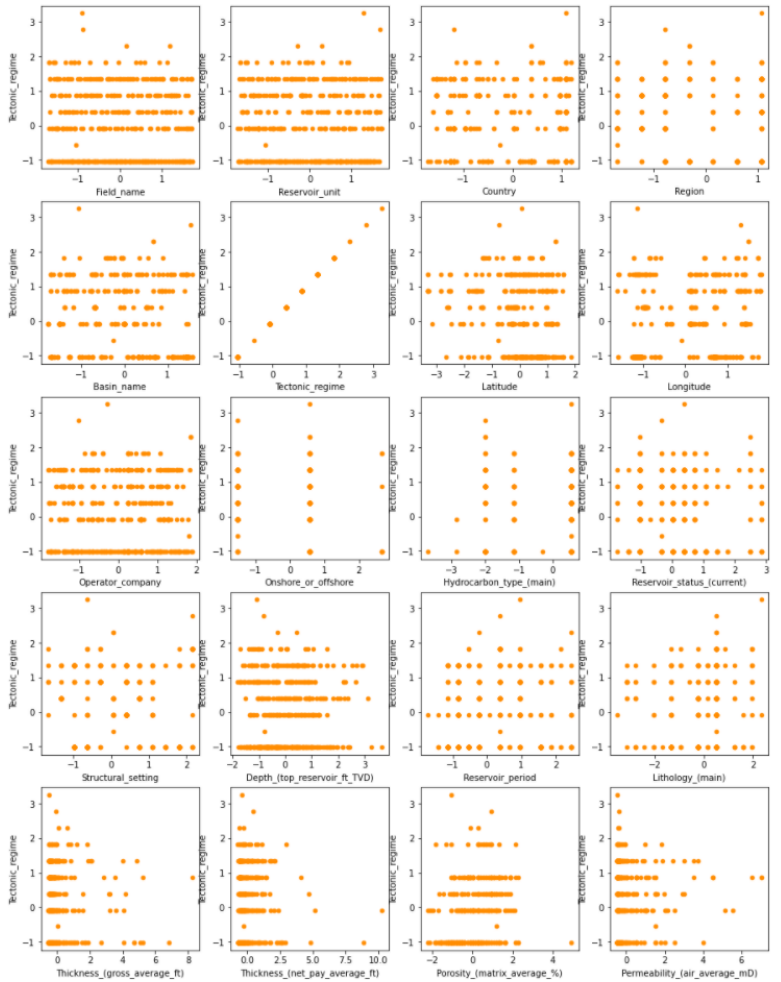


Fig. 4 – pairwise scatterplots with tectonic regime feature.

Try to find a correlation

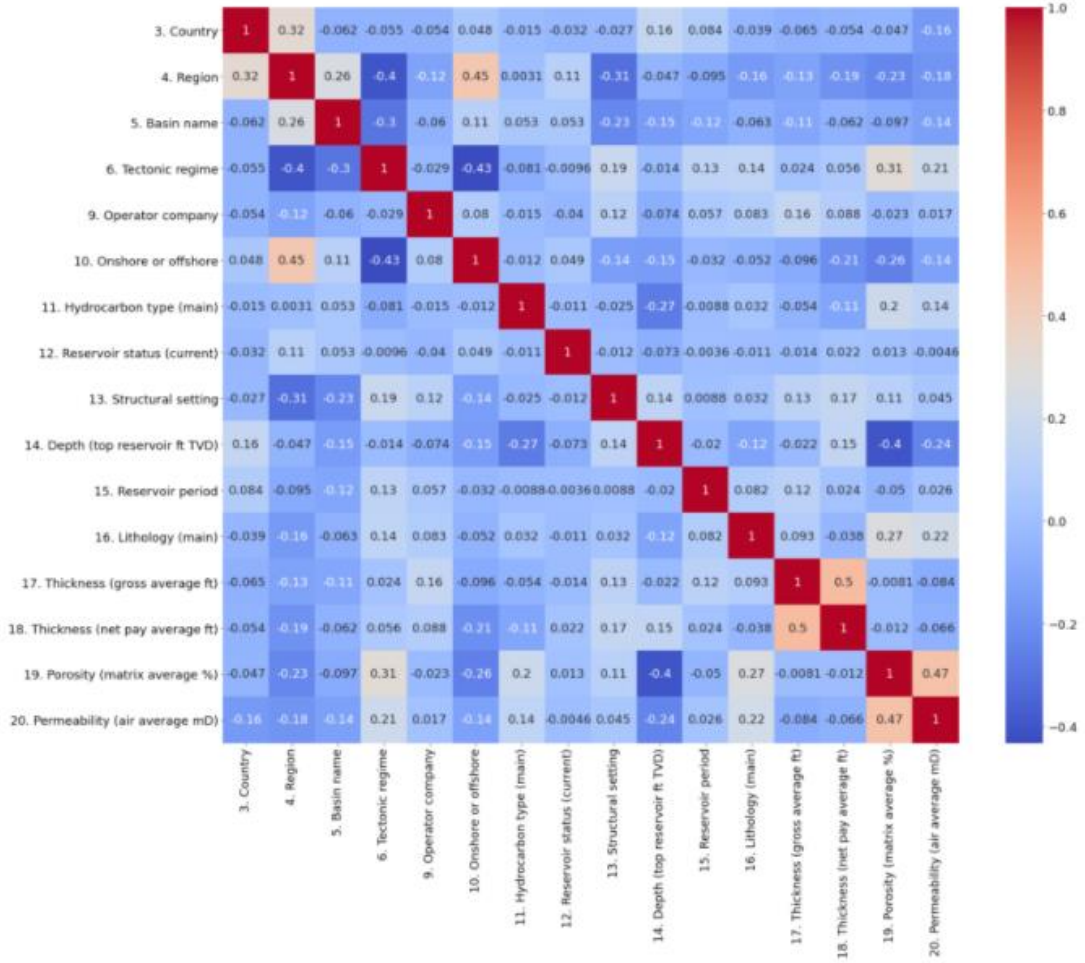


Fig. 5 – heatmap of features correlation.

Outliers in numerical features

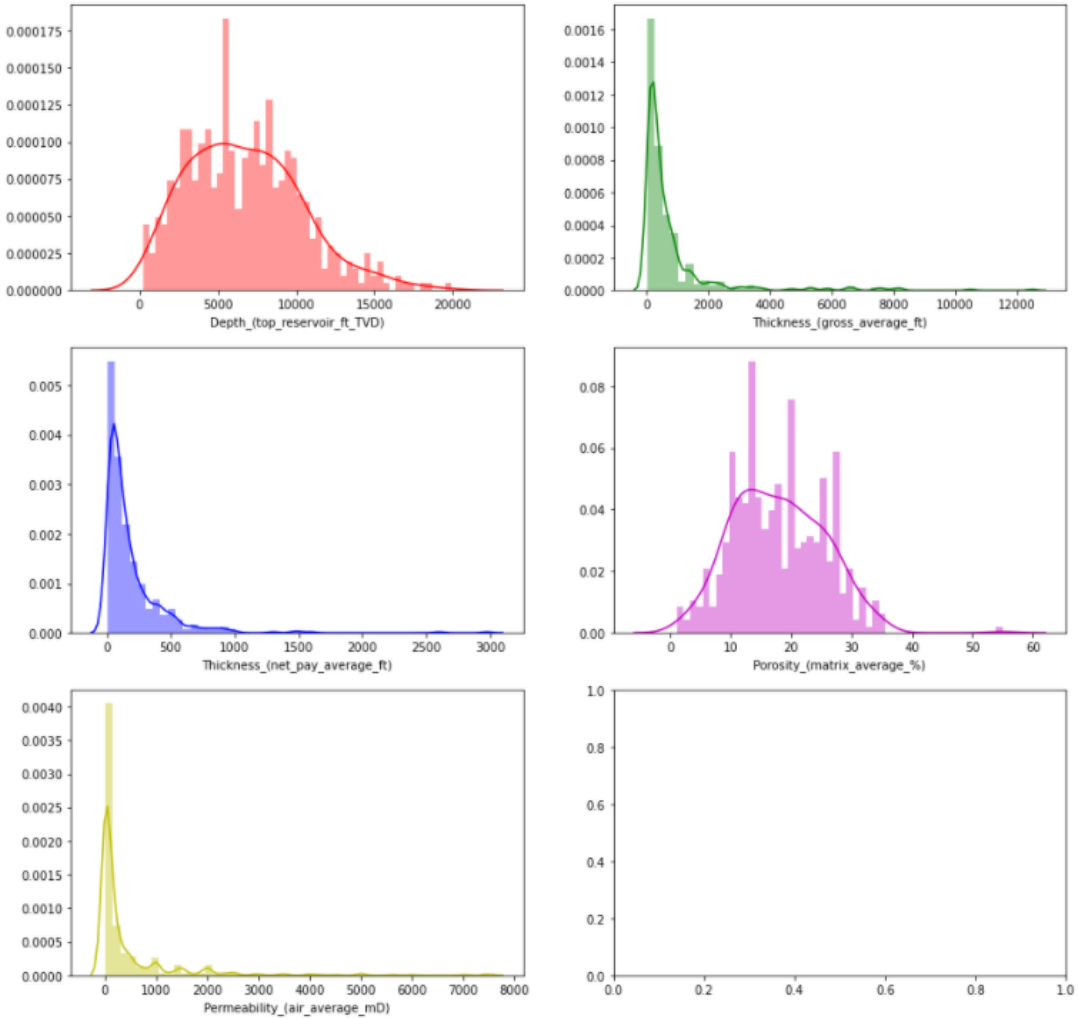


Fig. 6 – distribution plots of numerical features.

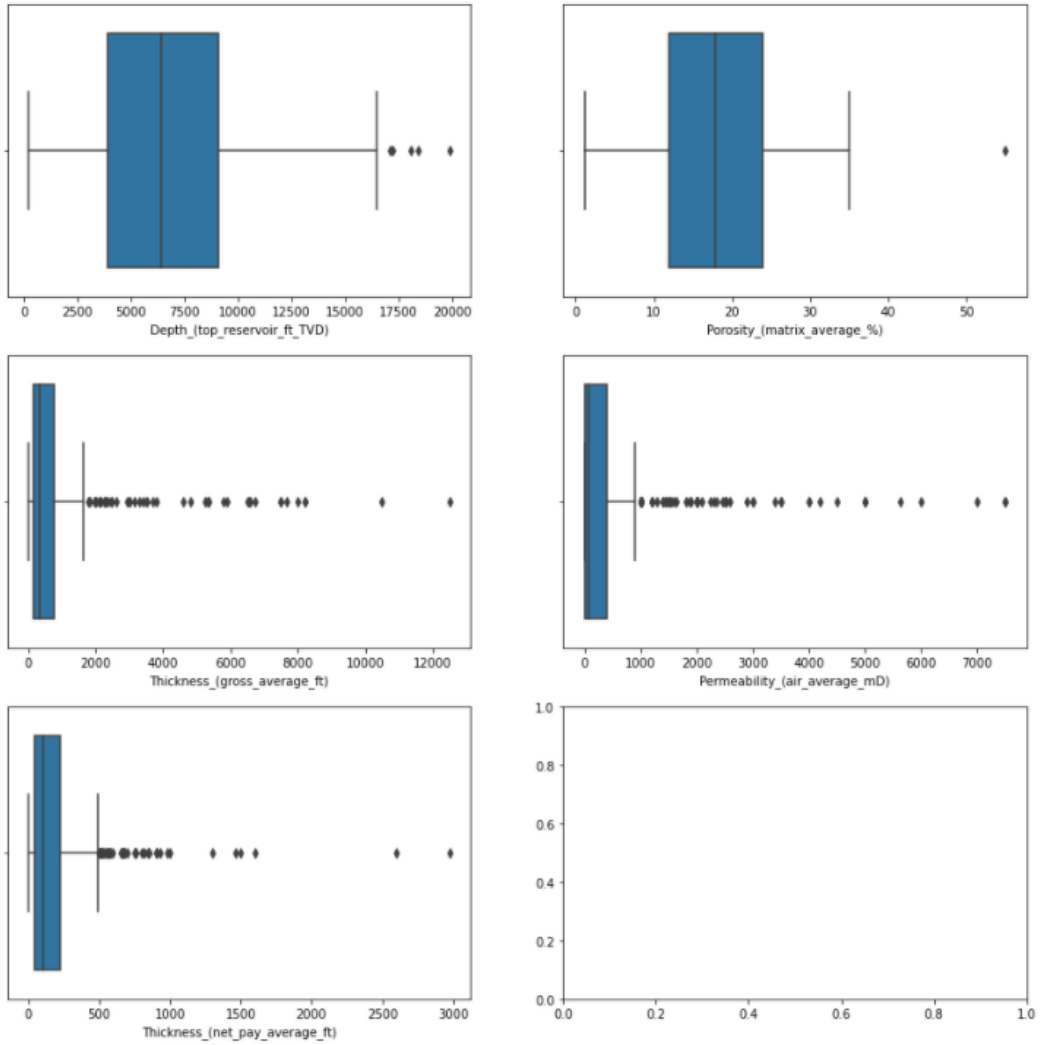


Fig. 7 – boxplots of numerical features.

PCA with coloring by tectonic regime

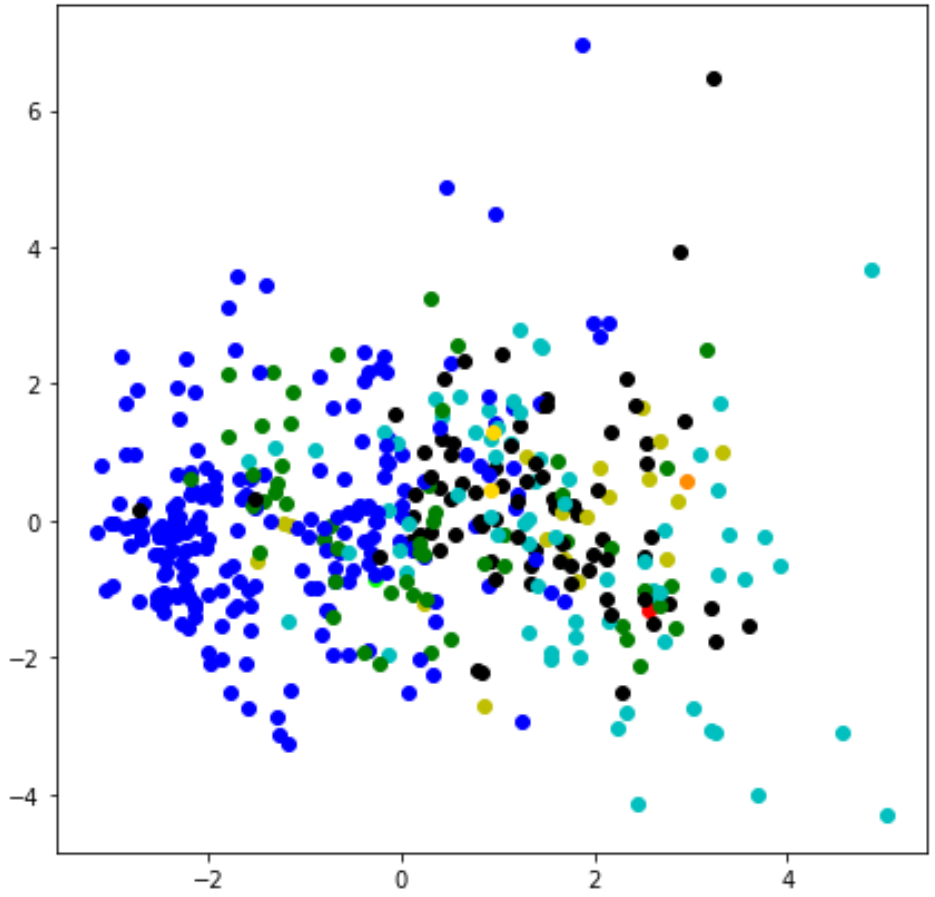


Fig. 8 – 2D PCA with coloring by tectonic regime.

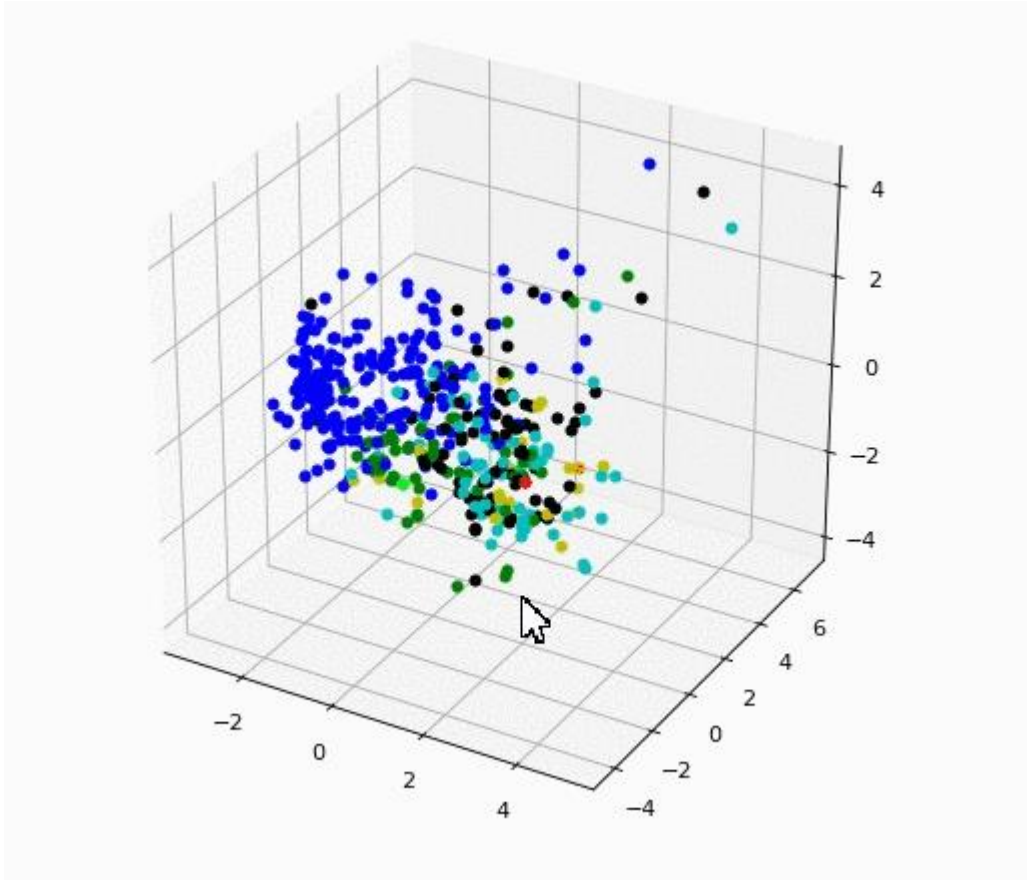
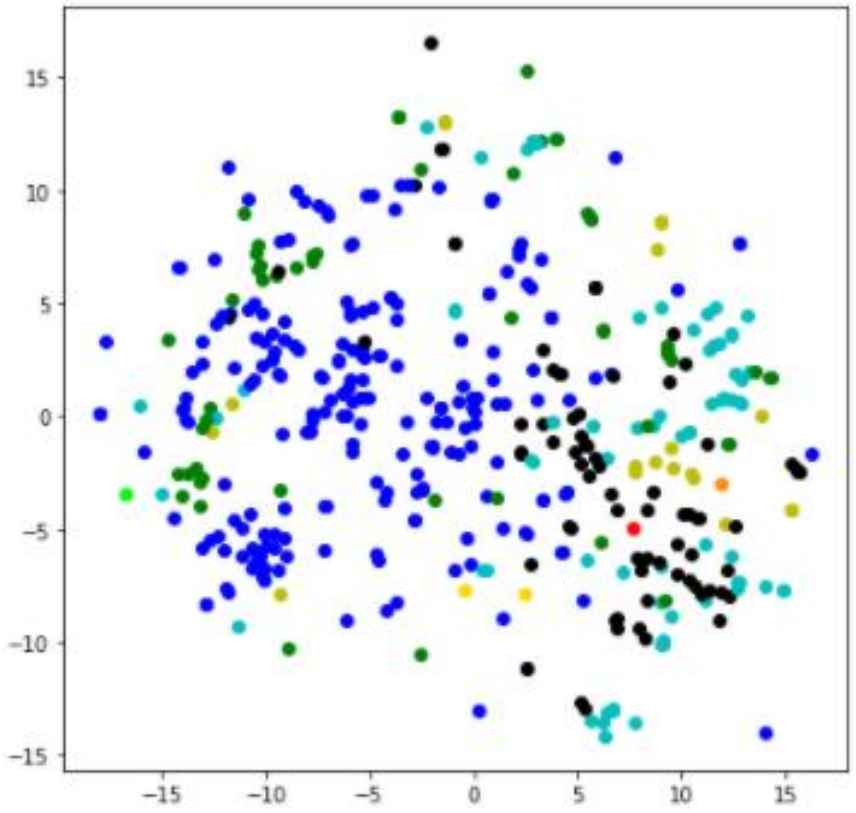
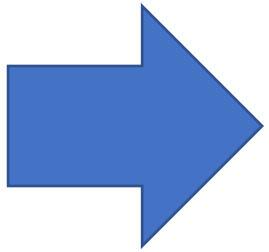


Fig. 9 – 3D PCA with coloring by tectonic regime.

t-SNE with coloring by tectonic regime





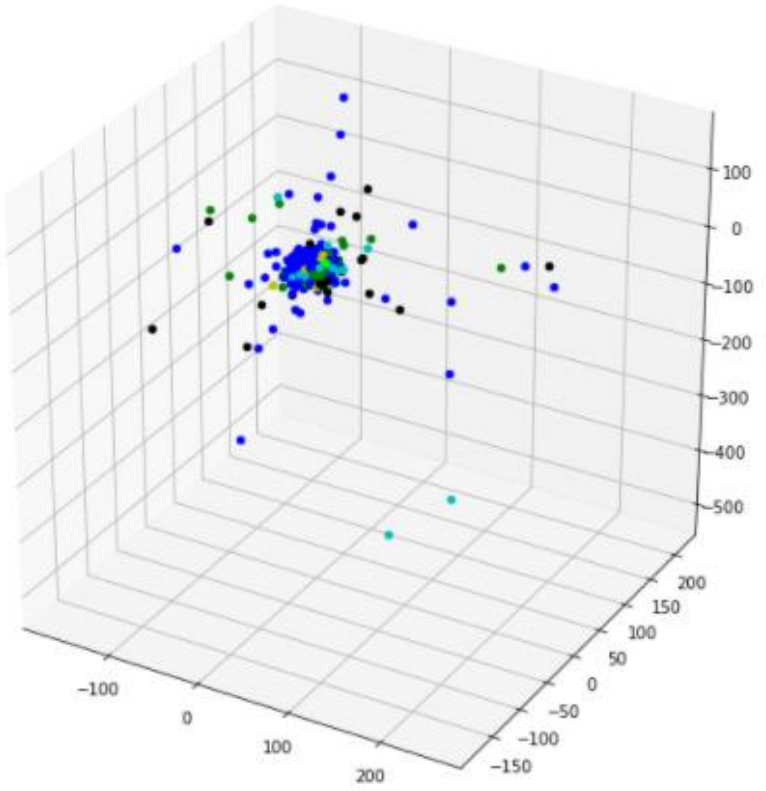
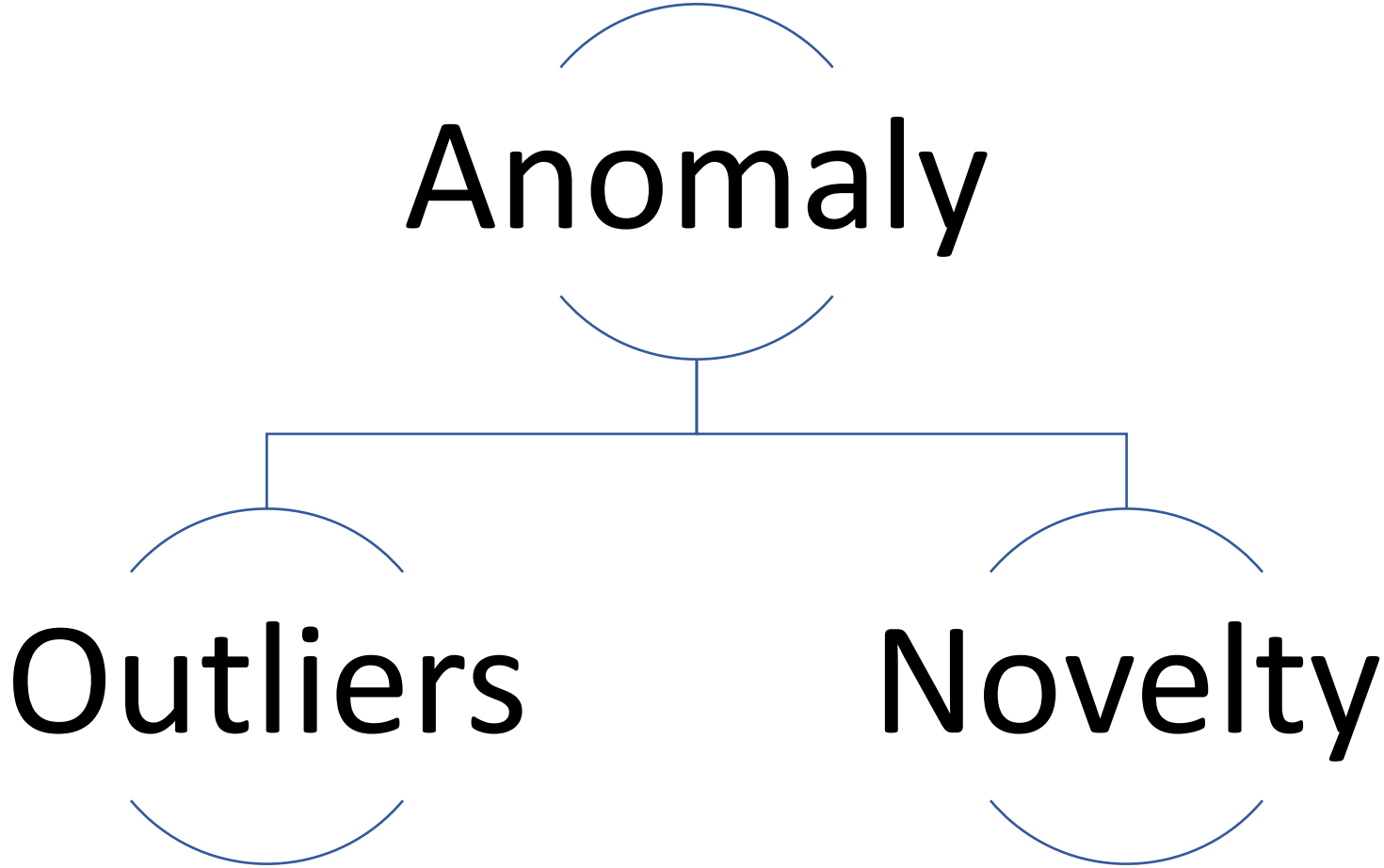


Fig. 10 – 2D t-SNE with coloring by tectonic regime.

Fig. 11 – 3D t-SNE with coloring by tectonic regime.

More on t-SNE and PCA comparison: https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne

Anomaly detection


Anomaly
Outliers
Novelty

One-class SVM

- SVM attempts to identify the hyperplane that best represents the largest separation, or margin, between border-line data points (i.e., ‘‘support vectors”).
- One-class SVM creates a finds the maximal gap hyperplane that separates data from the origin.
- The resulting closed boundary is found using the following formulation [2]

$$min_{w,\xi}\left(\frac{1}{2}\|w\|^2+\frac{1}{\upsilon n}\sum_{i=1}^{n}\xi_i+b\right)$$

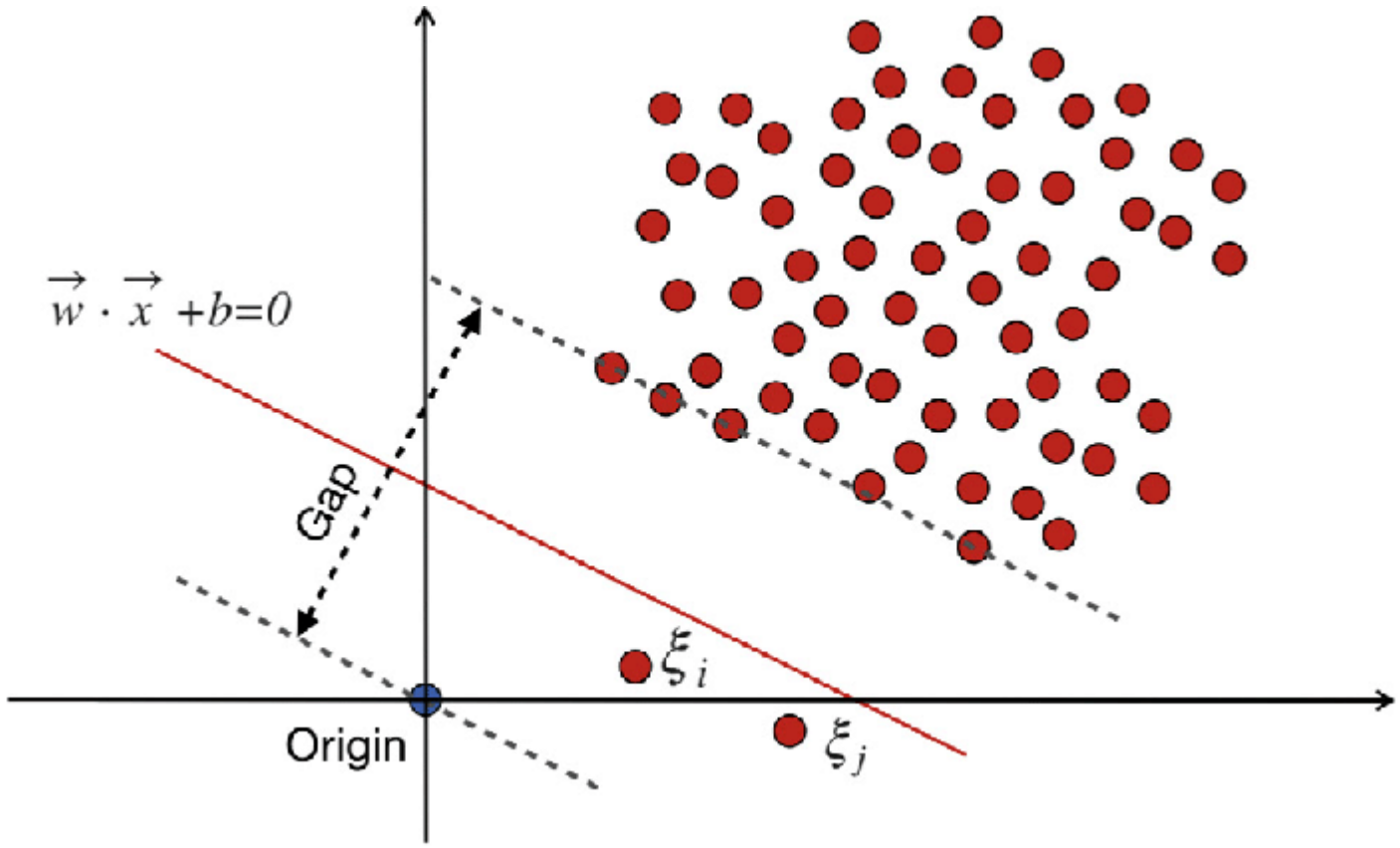


Fig. 12 – overview of one-class Support Vector Machine (SVM).

[2] Bonsang Koo, Byungjin Shin “Applying novelty detection to identify model element to IFC class misclassifications on architectural and infrastructure Building Information Models”
More on one-class SVM hyperparameter tuning: http://activisiongamescience.github.io/2015/12/23/Unsupervised-Anomaly-Detection-SOD-vs-One-class-SVM

Applying SVM to existing dataset

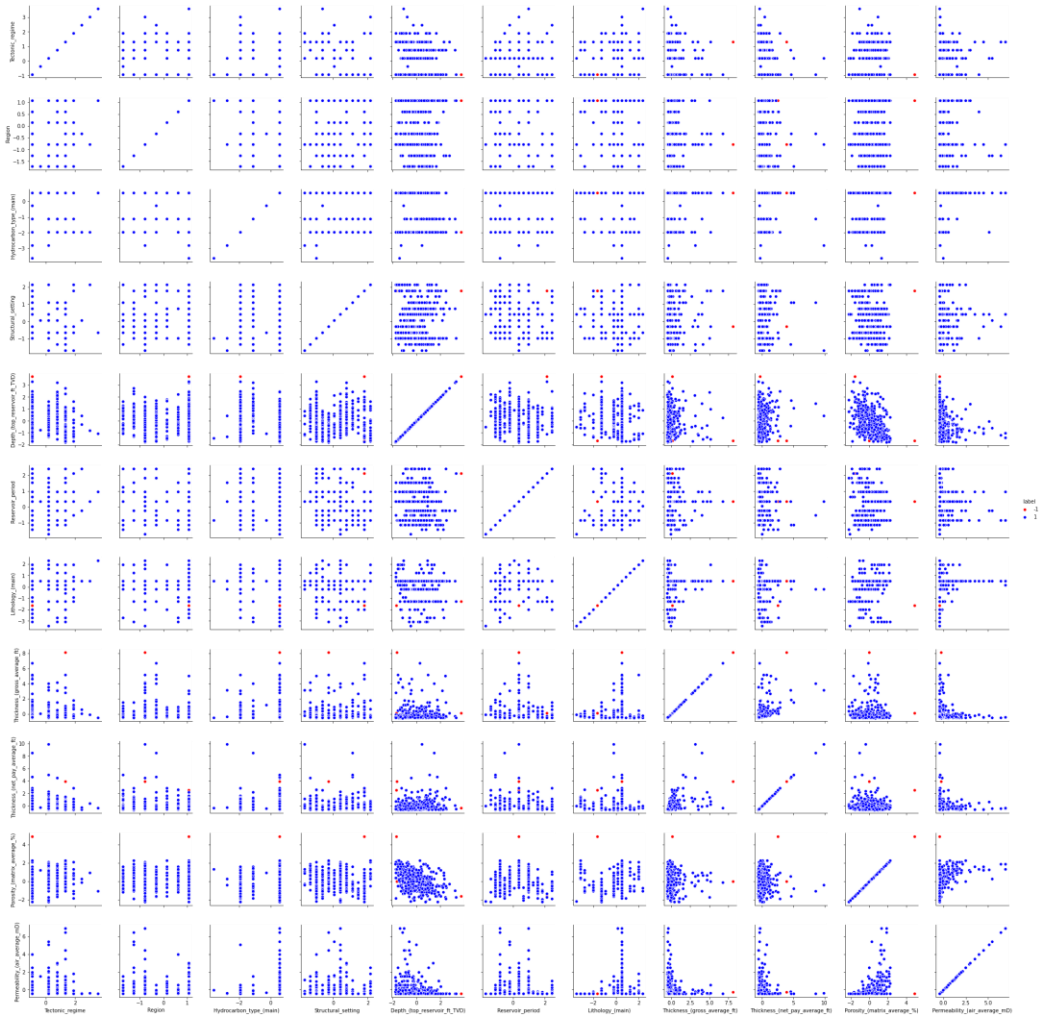


Fig. 13 – pairwise scatterplots with colored anomalies.

- Selected columns to fit the model:
  - Tectonic regime
  - Region
  - Hydrocarbon type (main)
  - Structural setting
  - Depth (top reservoir ft TVD)
  - Reservoir period
  - Lithology (main)
  - Thickness (gross average ft),
  - Thickness (net pay average ft)
  - Porosity (matrix average %)
  - Permeability (air average mD)

- We can plot 11x11 scatterplot in order to visualize anomalies within pairwise distributions.

Applying SVM to existing dataset

- Some plots don’t contain anomaly points.
- Very often relatively distant points are considered as anomalies.

- Valid samples percent = 99.32…
- Anomalies percent = 0.67…

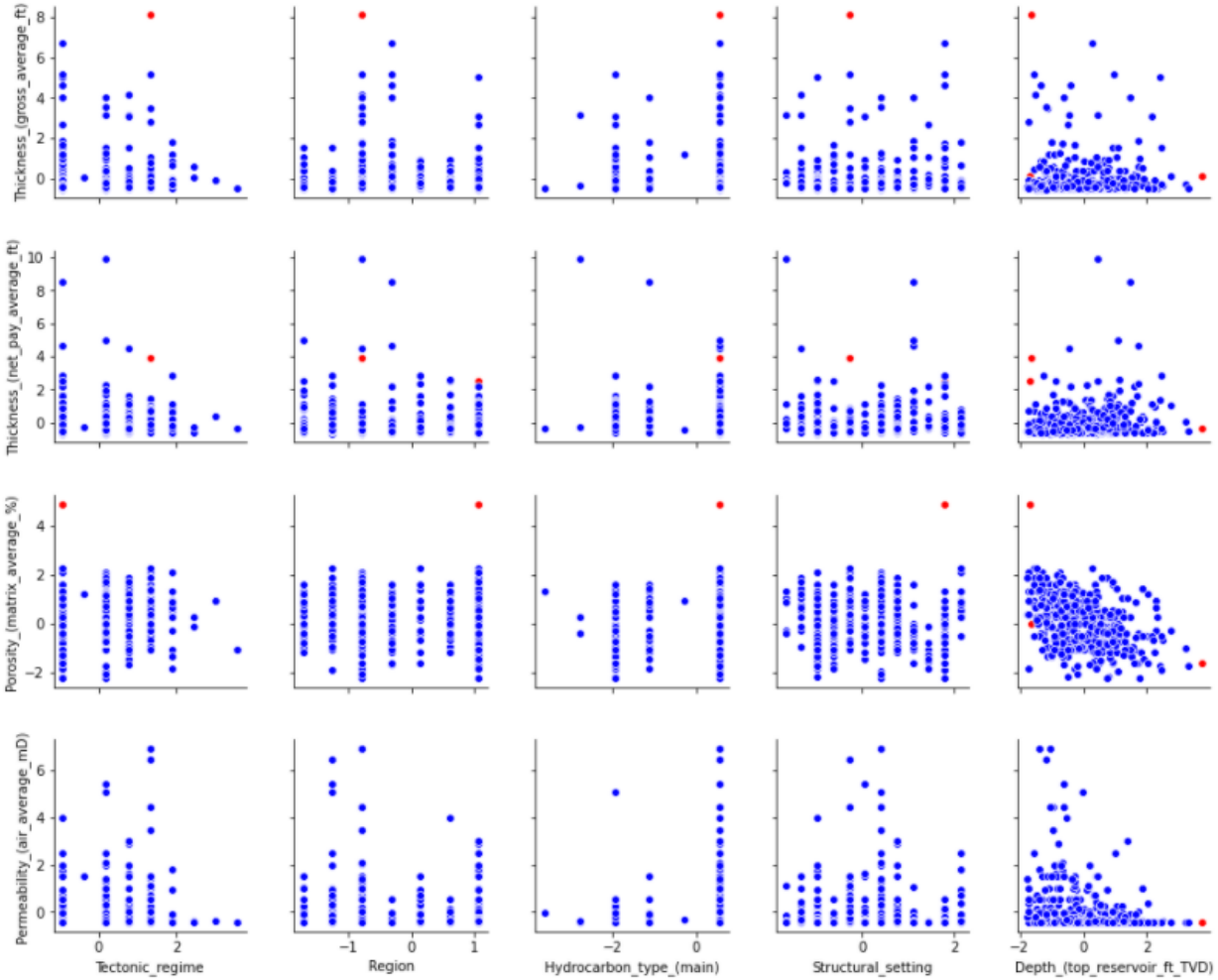


Fig. 14 – particular pairwise scatterplots with colored anomalies.

Manual data with anomalies

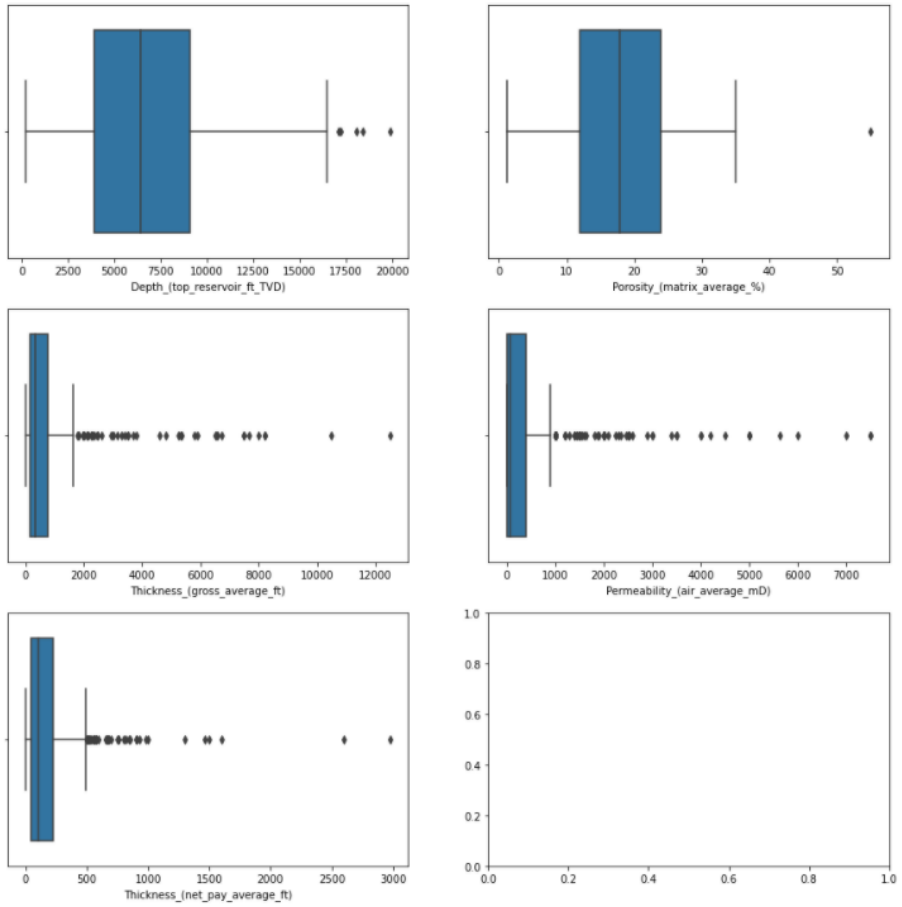


Fig. 15 – boxplots of numerical features.

- Take N valid objects from source data sets.
- Randomly change numeric features to extreme values.

Manual data with anomalies

- All generated data were considered anomalies.
- Of course these samples were simple and maybe obvious for the model.

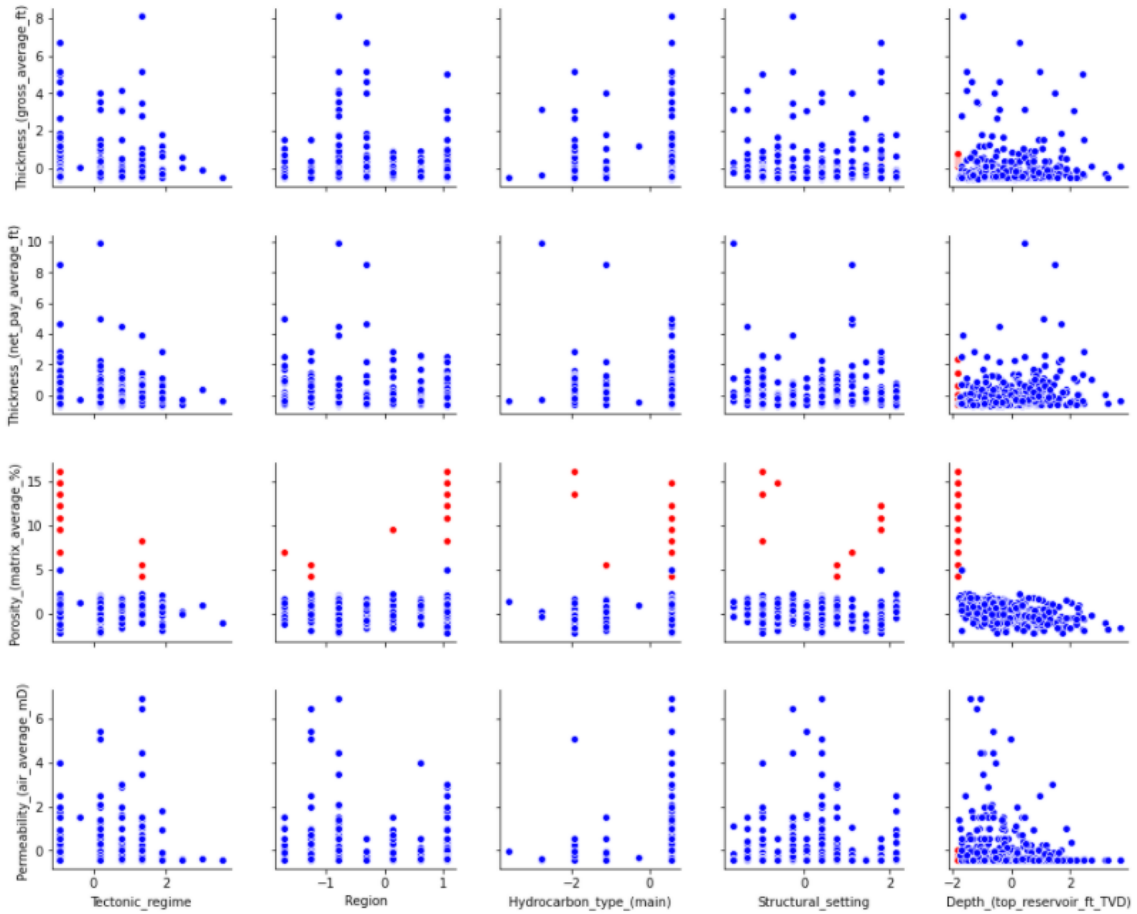


Fig. 16 – particular pairwise scatterplots with manually added anomalies.

Data generation

Gaussian mixture model

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [3].

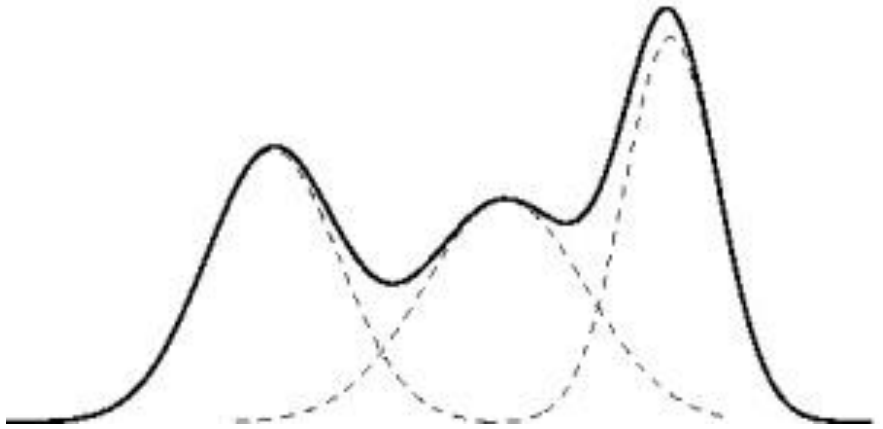


Fig. 17 – Gaussian mixture model.

.

[3] “Gaussian Mixture Models” https://prateekvjoshi.com/2013/06/29/gaussian-mixture-models/

Bayesian networks model

A Bayesian network (BN) is a compact representation of a probability distribution over a set of discrete variables. The structure of a BN is a directed acyclic graph (DAG), where the arcs have a formal interpretation in terms of probabilistic conditional independence [4].

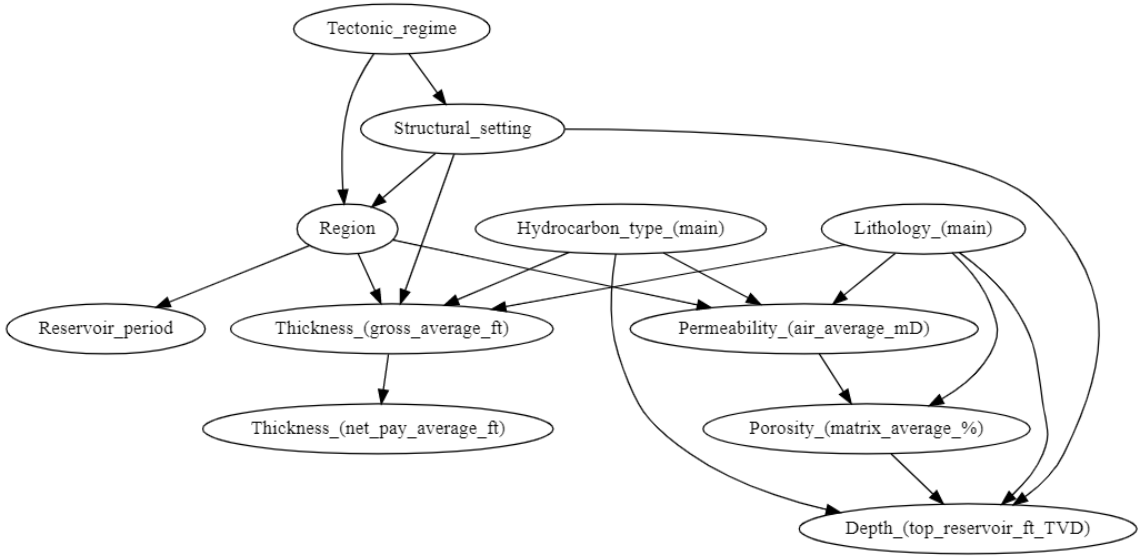


Fig. 18 – Example of a network's DAG with k2 score.

.

[4] Concha Bielza, Pedro Larrañaga (2014) "Bayesian networks in neuroscience: a survey"

Discretization strategy

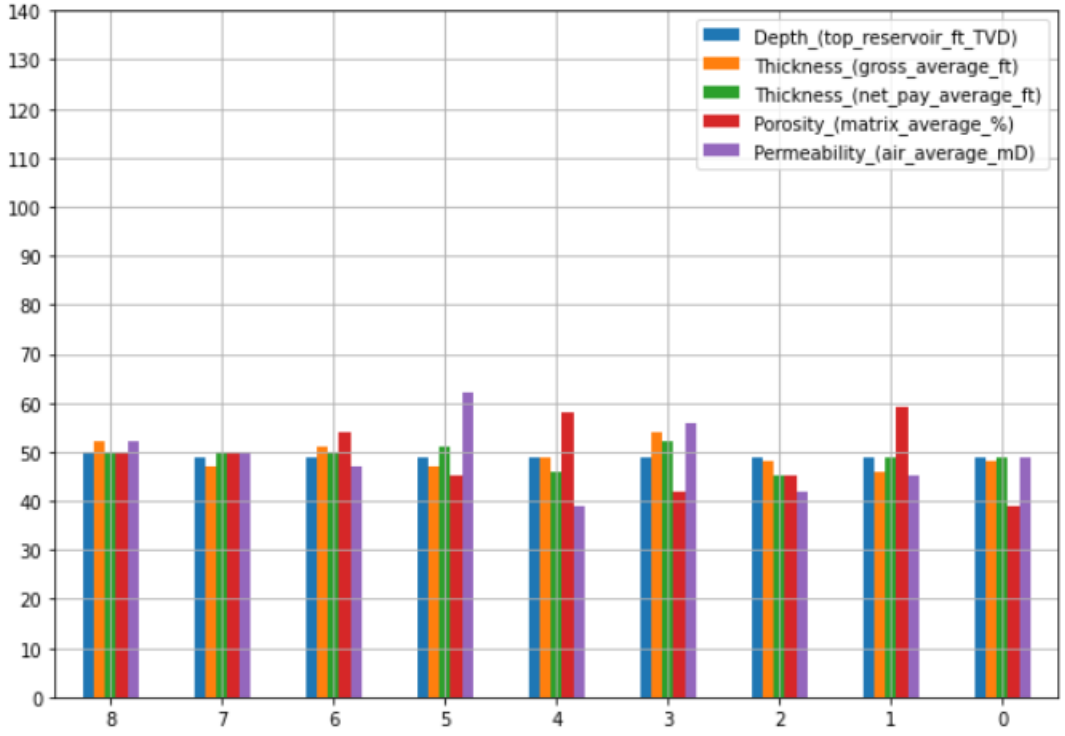


Fig. 19 – quantile strategy.

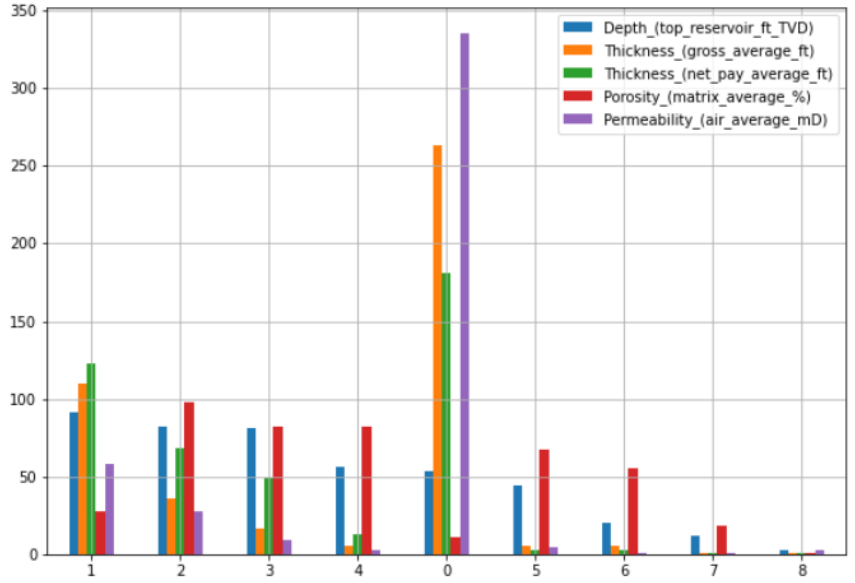


Fig. 20 – kmeans strategy.

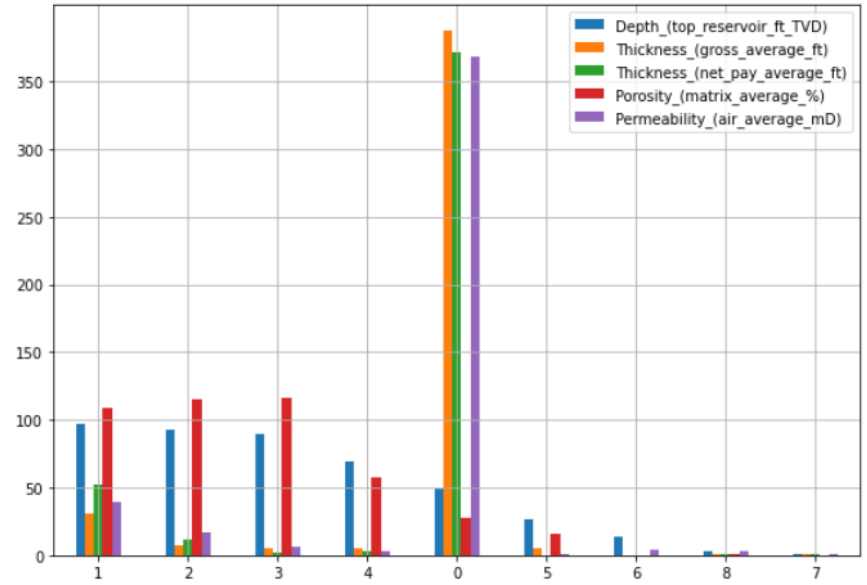


Fig. 21 – uniform strategy.

n_bins in KBinsDiscretizer

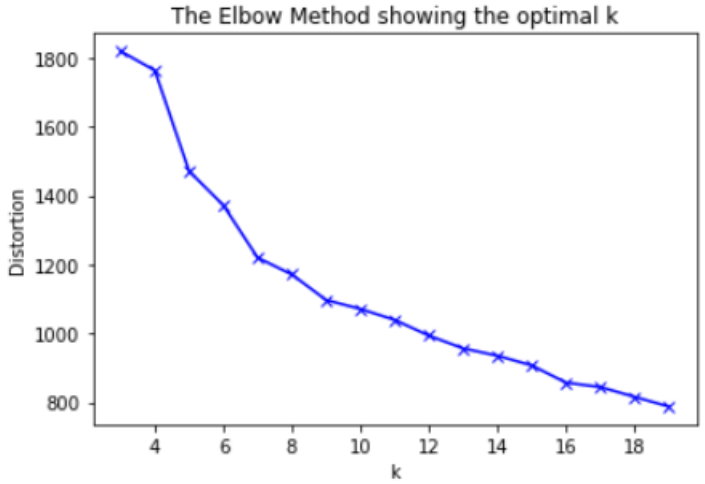


Fig. 22 – k-means for all columns.

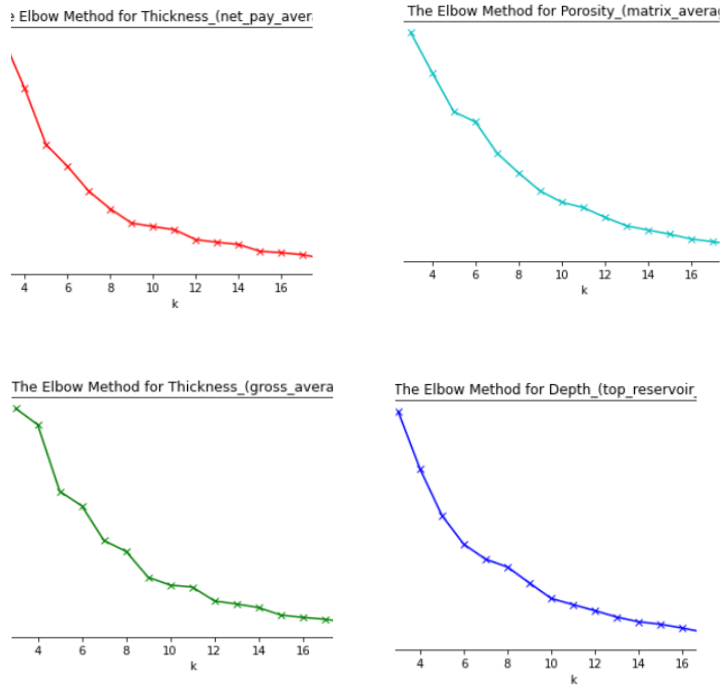


Fig. 23 – k-means for each column.

Different BN scores

There are 2 most popular scoring functions:

- K2
- BDeu
  - To obtain the BDeu score, we need a parameter called equivalent sample size α that expresses the strength of our prior belief in the uniformity of the conditional distributions of the network. A quick look at the Bayesian network learning literature reveals that there is no generally accepted “uninformative” value for the α parameter [5].

[5] Tomi Silander, Petri Kontkanen, Petri Myllymaki (2007) “On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter”

BDeu hyperparameter search

Equivalent sample size significantly (ESS) impacts on the DAG structure

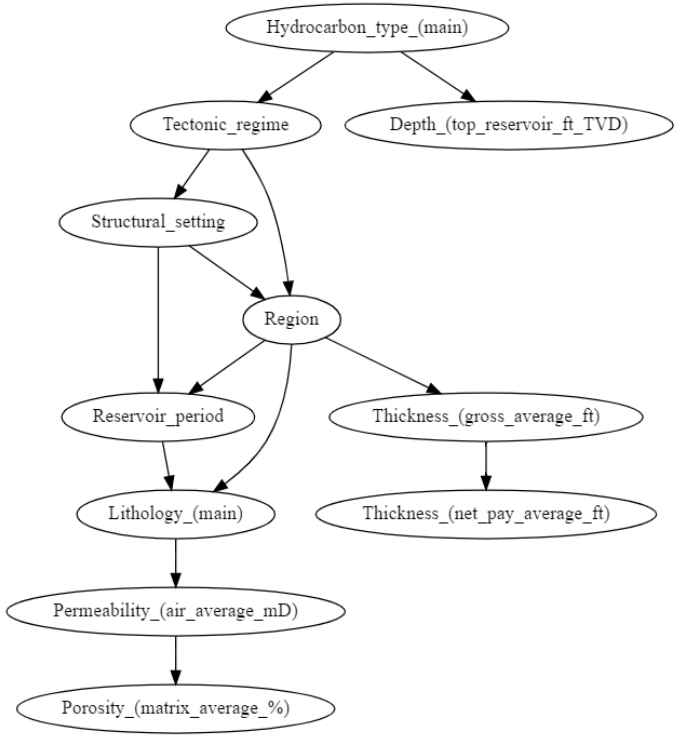


Fig. 24 – DAG for bn with BDeu score and ESS = 78.

Fig. 25 – DAG for bn with BDeu score and ESS = 10.

We performed exhaustive search in range [20, 100]

Comparison of BN graphs

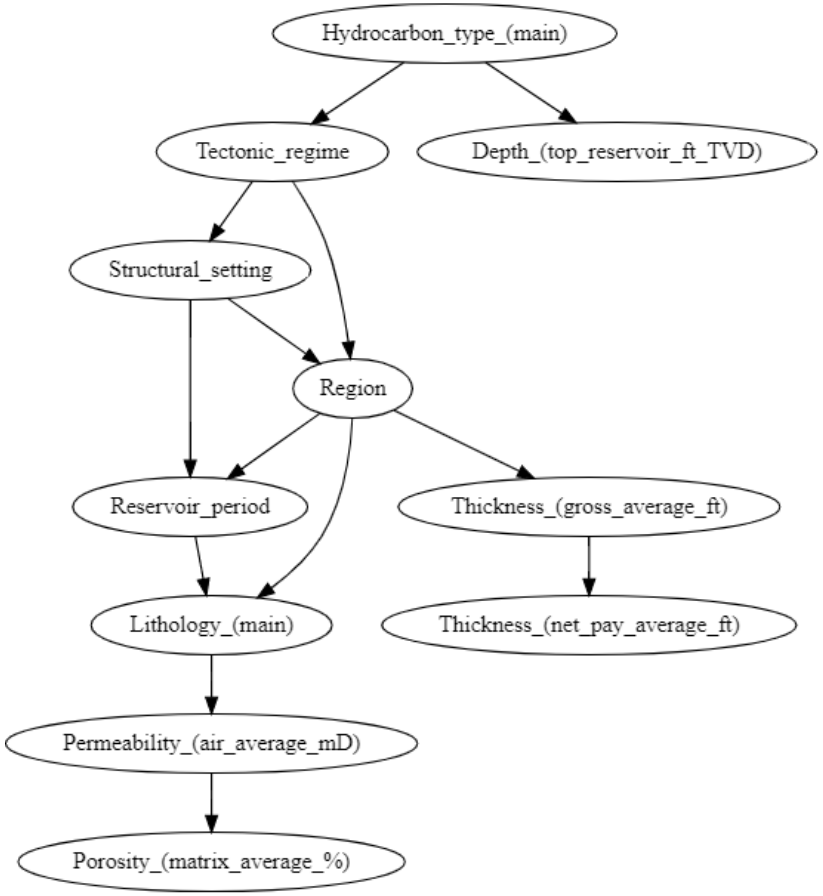


Fig. 26 – DAG for bn with BDeu score and single discretizer.

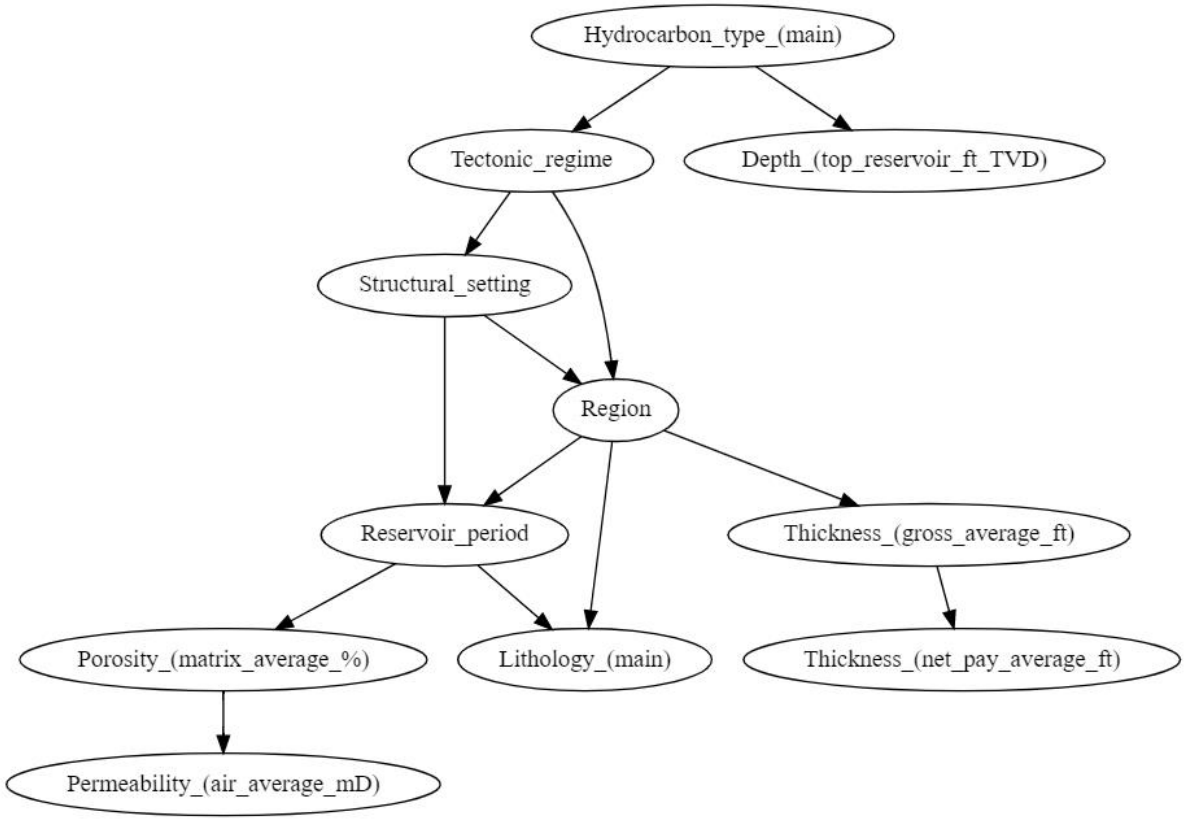


Fig. 27 – DAG for bn with BDeu score and several discretizers.

Comparison of BN models performances

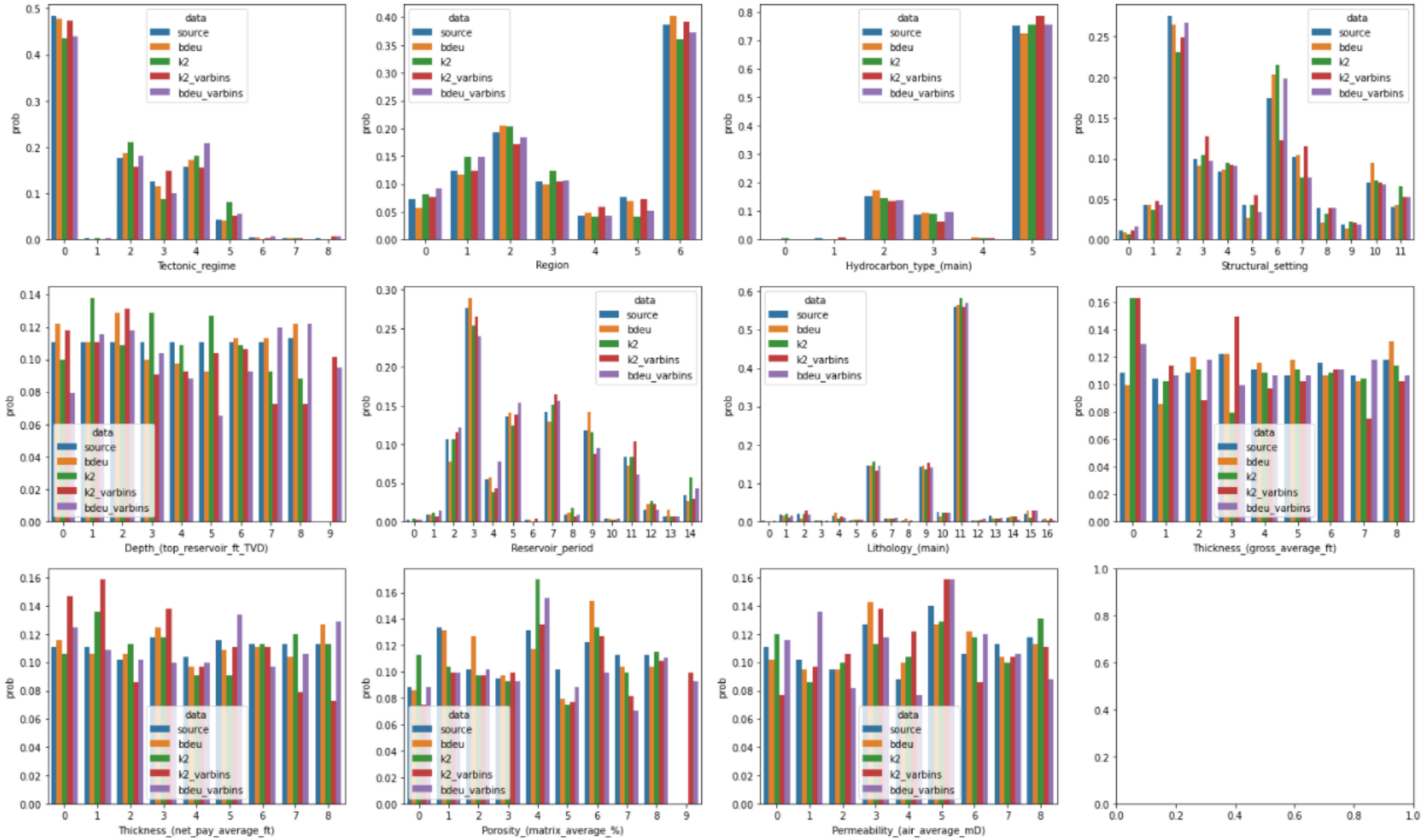


Fig. 28 – Histograms of one-feature distributions.

SRMSE

- SRMSE metric allows to measure error of joint distributions.

$$SRMSE = \sqrt{\sum_{m_1=1}^{M_1} \dots \sum_{m_n=1}^{M_n} \frac{(f_{m_1 \dots m_n} - \hat{f}_{m_1 \dots m_n})^2}{M_1 \times \dots \times M_n}}$$

Selected sets:

  - The set of categorical features
  - The set of numerical features
  - The set of correlated features (porosity and permeability)

- As a result the model with BDeu score has a little better metric value, then the model with k2 score.

Data sampling pipeline

	Hydrocarbon_type_(main)	Depth_(top_reservoir_ft_TVD)	Tectonic_regime	Structural_setting	Region	Thickness_(gross_average_ft)	Thickness_(net_pay_average_ft)	Reservoir_period	Lithology_(main)	Permeability_(air_average_mD)	Porosity_(matrix_average_%)
0	OIL	4090.0	COMPRESSION	FORELAND	MIDDLE EAST	677.0	101.50	JURASSIC	LIMESTONE	161.0	22.75
1	GAS-CONDENSATE	2960.5	EXTENSION	RIFT	AFRICA	1152.0	1700.50	CRETACEOUS	SANDSTONE	361.0	20.00
2	GAS	8763.5	STRIKE-SLIP	WRENCH	AFRICA	160.0	11.50	NEOGENE	THINLY-BEDDED SANDSTONE	850.0	17.50
3	GAS	8763.5	STRIKE-SLIP	WRENCH	AFRICA	245.0	29.75	NEOGENE	SANDSTONE	4350.0	41.50
4	OIL	2960.5	COMPRESSION	SALT	MIDDLE EAST	7000.0	71.50	CRETACEOUS	LIMESTONE	35.0	26.25

Fig. 29 – samples from the model with BDeu score.

	Lithology_(main)	Hydrocarbon_type_(main)	Tectonic_regime	Structural_setting	Region	Permeability_(air_average_mD)	Porosity_(matrix_average_%)	Depth_(top_reservoir_ft_TVD)	Thickness_(gross_average_ft)	Thickness_(net_pay_average_ft)	Reservoir_period
0	LIMESTONE	OIL	COMPRESSION	FORELAND	MIDDLE EAST	14.000	17.50	2960.5	245.0	205.00	JURASSIC
1	SANDSTONE	GAS-CONDENSATE	EXTENSION	RIFT	AFRICA	1.305	5.05	8763.5	46.5	29.75	CRETACEOUS
2	THINLY-BEDDED SANDSTONE	GAS	STRIKE-SLIP	WRENCH	AFRICA	361.000	17.50	10170.0	7000.0	71.50	NEOGENE
3	SANDSTONE	GAS	STRIKE-SLIP	WRENCH	AFRICA	5.300	17.50	8763.5	677.0	101.50	NEOGENE
4	LIMESTONE	OIL	COMPRESSION	SALT	MIDDLE EAST	161.000	20.00	5153.5	46.5	48.25	CRETACEOUS

Fig. 29 – samples from the model with k2 score.

Applying SVM to generated samples

10000 generated samples

BDeu valid samples percentage = 99.79

k2 valid samples percentage = 99.6

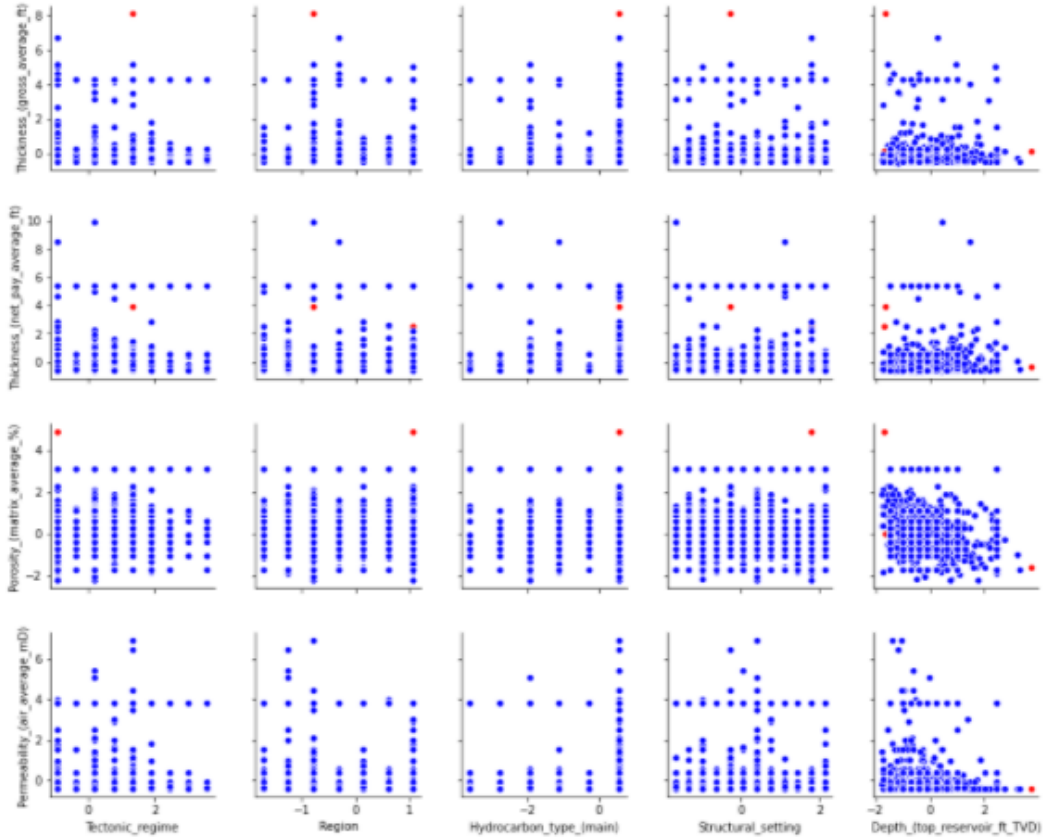


Fig. 30 – particular pairwise scatterplots with anomalies in sampled data.

Filling missing values

Prediction models

For categorical missing values	For numerical missing values
RandomForestClassifier	RandomForestRegressor
GradientBoostingClassifier	GradientBoostingRegressor
ExtraTreesClassifier	ExtraTreesRegressor
AdaBoostClassifier	AdaBoostRegressor
XGBClassifier	XGBRegressor
Bayesian network	Ridge regression

Metrics

Accuracy, F1-score

RMSE

Data and Parameters

For data preparation:

- Label encoding
- Scaler
- Normaliser

For models:

- Hyper-parameters tuning
- Cross-validation
- Important features

Comparison of applied models

- Validation score for categorical value

	RF Classifier	GB Classifier	ExtraTrees Classifier	XGB Classifier	AdaBoost Classifier
Accuracy	0.9101	0.9101	0.9213	0.8989	0.5730
F1	0.9087	0.9113	0.9192	0.8978	0.4497

- Validation RMSE score for numerical value

RF Regressor	GB Regressor	ExtraTrees Regressor	AdaBoost Regressor	XGB Regressor	Ridge
4.8994	5.3328	4.2872	5.1466	5.2331	5.2172

Test score

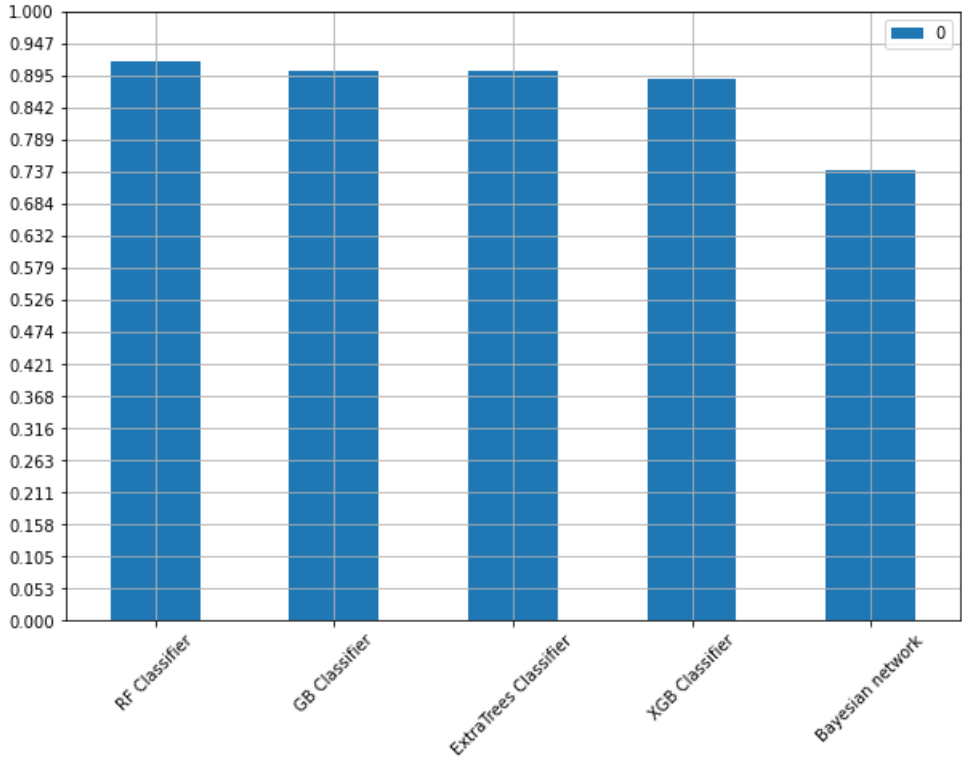


Fig. 31 – test score for the classification task.

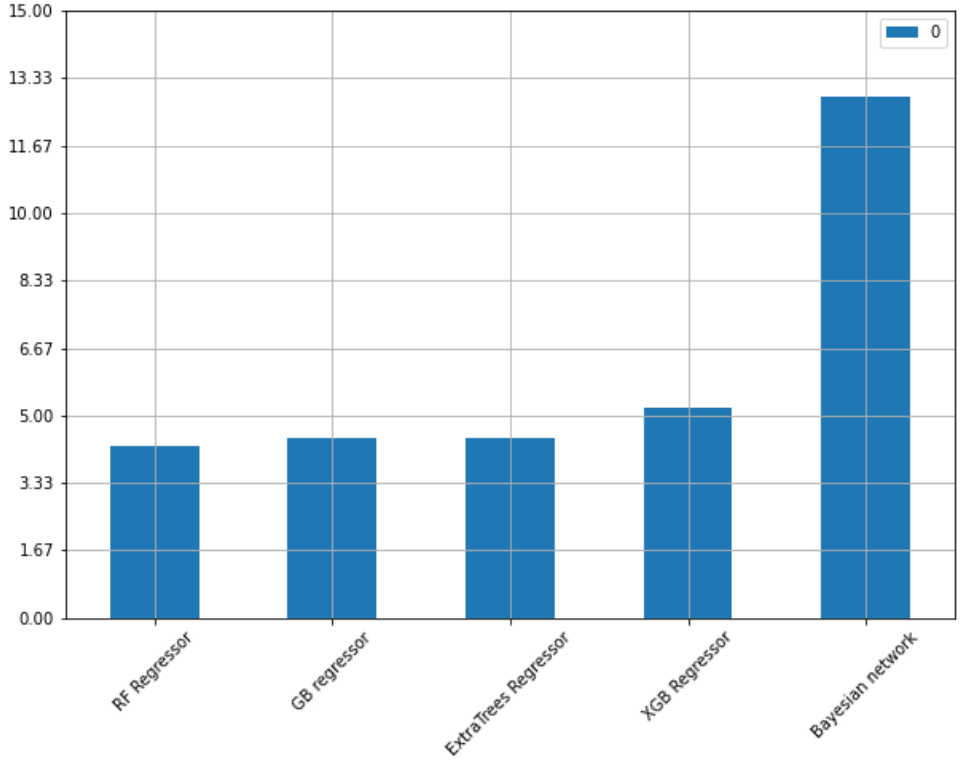


Fig. 32 – test score for the regression task.

Comparison of applied models

- Test score for categorical value

	RF Classifier	GB Classifier	ExtraTrees Classifier	XGB Classifier	Bayesian network
Accuracy	0.9166	0.9030	0.9025	0.8899	0.75

- Test RMSE score for numerical value

RF Regressor	GB Regressor	ExtraTrees Regressor	XGB Regressor	Bayesian network
4.2242	4.4254	4.4409	5.2029	12.8659

Thank you for your attention!