TalTech Virumaa kolledž

Masinõppe meetodid ja prognoosmudelid tehnoloogilistes protsessides

Lektor: Olga Dunajeva, TalTech Virumaa kolledž, vanemlektor, PhD
olga.dunajeva@taltech.ee

02.10.2022

Содержание

- Модель машинного обучения
- Алгоритмы машинного обучения
- Машинное обучение в технологических процессах:
    - обнаружение неисправностей и аномалий
- Создание модели машинного обучения
- Валидирование модели

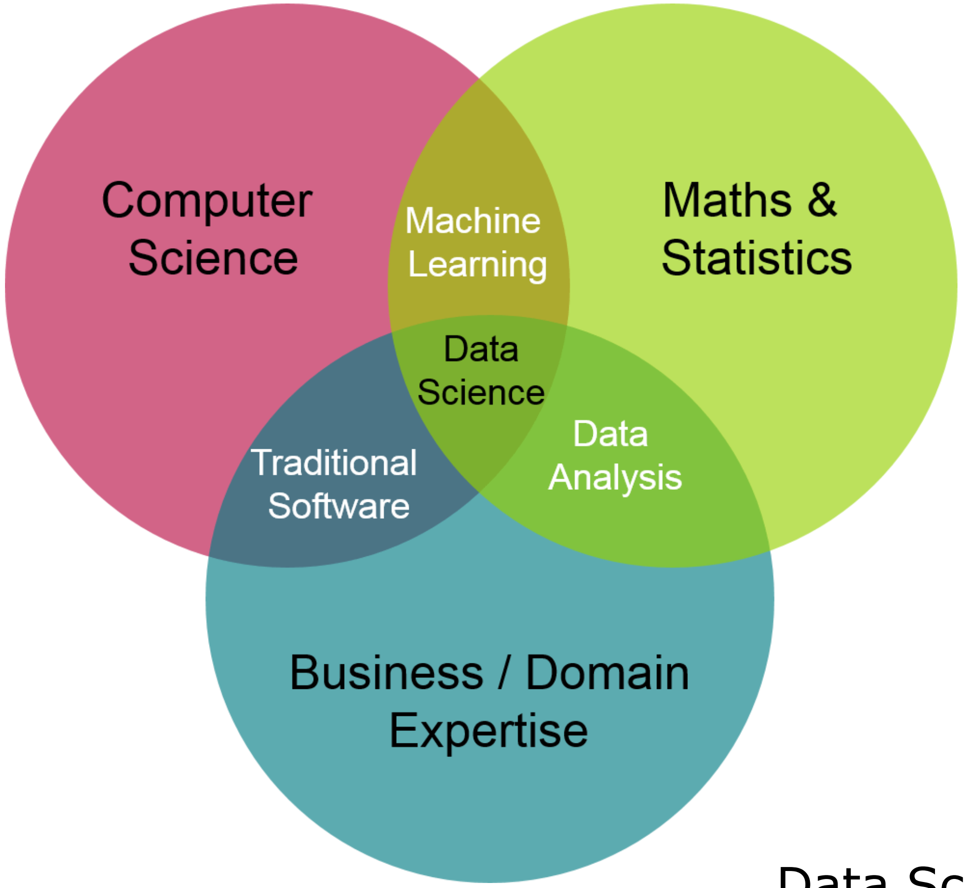
Расписание

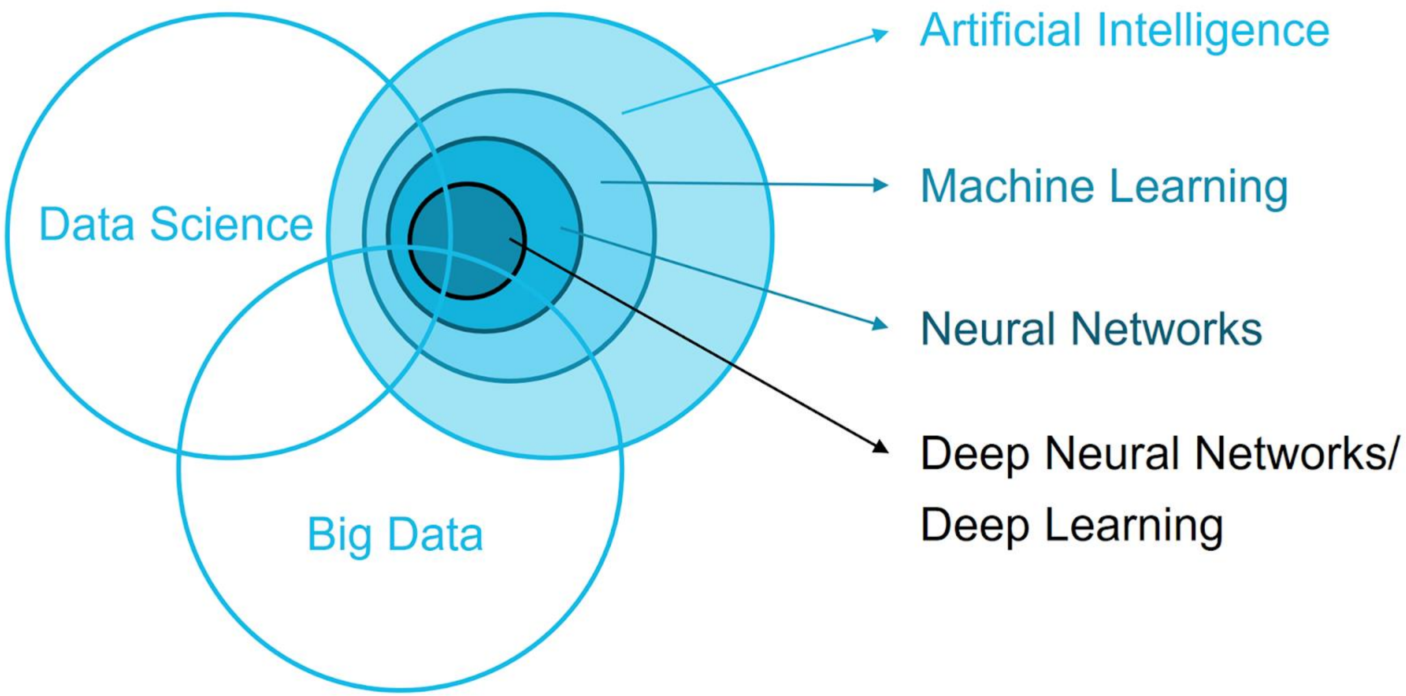
- 13:00 – 16:15 Masinõpe ja prognoosmudelid tehnoloogilistes protsessides , 4 акад. часа

Материалы

- Moodle`i kursus: Tööstusprotsessid ja andmete visualiseerimine
- Registreerimisvõti: **xtRg4y**

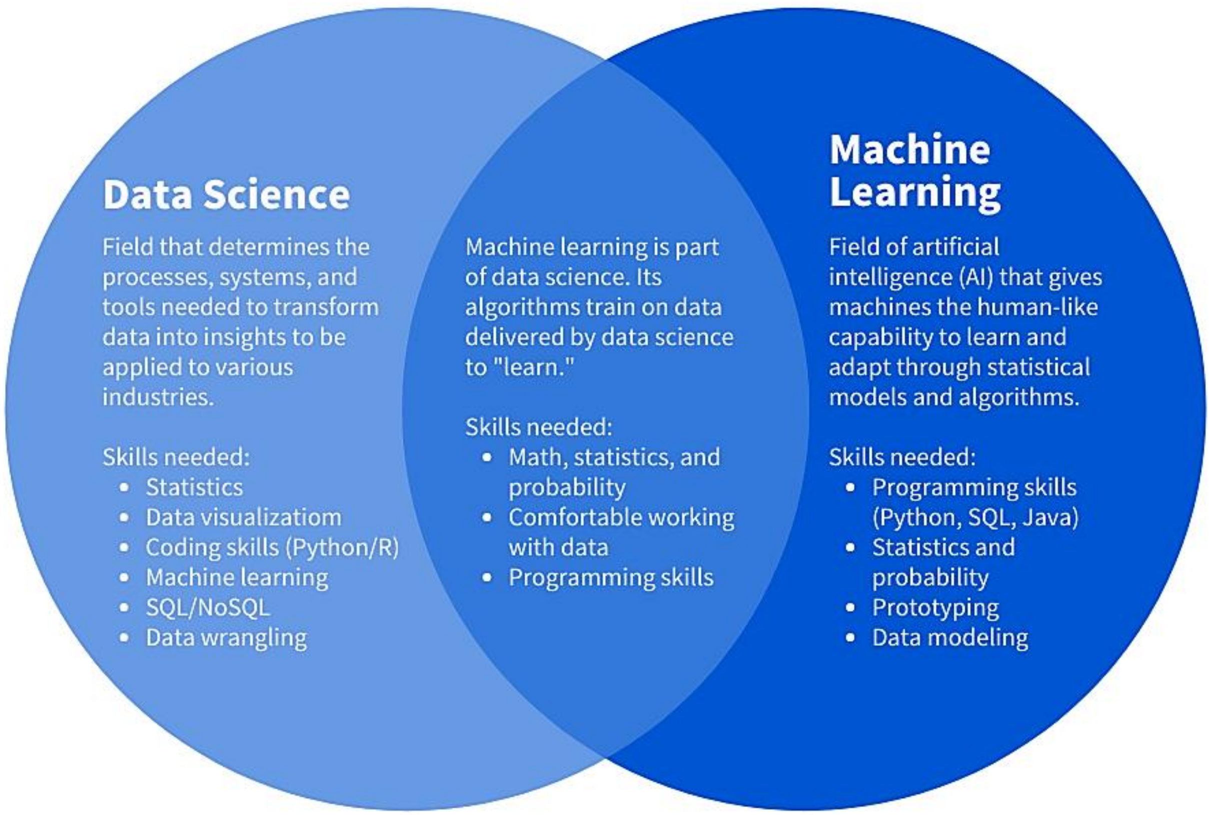
Машинное обучение: навыки и разделы





Data Scientist— лучший статистик среди программистов и лучший программист среди статистиков

Data Science vs Machine Learning

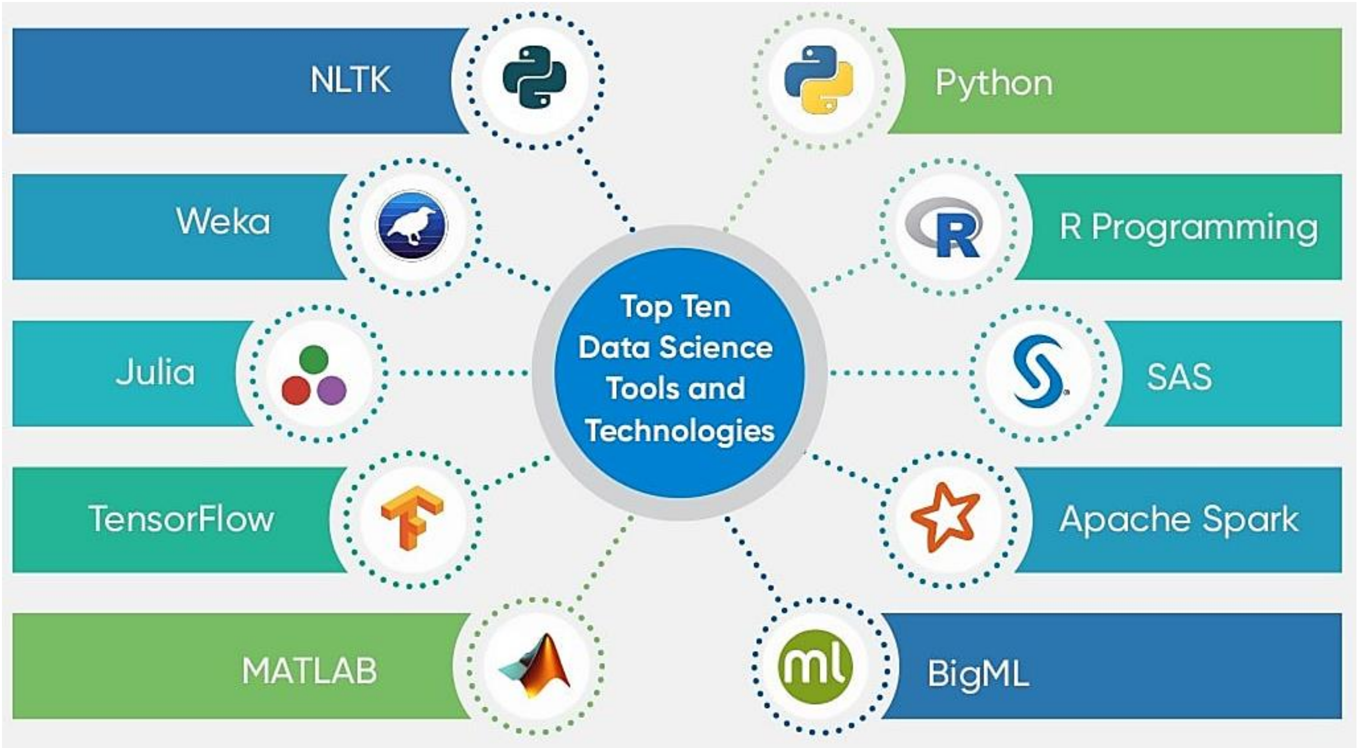


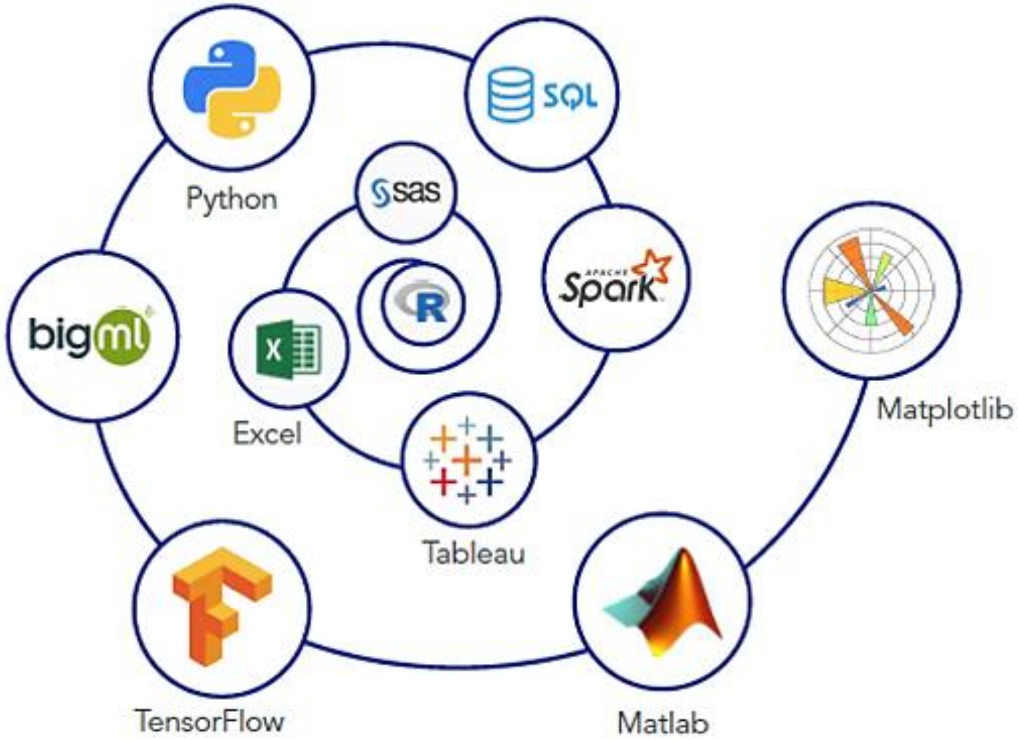
- **Машинное обучение** – наука о разработке алгоритмов и статистических моделей, которые компьютерные системы используют для выполнения задач, полагаясь на шаблоны в данных.
- Выявляя шаблоны в данных, системы могут прогнозировать результаты на основе заданного набора входных данных. Например, обучить медицинское приложение диагностировать по рентгеновским изображениям, сохраняя миллионы отсканированных изображений и соответствующие диагнозы.

Инструменты работы с данными

Power BI
Grafana
tableau
Microsoft Dynamics 365
Google Analytics
Google Data Studio
Shiny by RStudio
python™
Google Big Query
amazon REDSHIFT
Microsoft Azure
MySQL®
teradata.

Top Ten Data Science Tools and Technologies in 2022

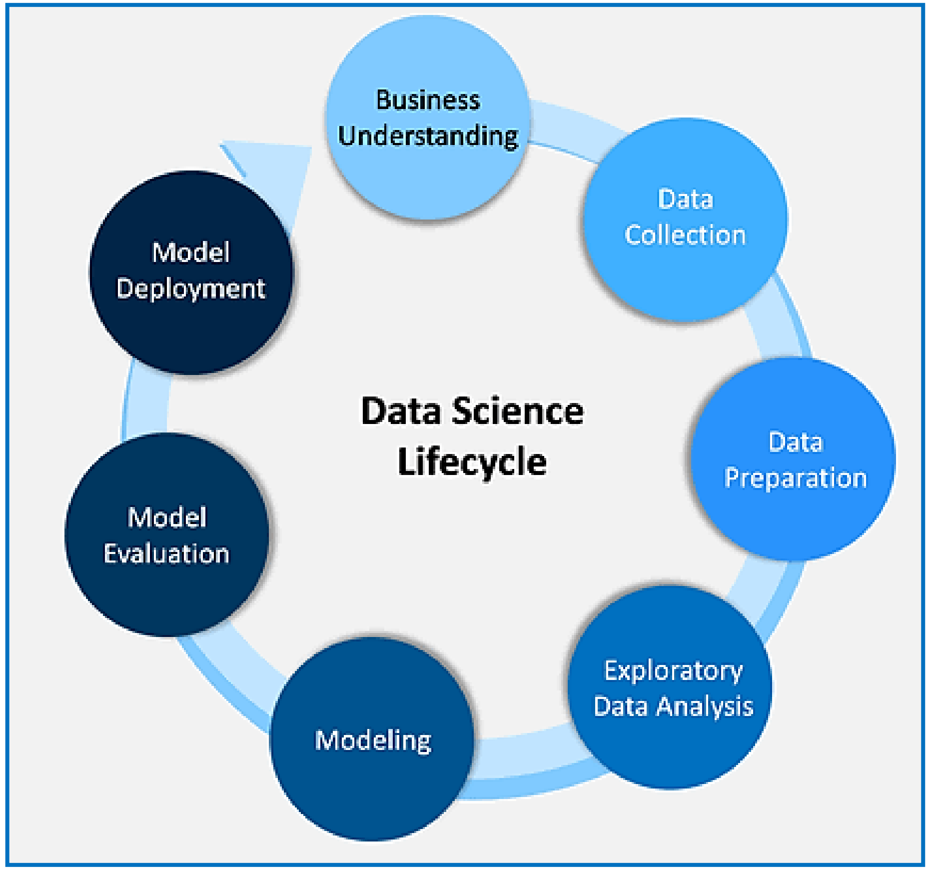




https://www.bacancytechnology.com/blog/data-science-for-business

https://medium.com/javarevisited/10-essential-tools-data-scientists-should-learn-in-2022-acbae6558643

Проект машинного обучения

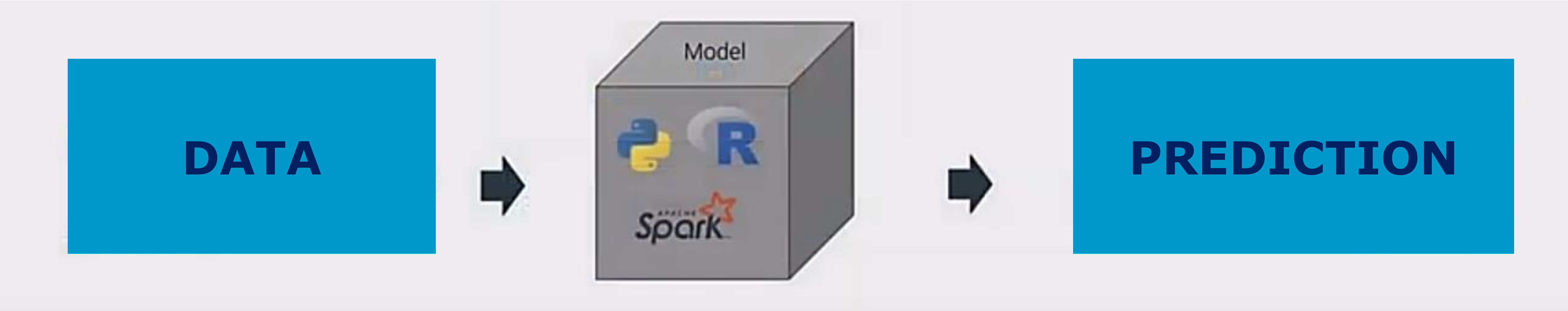


Cross Industry Process for Data Mining,
CRISP-DM

IBM Knowledge Center, CRISP-DM Help Overview

- Постановка задачи
- Сбор данных
- Поготовка данных (отсутствующие значения, ошибки, аномалии, создание/удаление атрибутов)
  Teie mudel on sama hea kui teie andmed
- Предварительный анализ (изучение распределений и связей)
- Конструирование модели (выбор алгоритма, настройка параметров)
- Валидирование модели (тестовые данные, метрики)
- Имплементирование модели, исправление

Модель машинного обучения



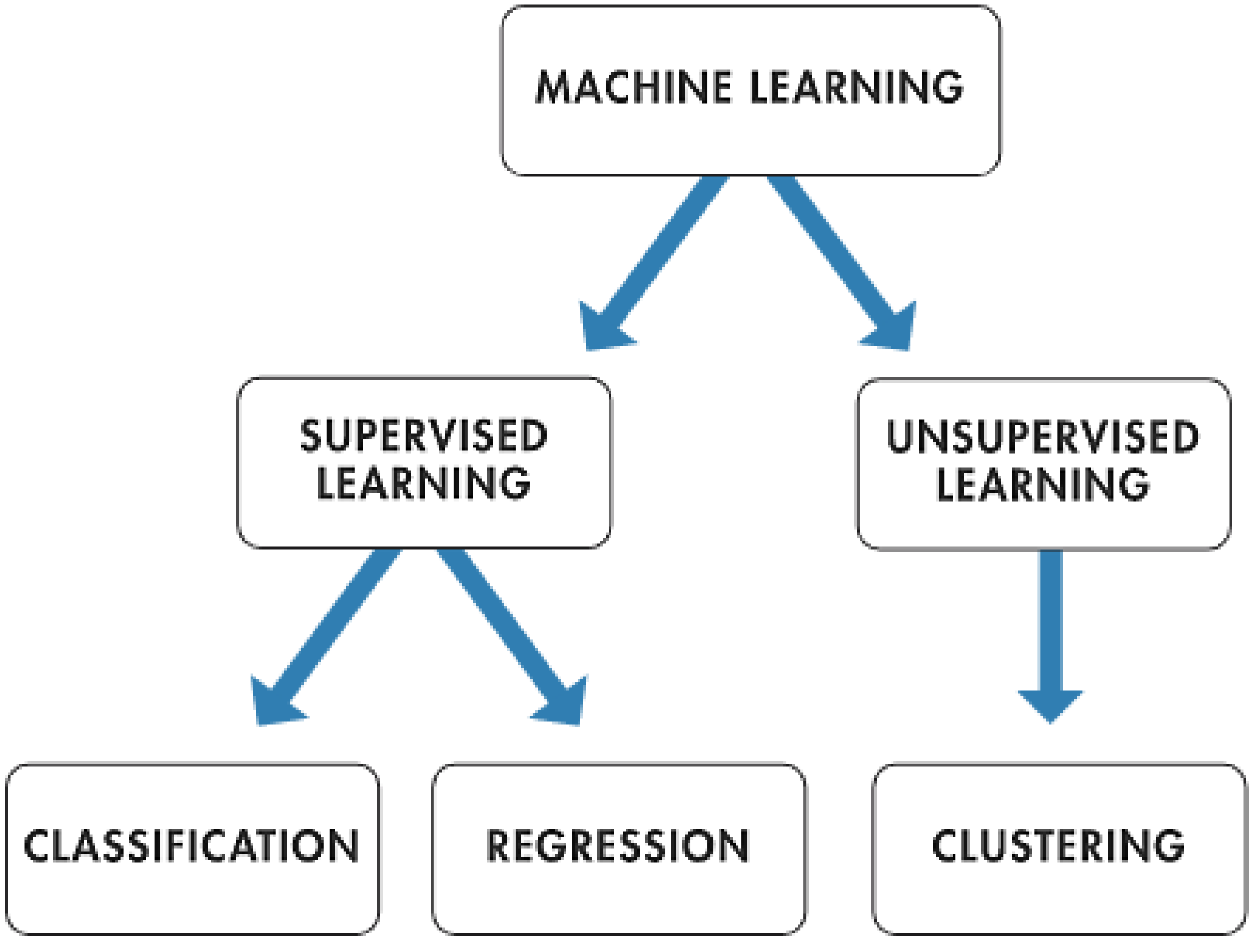
DATASET

Variables

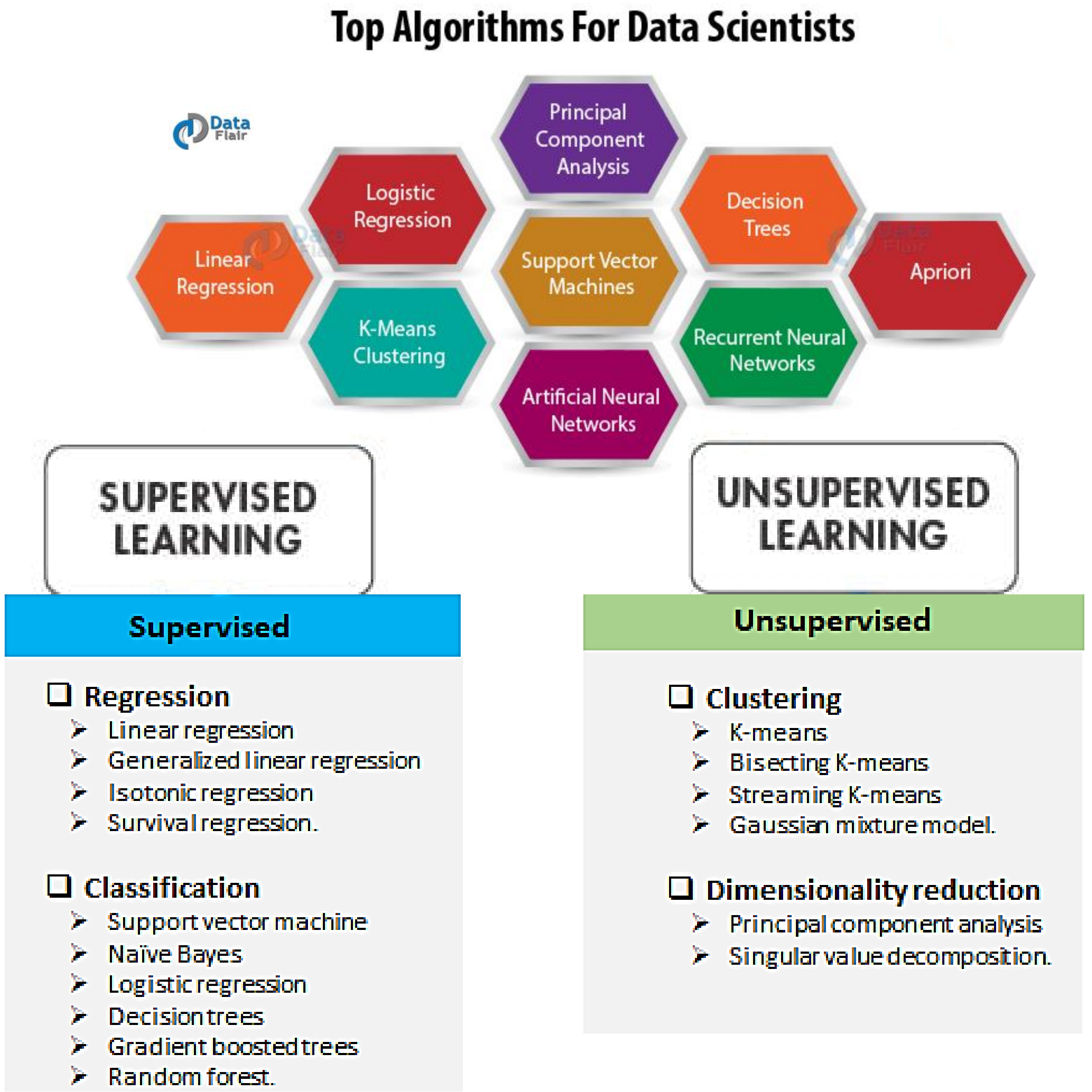
Cases

Store	Sales	Customers	Refunds
1	14	9	3
2	19	13	4
3	22	19	4
4	24	20	3
5	29	26	8
6	40	34	6

Виды машинного обучения



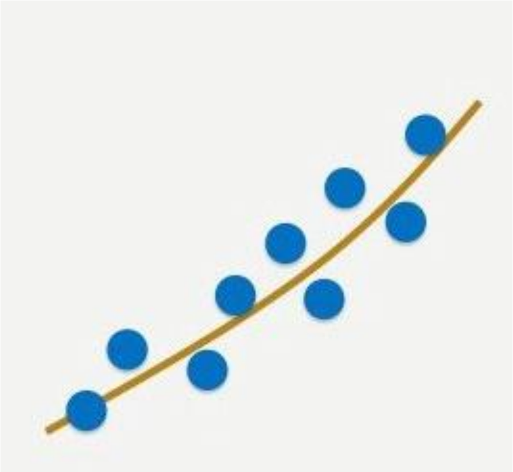
Виды машинного обучения

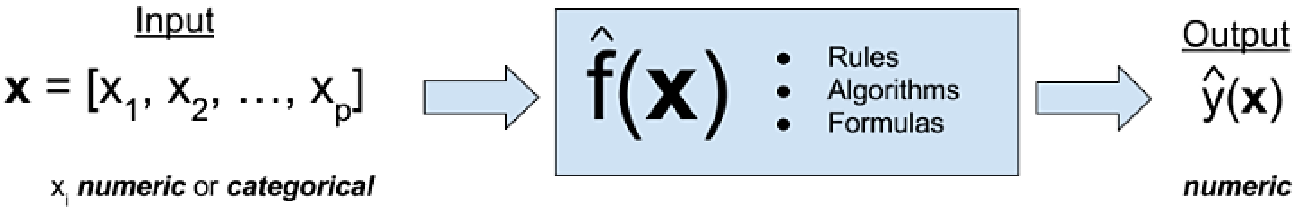


- **Supervised Learning**
  по маркированным данным
  (X_1, X_2, ..., X_m, Y)
  создание модели для прогнозирования Y
- **Unsupervised Learning**
  по немаркированным данным
  (X_1, X_2, ..., X_m)
  выявление шаблонов и закономерностей

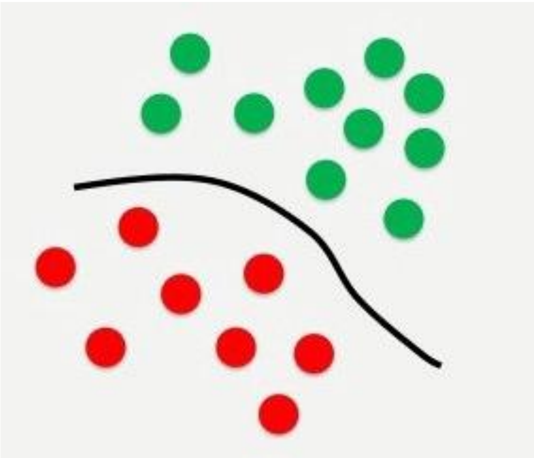
Модели машинного обучения

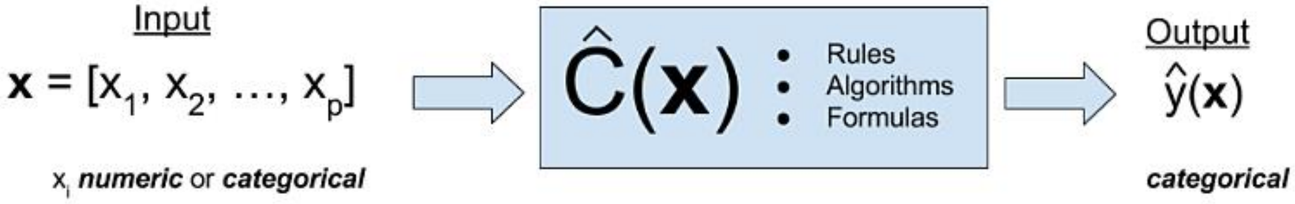
- **Модель регрессии**
  обучение на размеченных данных
  ($X_1, X_2, ..., X_m$, Y), где Y числовая переменная. <u>Цель:</u> получение наиболее точного прогноза или модели с минимальной ошибкой RMSE.



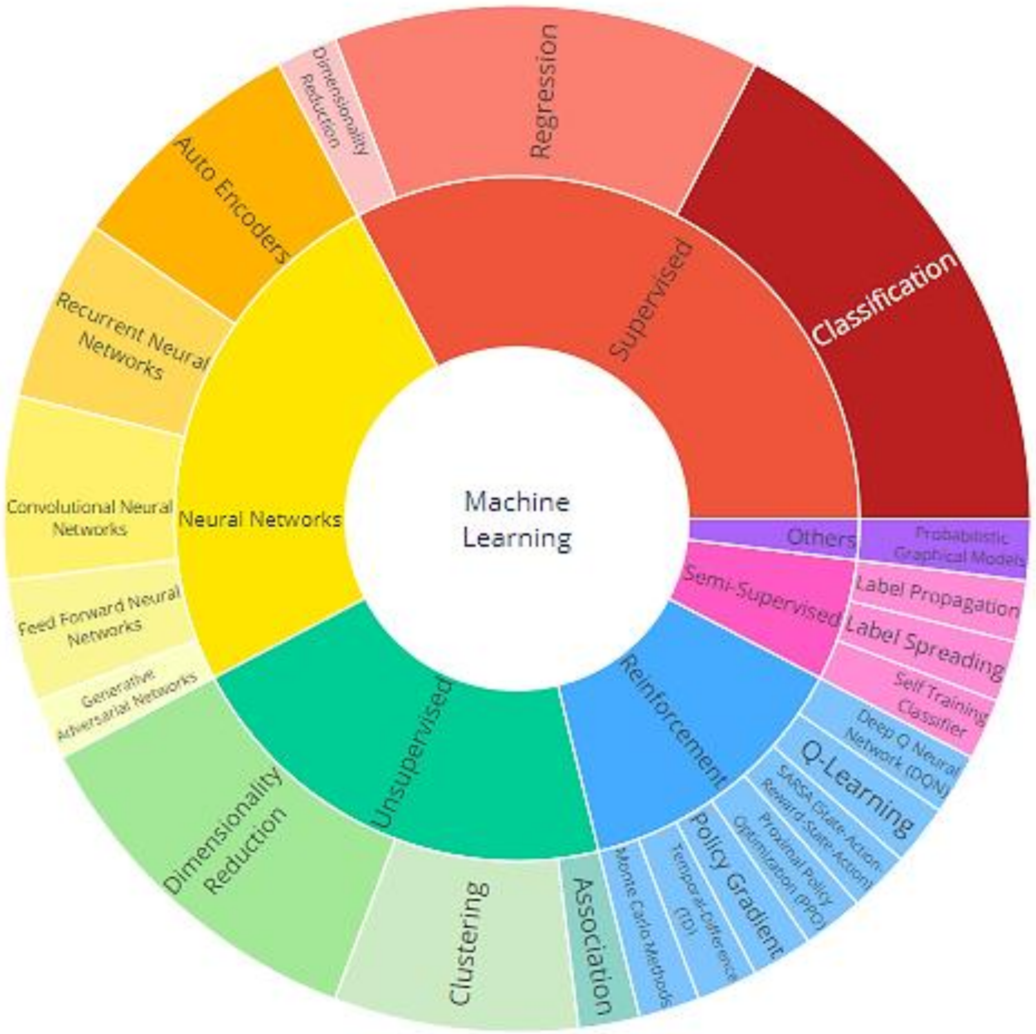


- **Модель классификации**
  обучение на размеченных данных
  ($X_1, X_2, ..., X_m$, Y), где Y категорная переменная.
  <u>Eesmärk:</u> получение модели для предсказания правильного класса с наибольшей вероятностью.





Алгоритмы машинного обучения

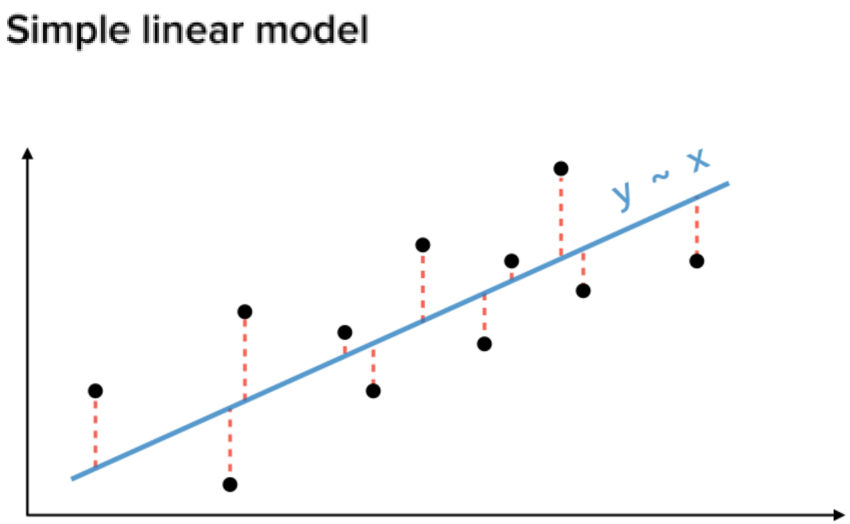


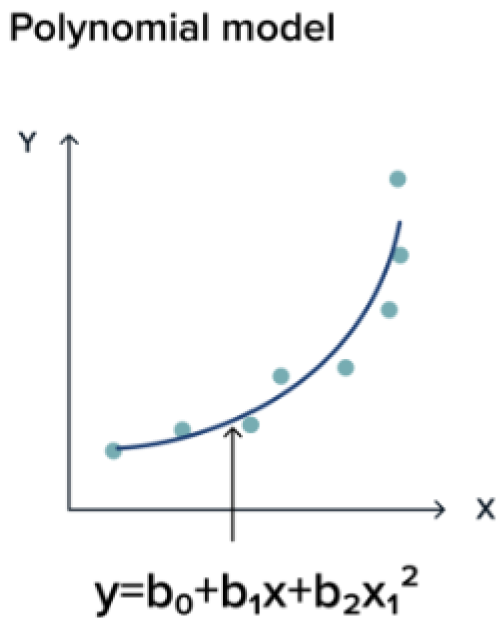
https://towardsdatascience.com/mars-multivariate-adaptive-regression-splines-how-to-improve-on-linear-regression-e1e7a63c5eae

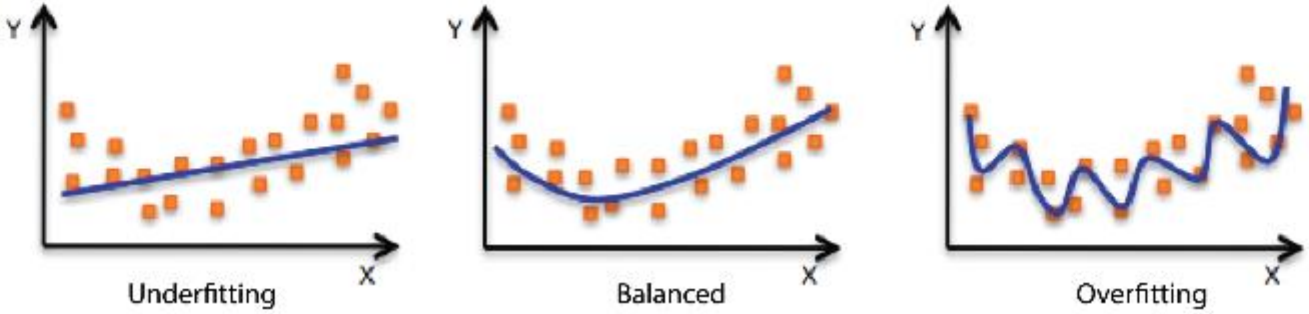
Алгоритмы машинного обучения

- **Модель (линейной) регрессии**

$$Y = X\beta + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$



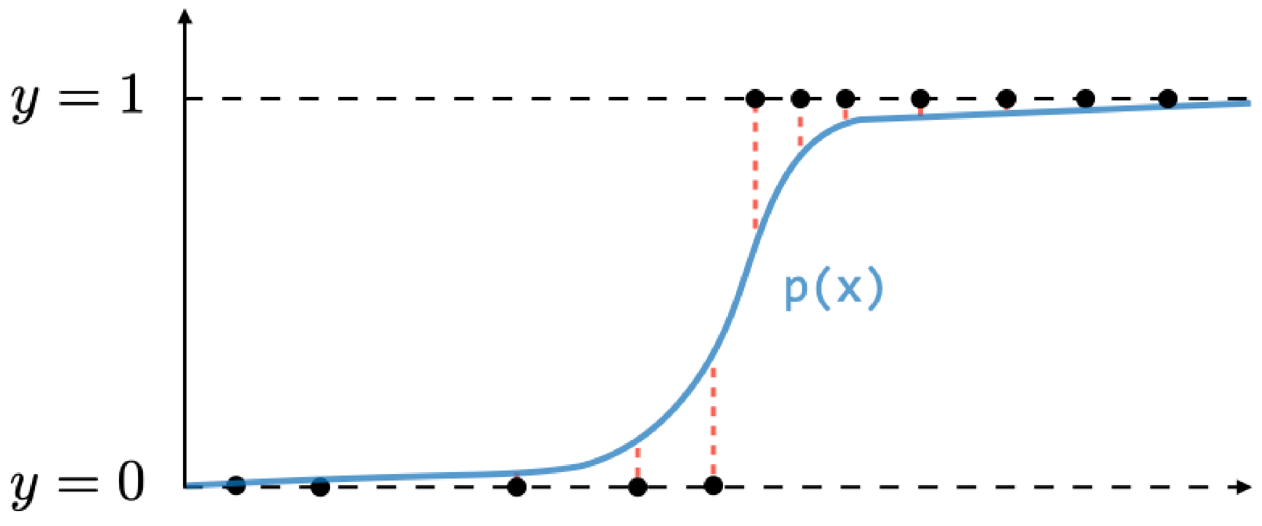




Модели классификации

- Логистическая регрессия

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

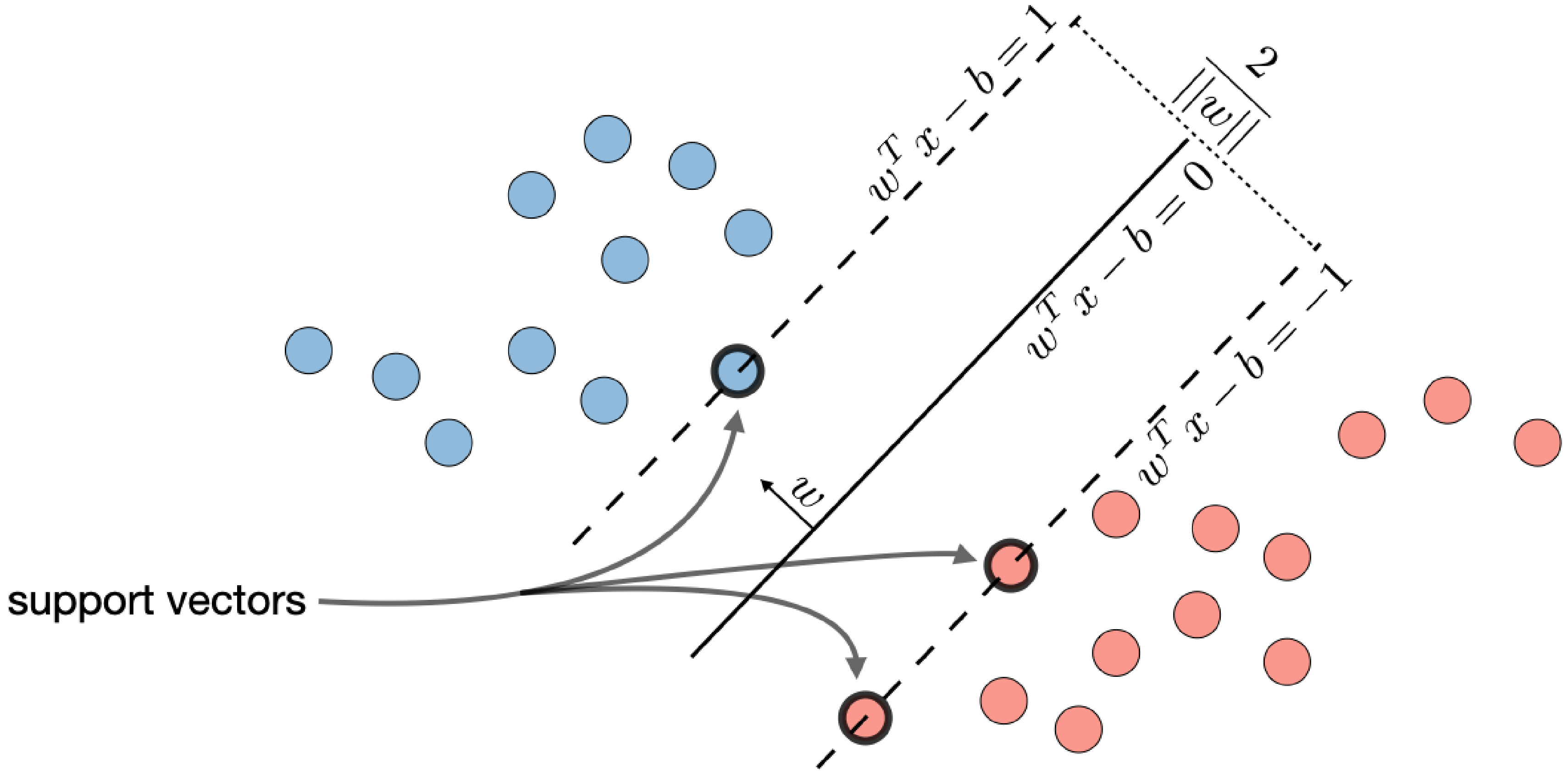


$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

$$y \in \{0, 1\} = \begin{cases} 0: negatiivne\ klass\ (mees) \\ 1: positiivne\ klass\ (naine) \end{cases}$$

Модели классификации

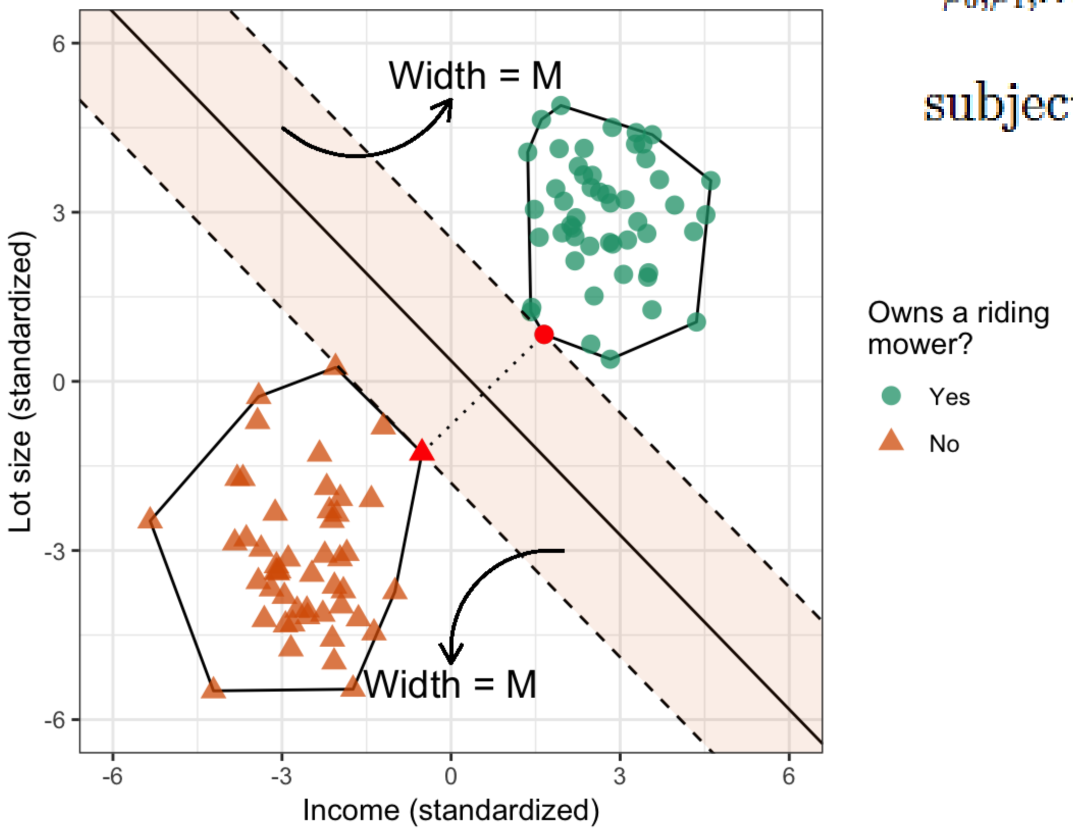
- **SVM - Support vector machines – метод опорных векторов**



Модели классификации

- **SVM - Support vector machines – метод опорных векторов**

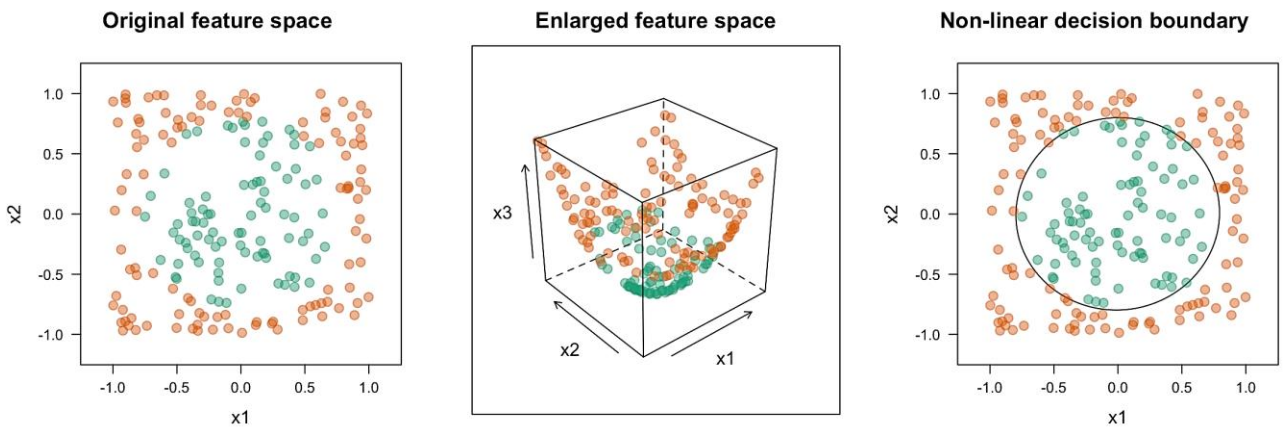
Задача оптимизации



$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \begin{cases} \sum_{j=1}^{p} \beta_j^2 = 1, \\ y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \geq M, \quad i = 1, 2, \ldots, n \end{cases}$$

Kernel trick

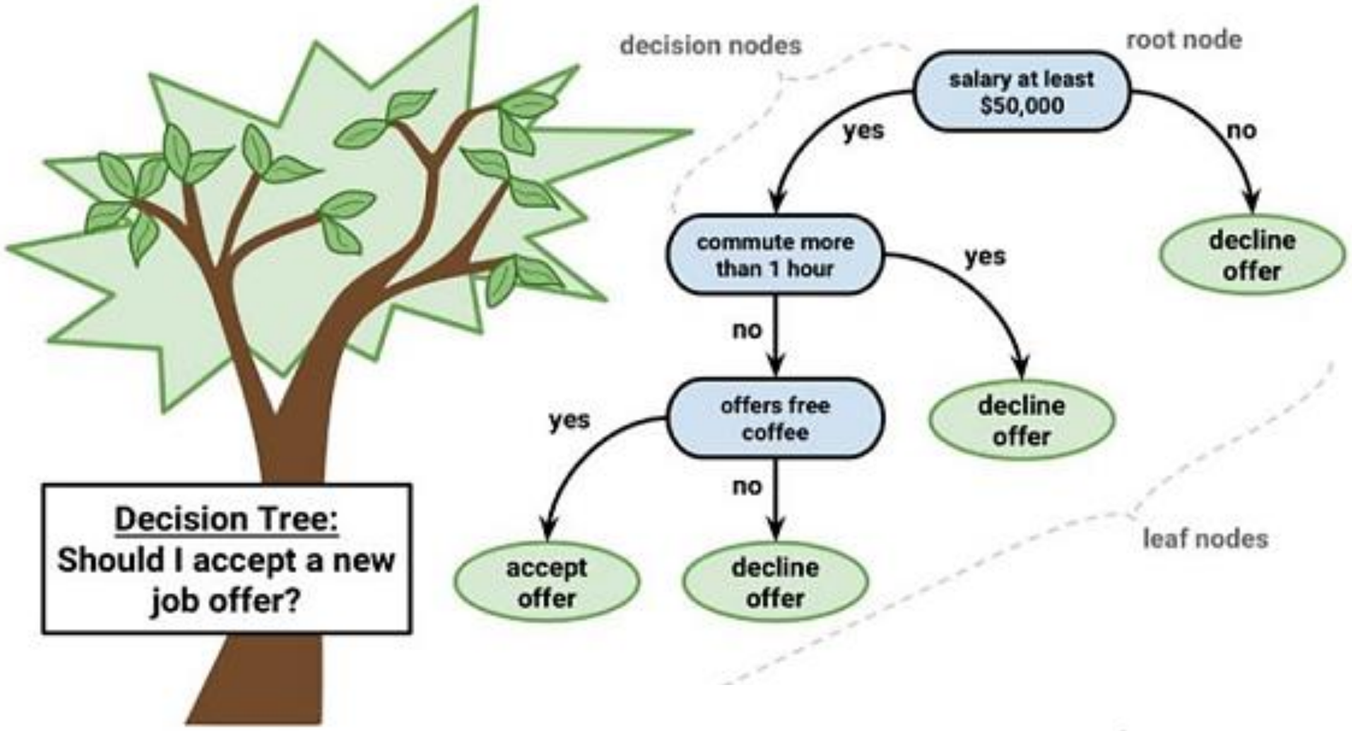


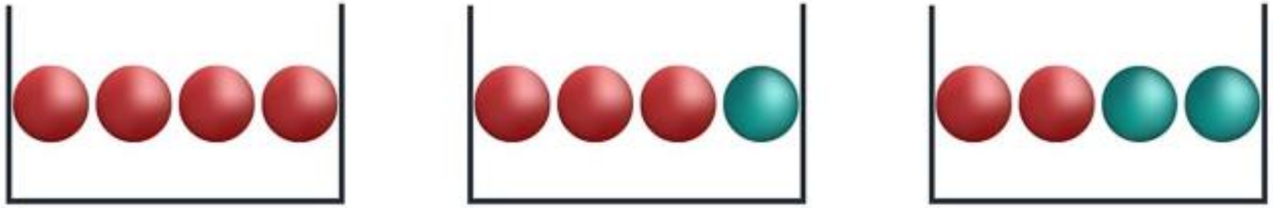
Модели классификации

- Деревья решений

$$Entropy = \sum_i -p_i \log_2 p_i$$

$$Gini = 1 - \sum_{i=1} (p_i)^2$$

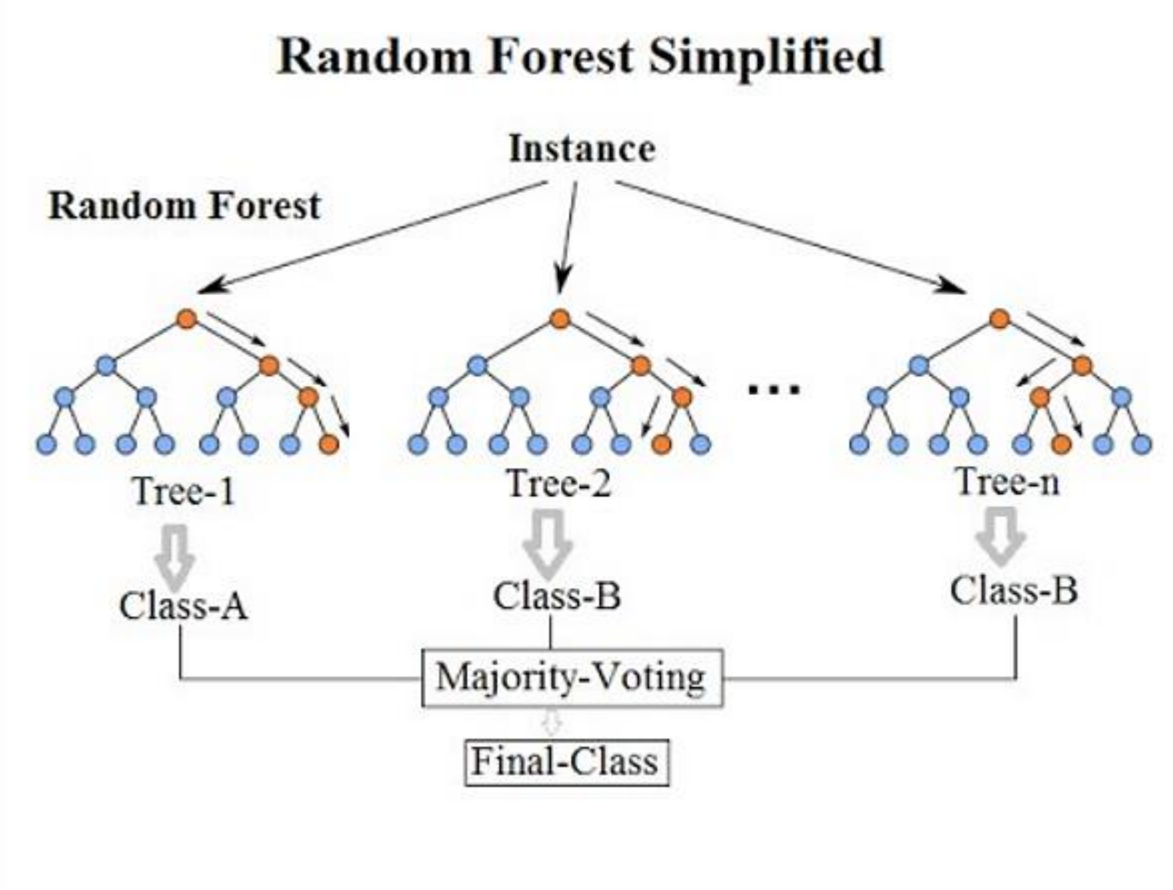


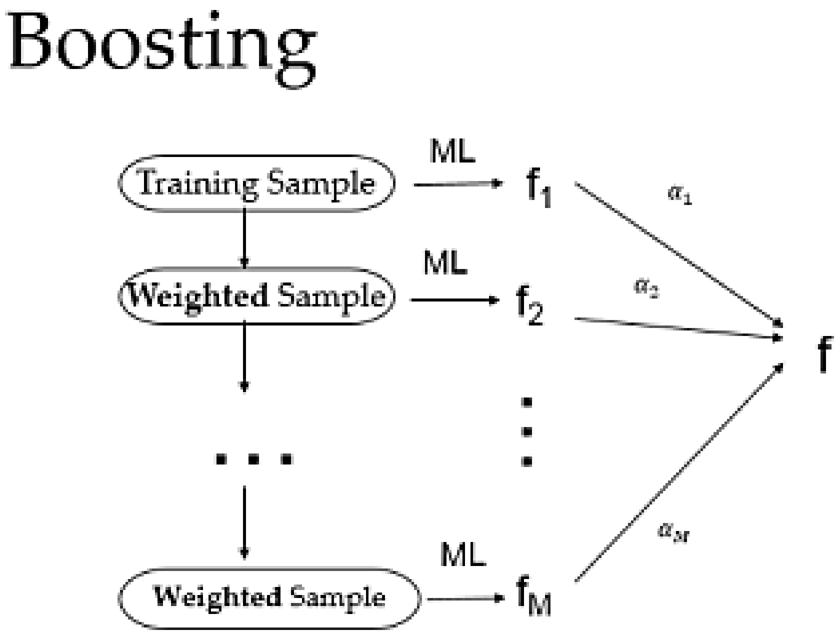


Impurity:	madal	keskmine	kõrge
Entropy:	$-1 \log_2 1 = 0$	$-1/4 \log_2 1/4 - 3/4 \log_2 3/4 = 0.81$	$-2*2/4 \log_2 2/4 = 1$

Модели классификации

- **Random Forest**

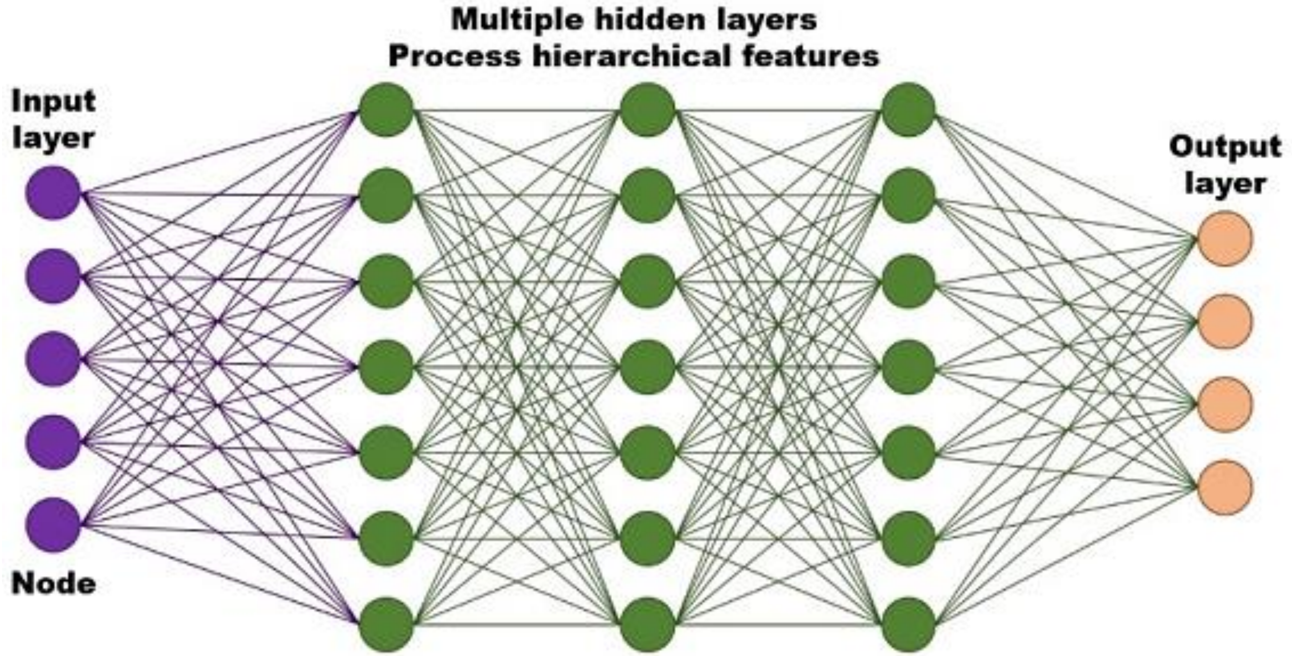


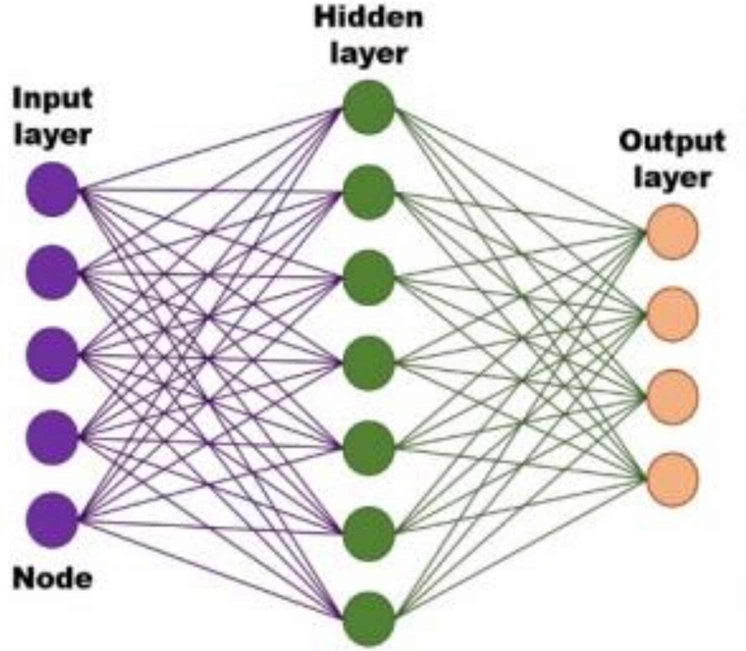


Модели классификации

- **Neural Networks**

$$P(y_i = Ck|\boldsymbol{X} = \boldsymbol{x_i}) = \phi_o(\alpha_k + \sum_h w_{hk}\phi_h(\alpha_h + \sum_{j=1}^{p} w_{ih}x_{ij}))$$



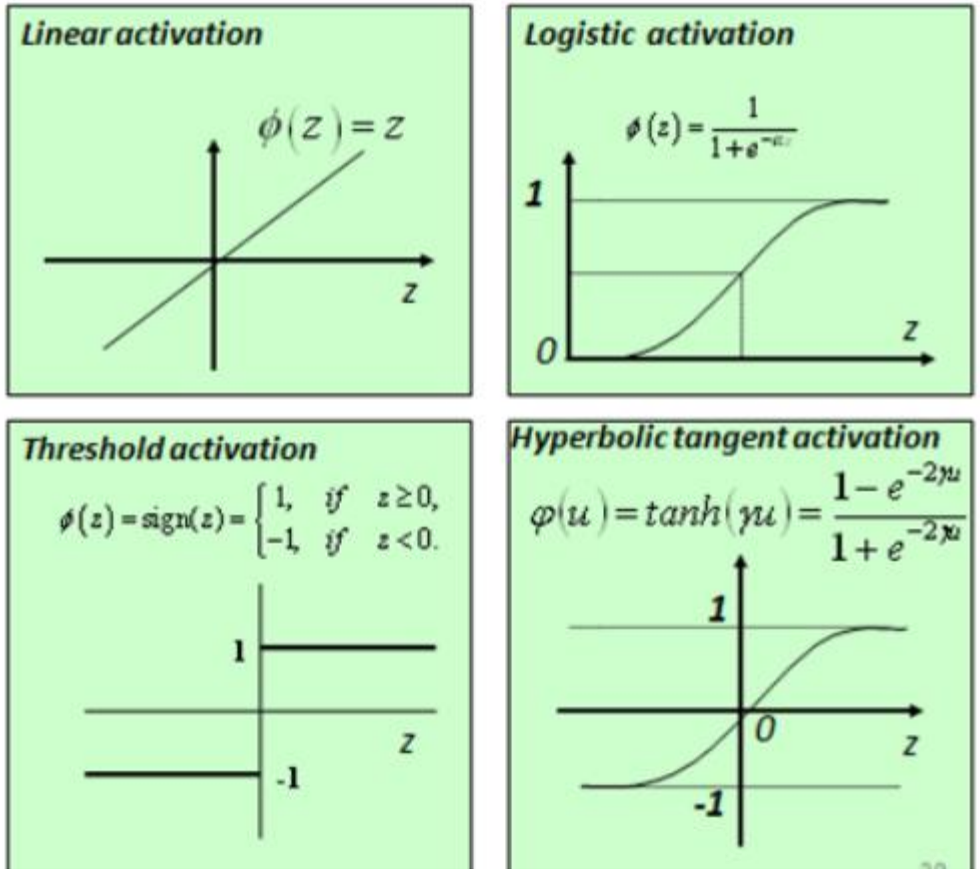


Модели классификации

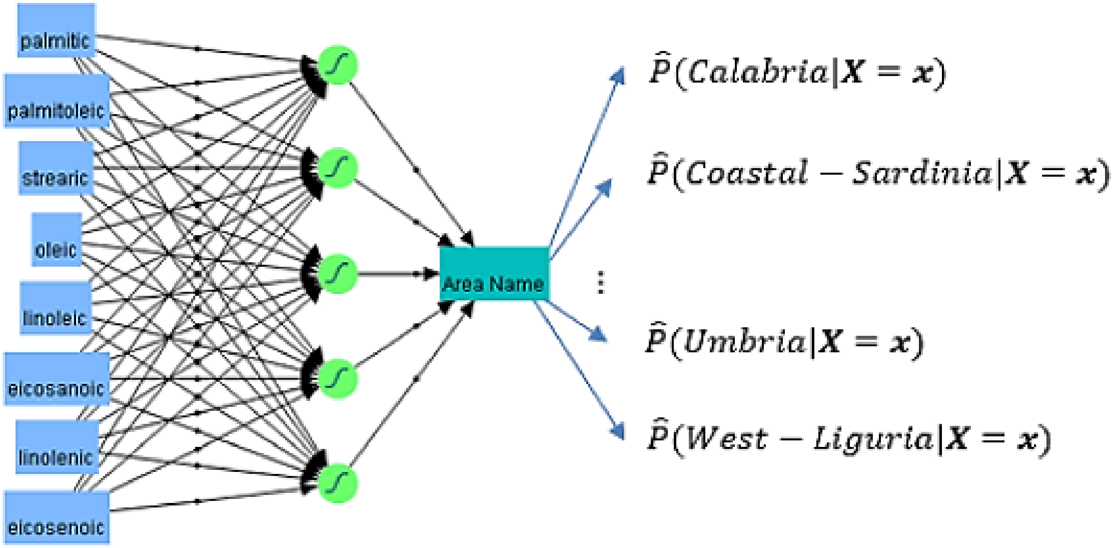
- Neural Networks

$$P(y_i = Ck | \boldsymbol{X} = \boldsymbol{x}_i) = \phi_o(\alpha_k + \sum_h w_{hk}\phi_h(\alpha_h + \sum_{j=1}^{p} w_{ih}x_{ij}))$$

- $\boldsymbol{\phi_h}$

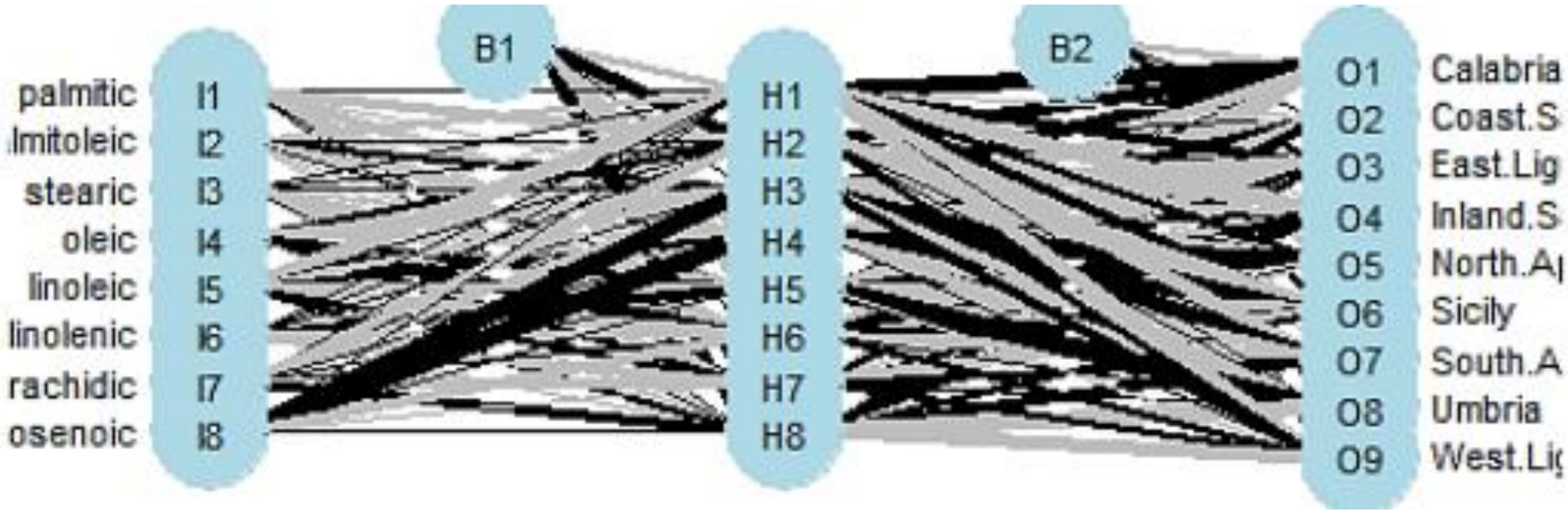


- $\boldsymbol{\phi_o}$
- Sigmoid $\dfrac{1}{1+e^{-x}}$
- (hyperbolic tangent +1)/2



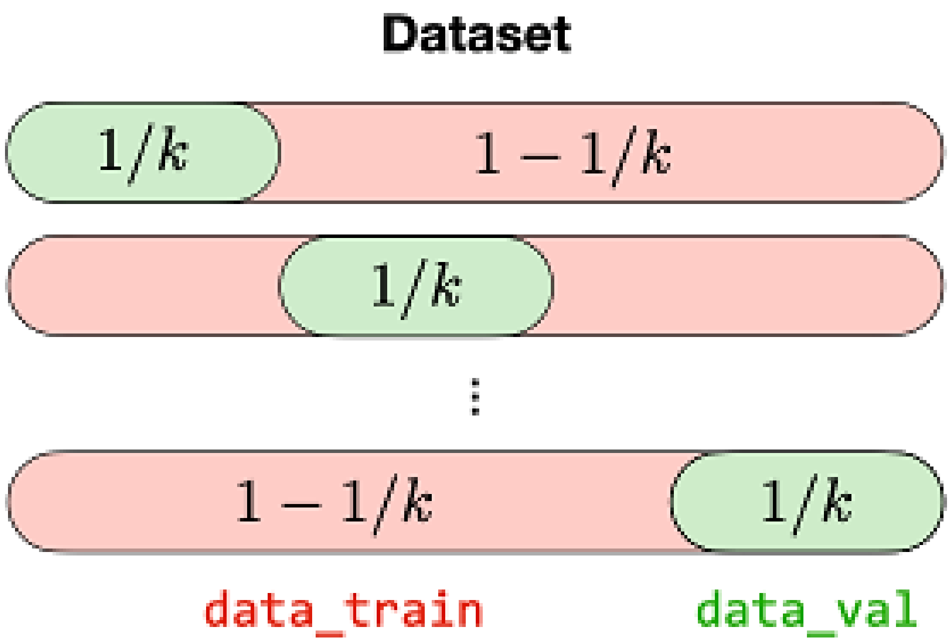
Модели классификации

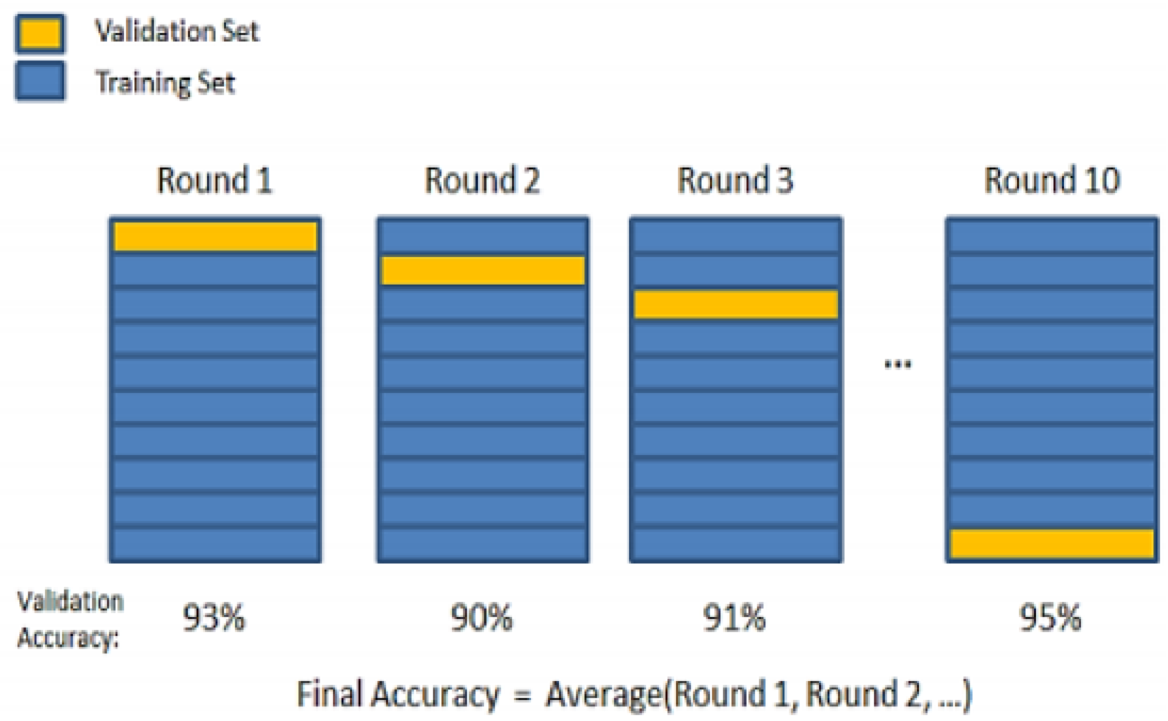
- Neural Networks



Валидирование модели

- **Cross-Validation**





Валидирование модели

- **Метрики регрессии**

Metric	Definition	Interpretation
RMSE	$\sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$	Root mean square error
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \widehat{y}_i\rvert$	Mean average error
R^2	$1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$ with sum of squares total (SST) and errors (SSE) defined as: • $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ • $\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	Proportion of the variance that is explained by the model

Валидирование модели

- **Метрики классификации**

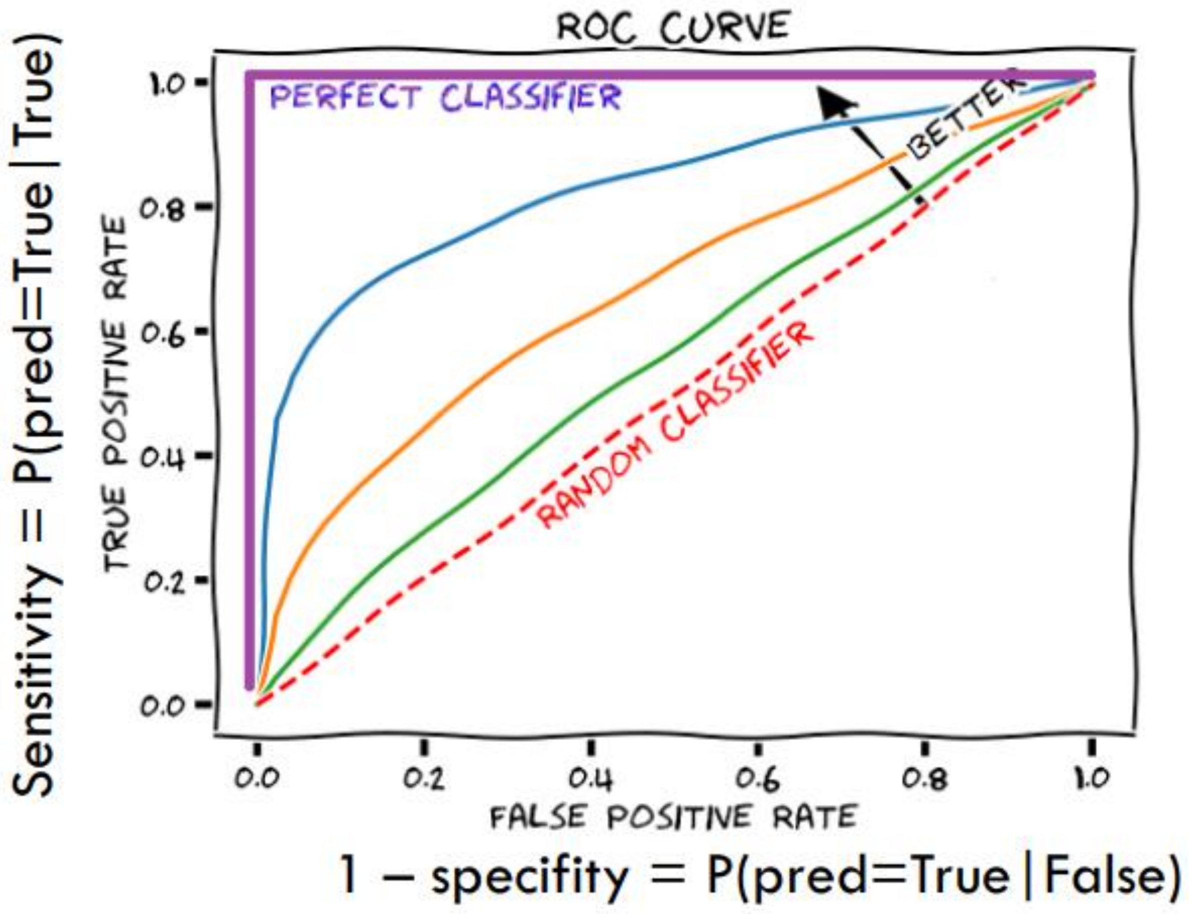
Predicted Class

Actual Class		Positive	Negative	
	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

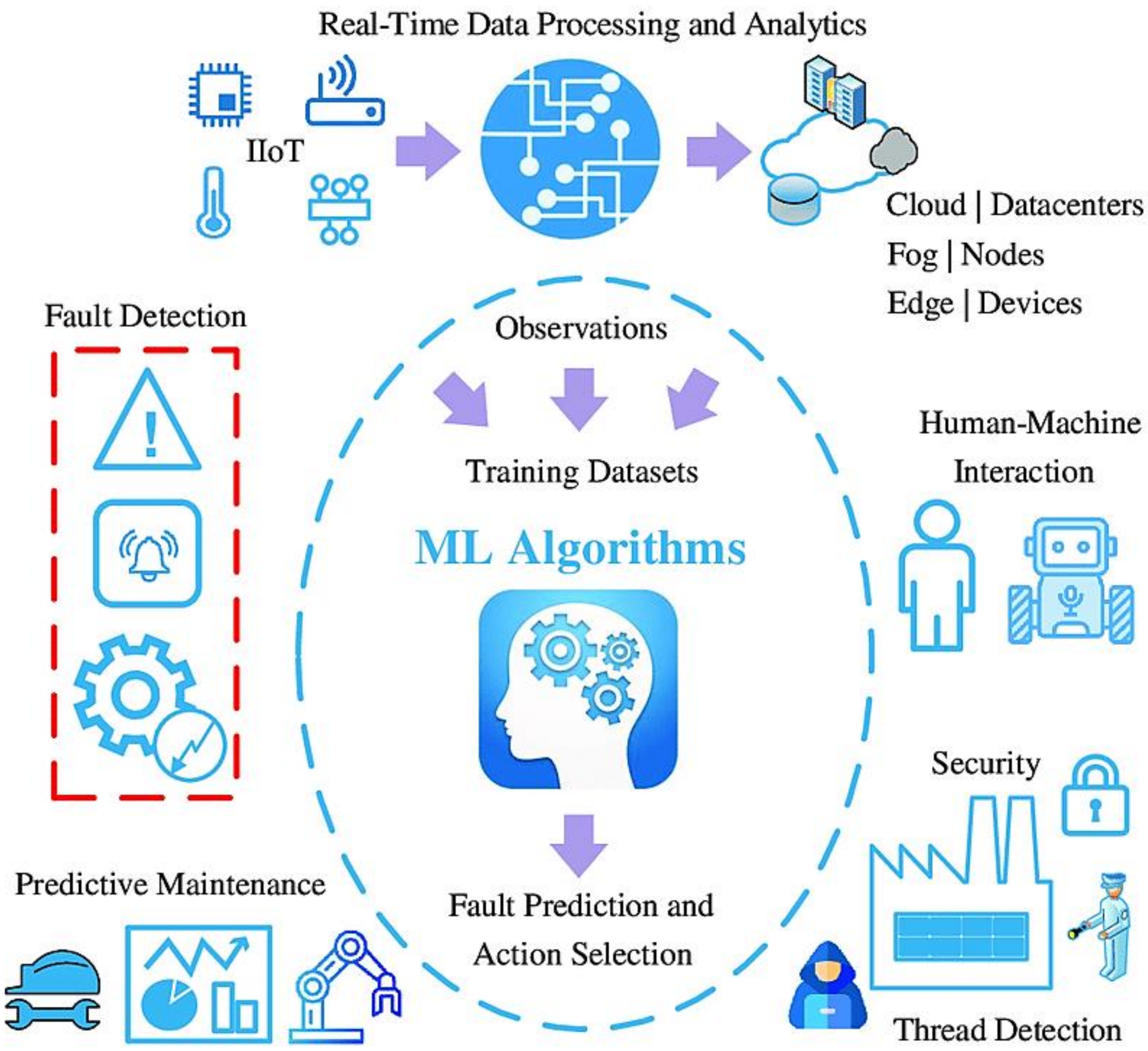
Metric	Definition	Interpretation
Confusion matrix	Matrix of TP, FP, TN, FN	Granular view of model performance
Accuracy	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	How accurate positive predictions are
Precision	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	How accurate positive predictions are
Recall Sensitivity TPR	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	Coverage of actual positive samples
Specificity TNR	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$	Coverage of actual negative samples
F1 score	$\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$	Harmonic mean of precision and recall
Matthew's correlation coefficient	$\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$	Hybrid metric between -1 and 1, useful for unbalanced classes
Receiving operating curve	Plot with TPR wrt FPR for different cutoff values	Identity line is a random guess classifier.
AUC	Area under the curve that plots TPR wrt FPR	Value between 0 and 1, where 0.5 is equivalent of a classifier that makes random predictions.

Валидирование модели

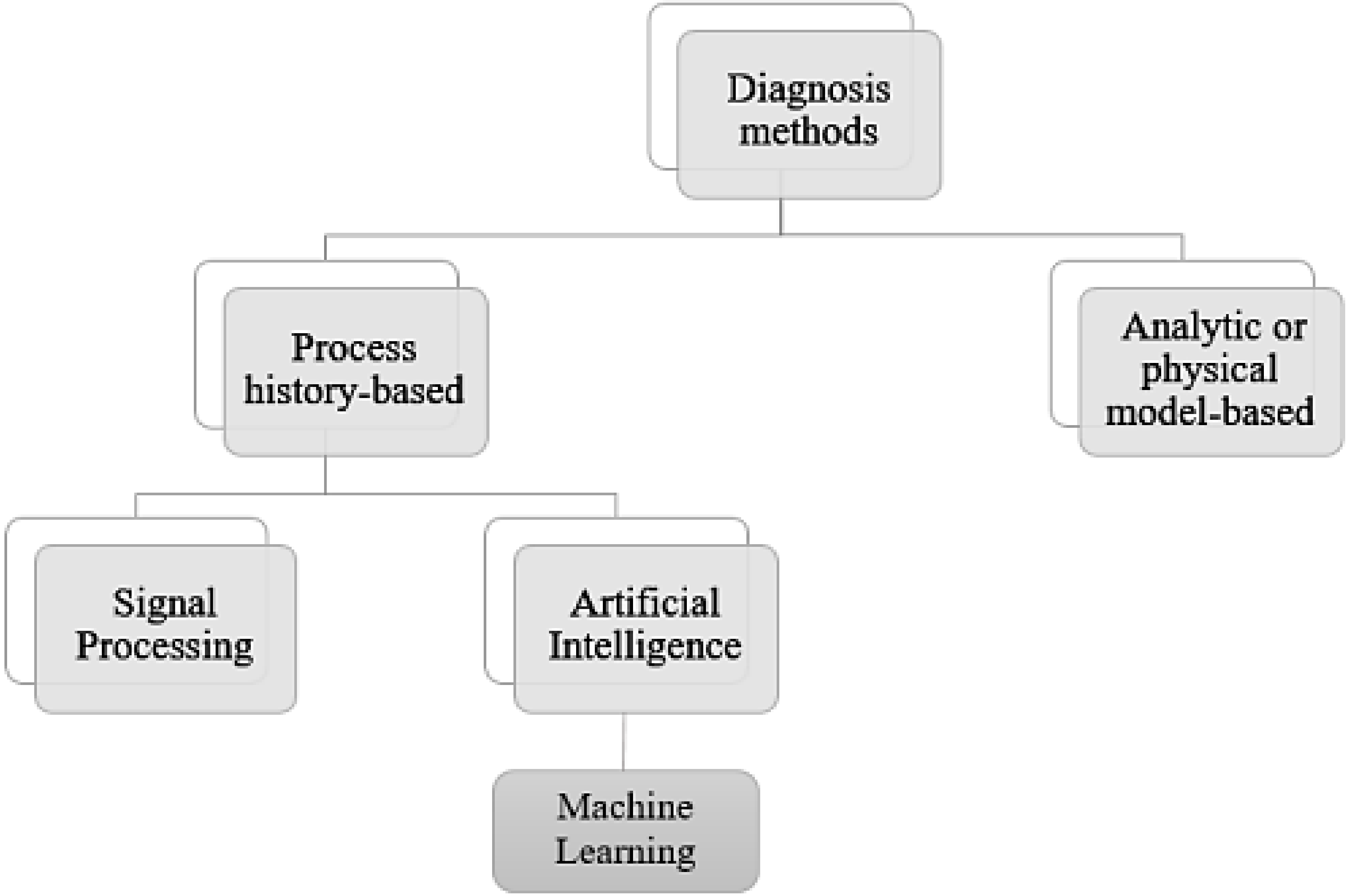
- **Метрики классификации**



Машинное обучение на предприятии



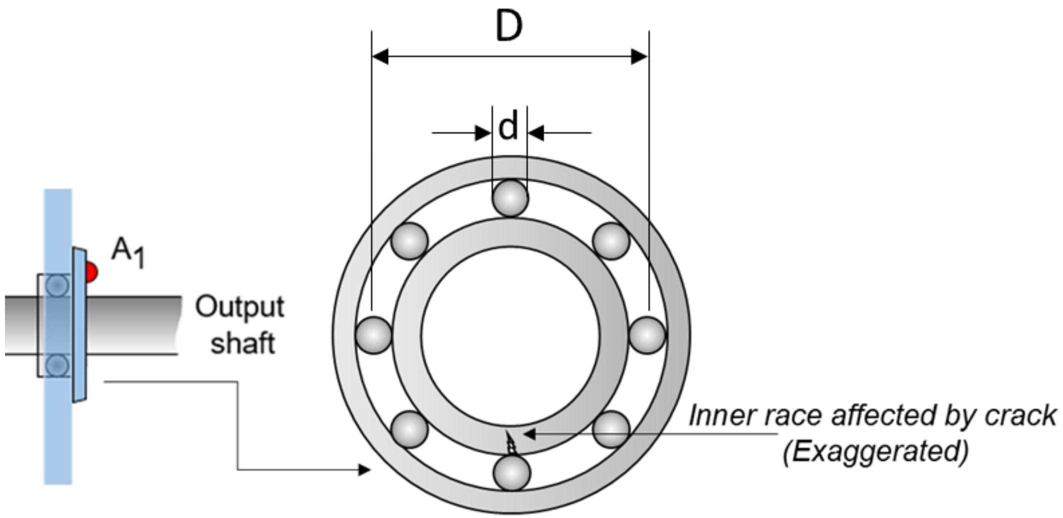
Методы диагностики неисправностей



The different methods in fault diagnosis

Методы диагностики неисправностей: примеры

Rolling Element Bearing Fault Diagnosis



Ball bearing cross−section (Magnified)

machines MDPI

Article

Fault Detecting Accuracy of Mechanical Damages in Rolling Bearings

Karolina Kudelina [1,*], Tatjana Baraškova [2], Veroonika Shirokova [2], Toomas Vaimann [1] and Anton Rassõlkin [1]

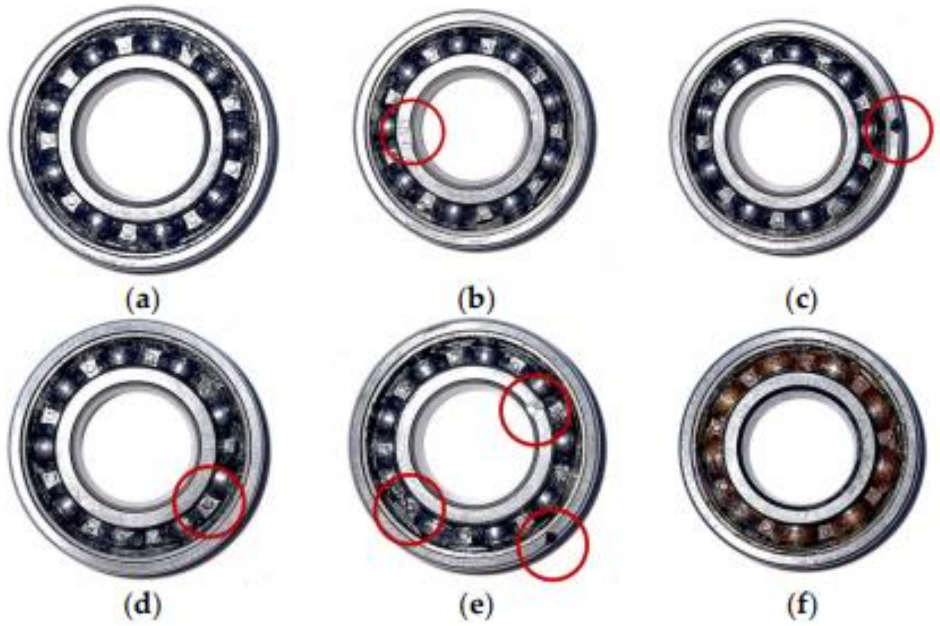
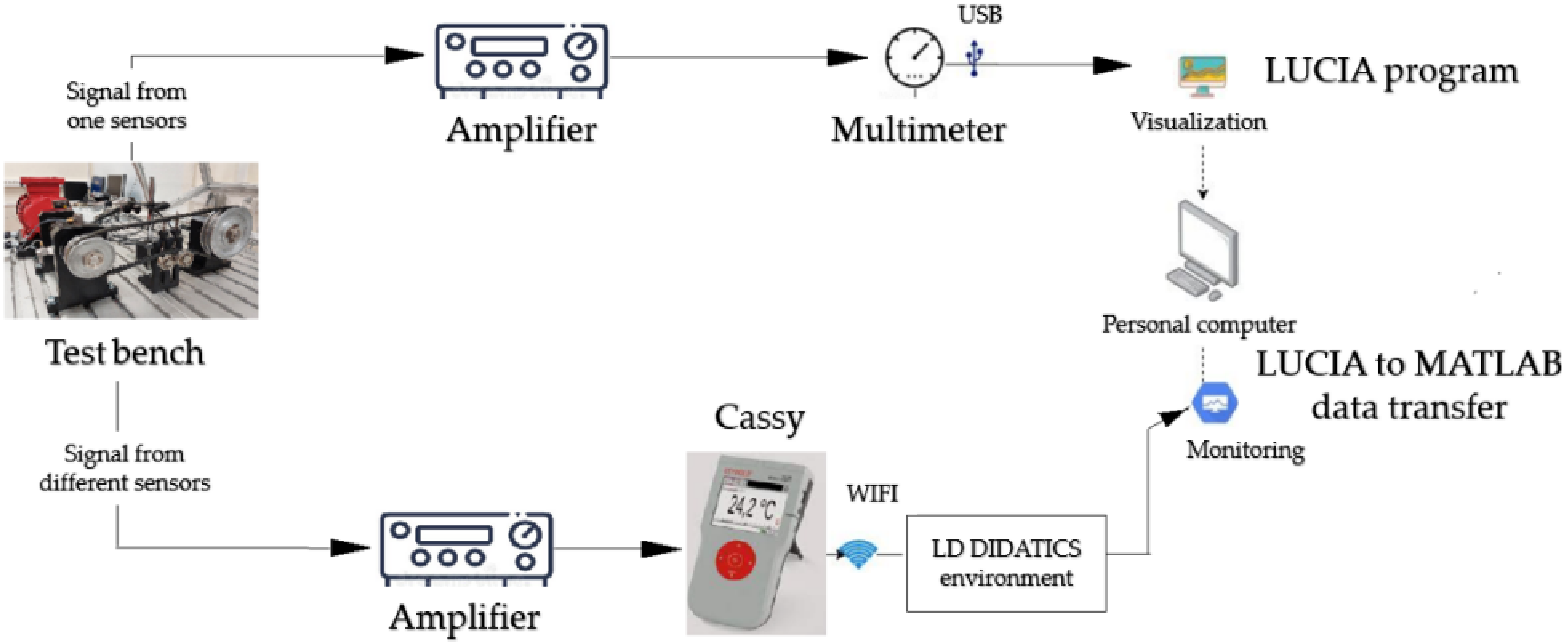


Figure 8. Studied bearings with different damages: (**a**) healthy bearing, (**b**) bearing with damaged inner ring, (**c**) damaged outer ring, (**d**) damaged cage, (**e**) complex damage, and (**f**) material fatigue.

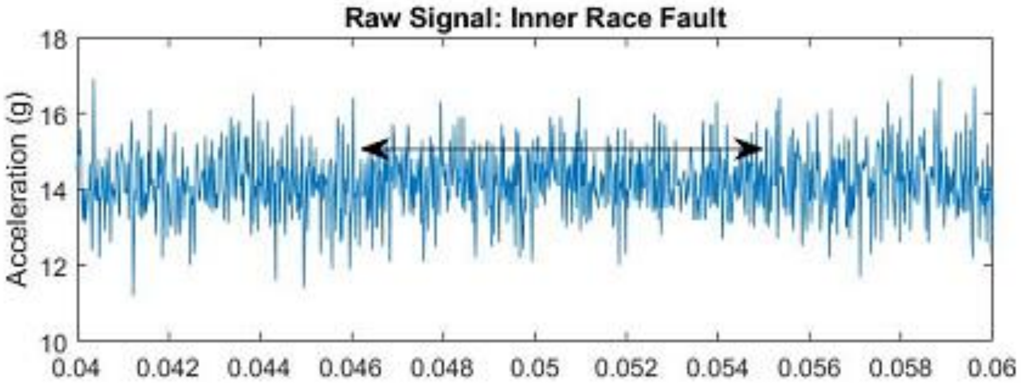
Методы диагностики неисправностей: примеры

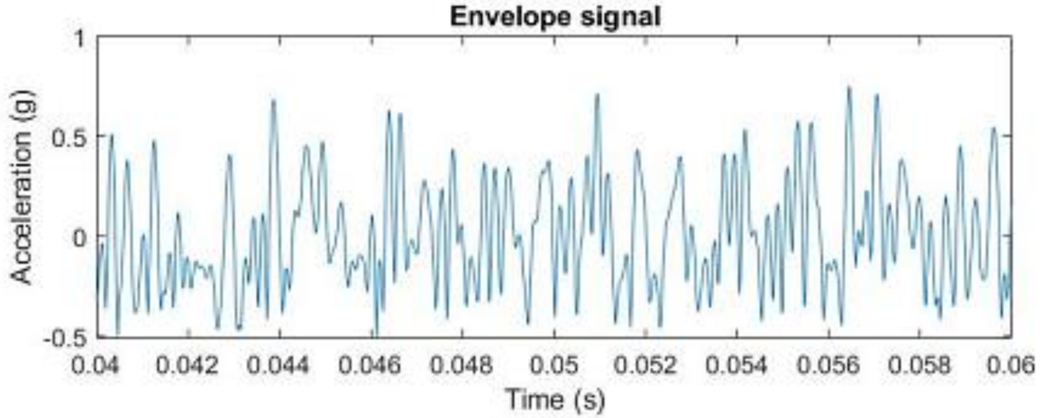
Rolling Element Bearing Fault Diagnosis



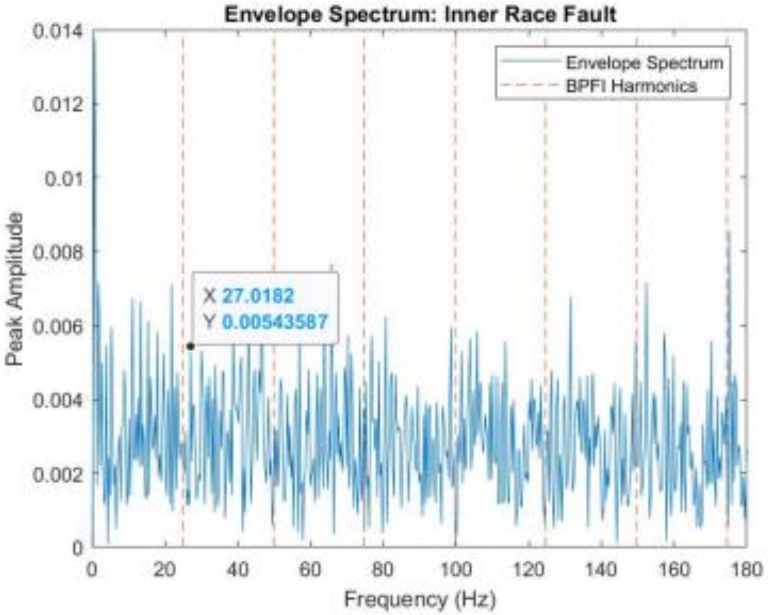
Методы диагностики неисправностей: примеры

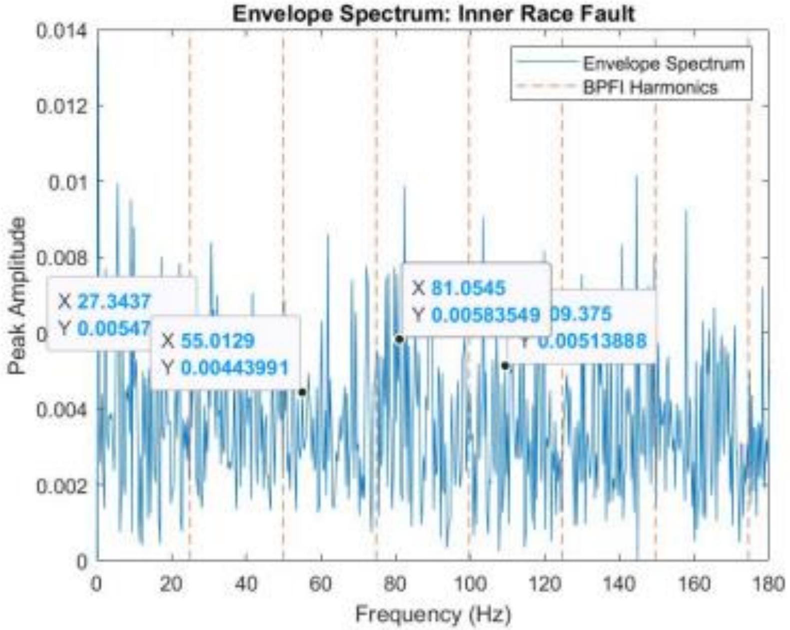
Rolling Element Bearing Fault Diagnosis





FFT





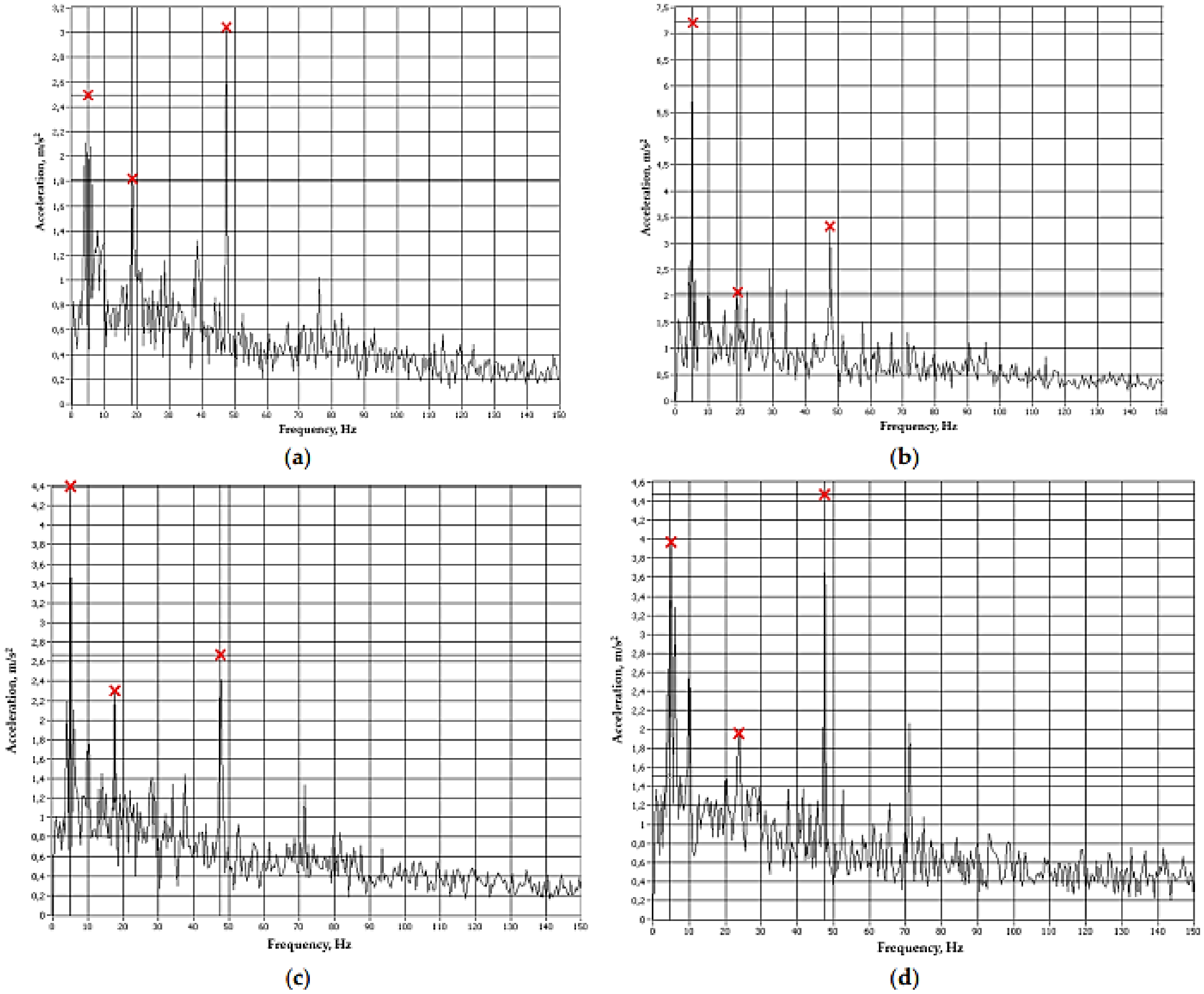
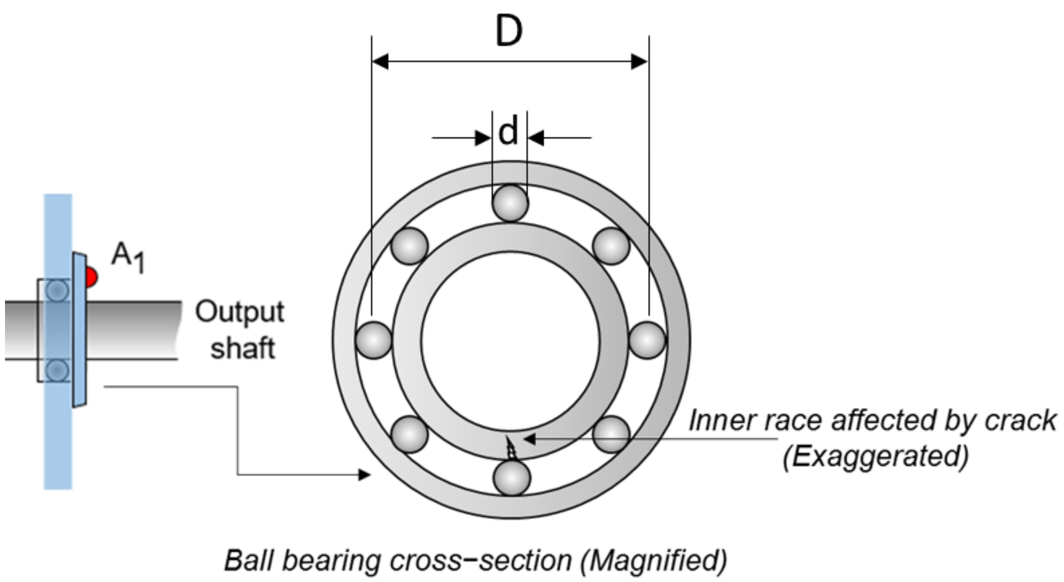


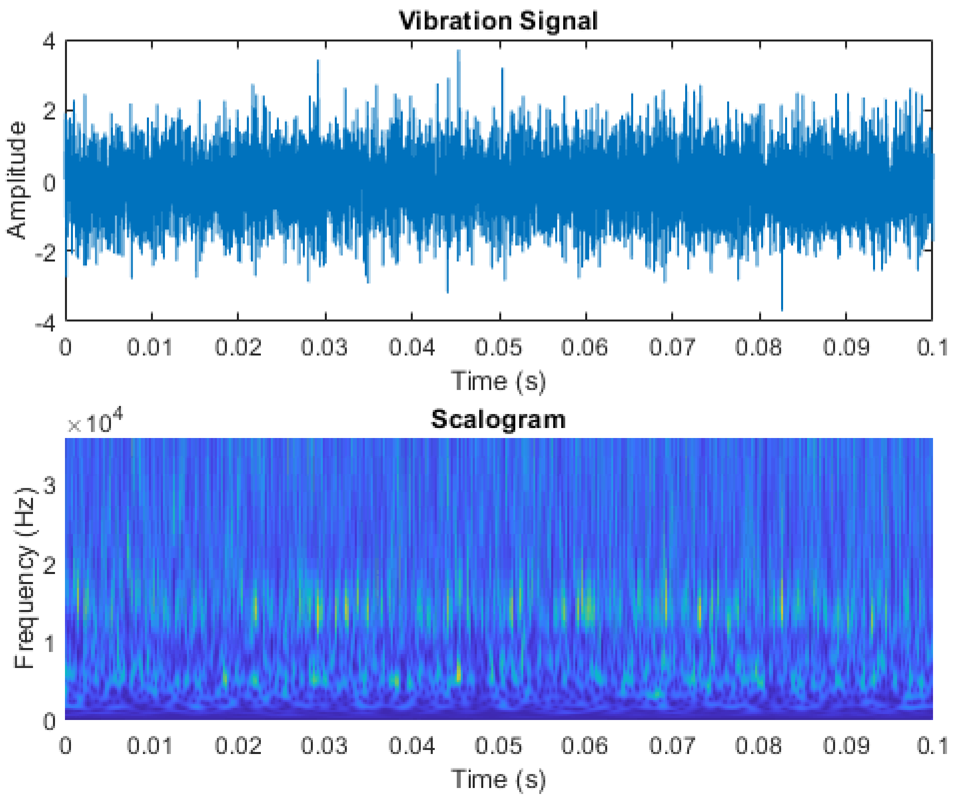
Figure 16. Vibration spectra of faulty bearings: (**a**) damaged outer ring, (**b**) damaged cage, (**c**) complex damage, and (**d**) material fatigue.

Методы диагностики неисправностей: примеры

Rolling Element Bearing Fault Diagnosis Using Deep Learning



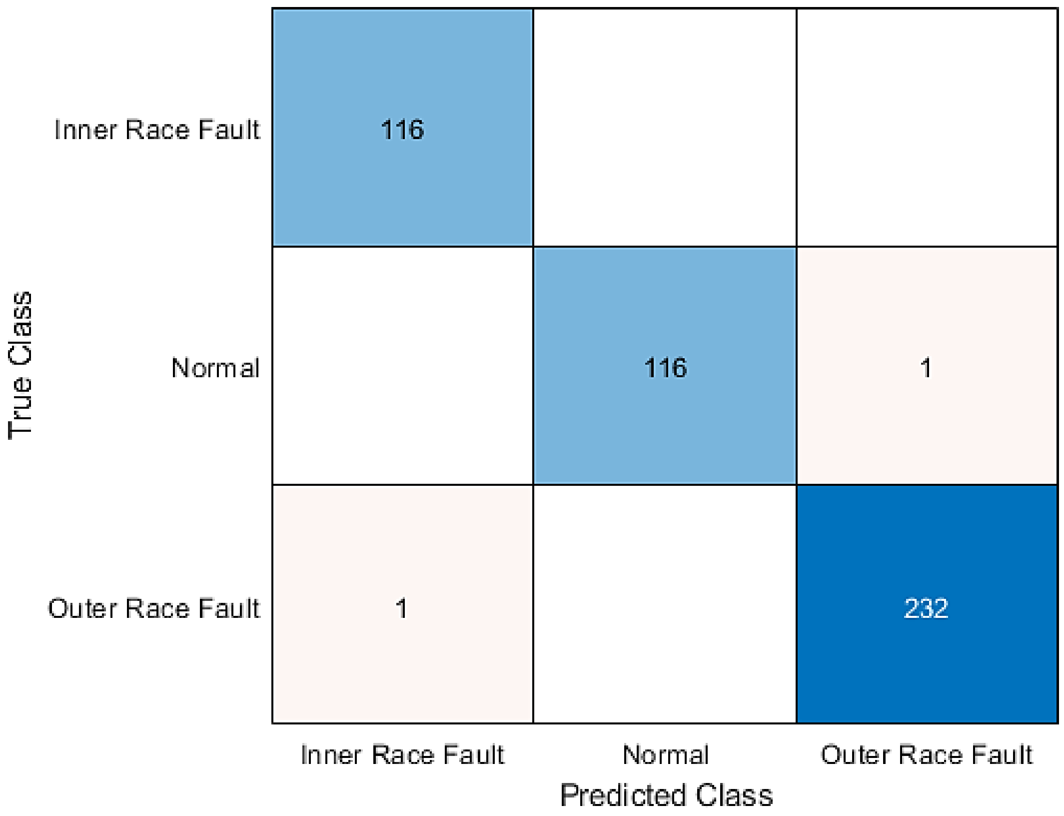
- pretrained SqueezeNet convolutional neural network
- SqueezeNet has been trained on over a million images



https://se.mathworks.com/help/predmaint/ug/rolling-element-bearing-fault-diagnosis-using-deep-learning.html

Методы диагностики неисправностей: примеры

Rolling Element Bearing Fault Diagnosis Using Deep Learning

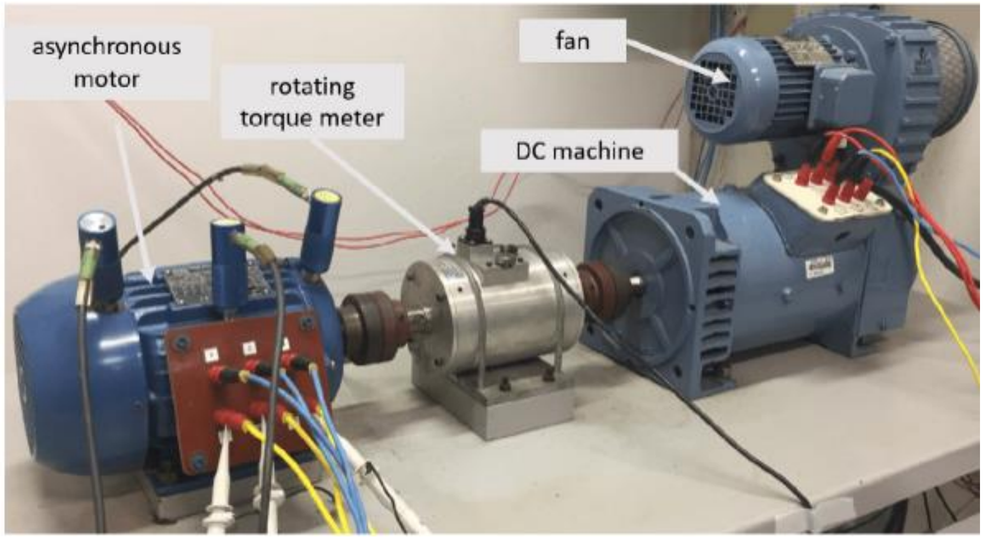


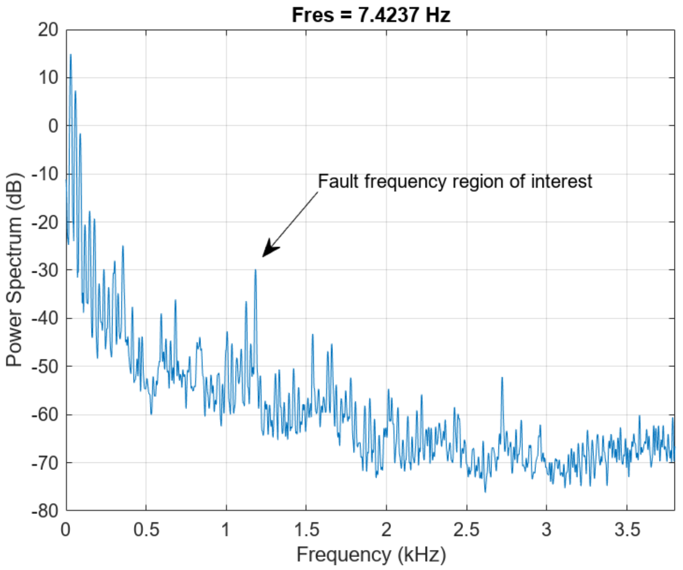
- accuracy = 0.9957
- average accuracy ~98%

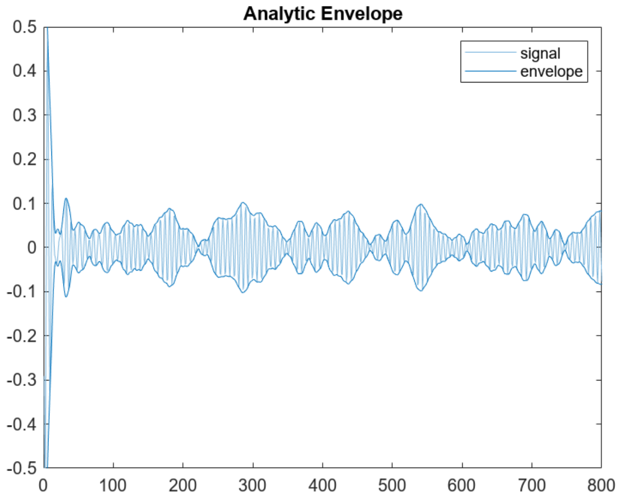
https://se.mathworks.com/help/predmaint/ug/rolling-element-bearing-fault-diagnosis-using-deep-learning.html

Методы диагностики неисправностей: примеры

Broken Rotor Fault Detection in AC Induction Motors Using Vibration and Electrical Signals







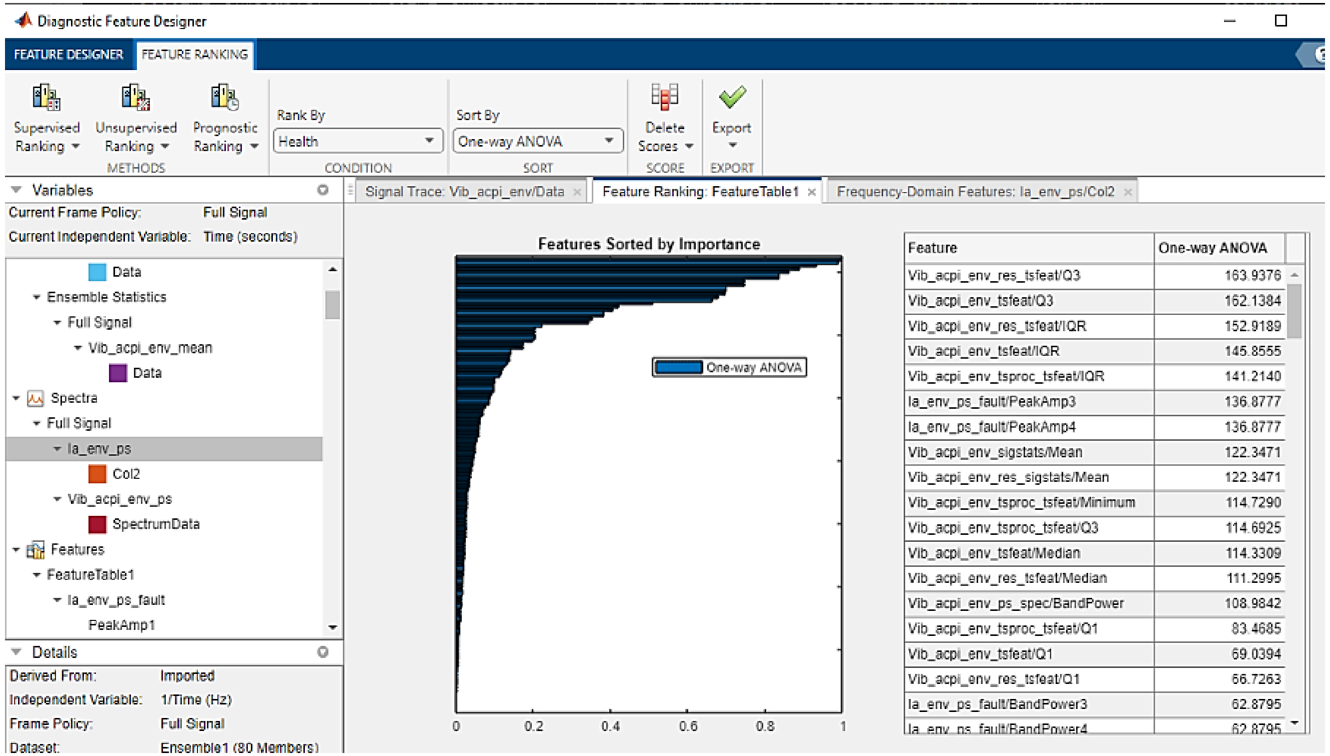
Ia	Vib_acpi	Vib_acpi_env	Ia_env_ps	Health	Load
{50001×1 timetable}	{7601×1 timetable}	{7601×1 timetable}	{25001×2 double}	"healthy"	"torque05"

https://se.mathworks.com/help/predmaint/ug/broken-rotor-fault-detection-in-ac-induction-motors-using-vibration-and-electrical-signals.html

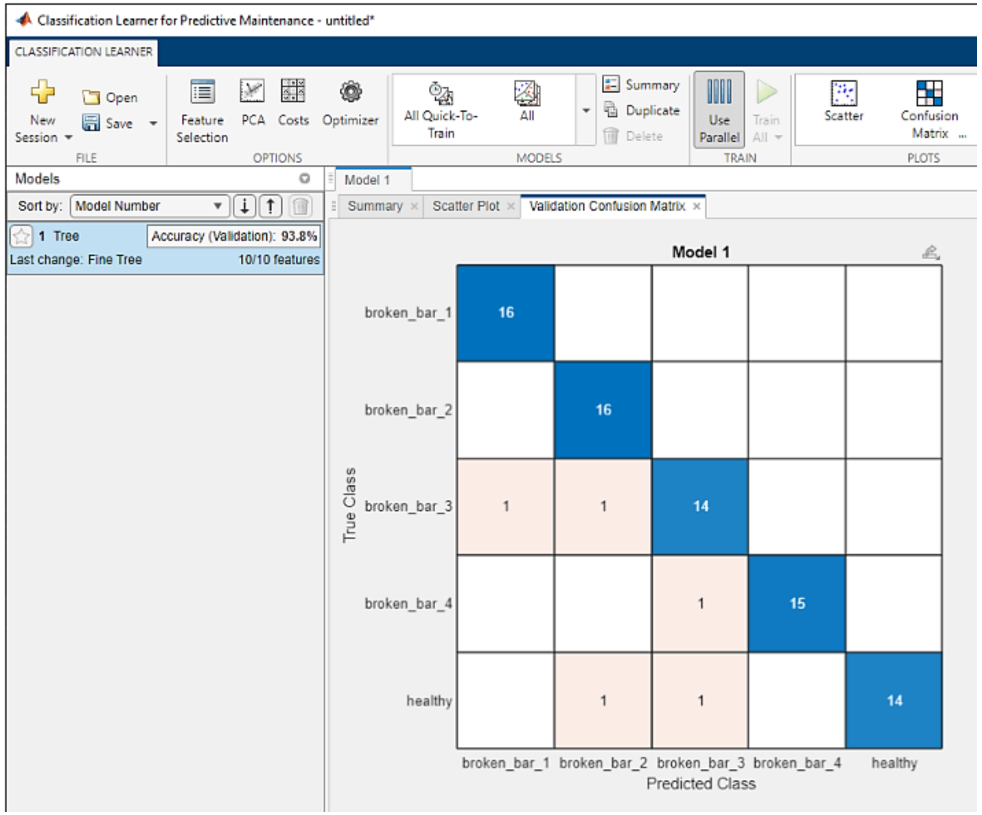
Методы диагностики неисправностей: примеры

Broken Rotor Fault Detection in AC Induction Motors Using Vibration and Electrical Signals

- **Diagnostic Feature Designer App**



- **Classification Learner App**

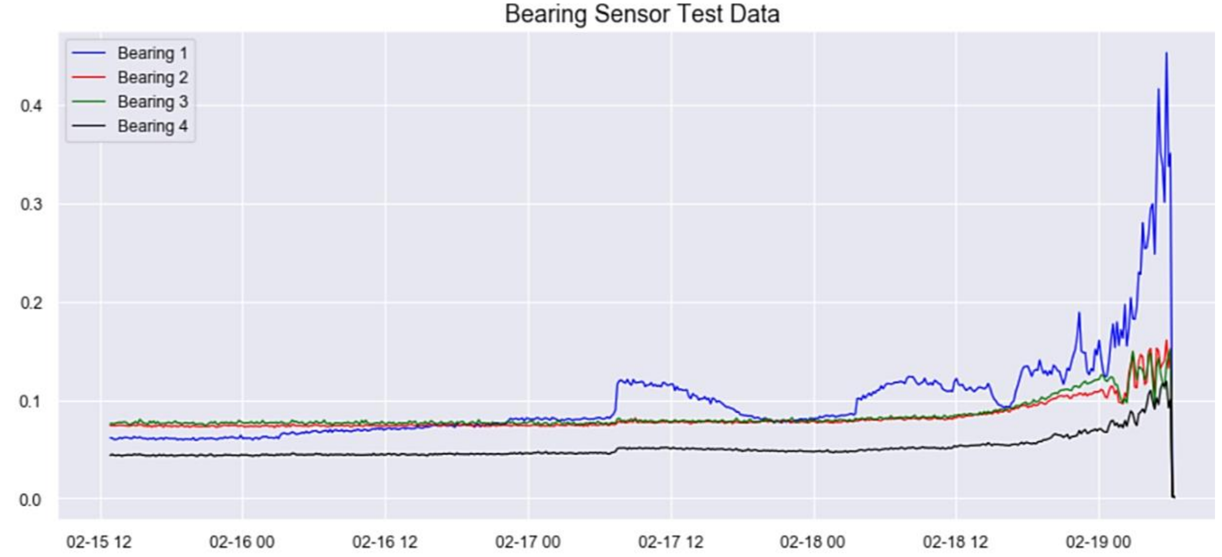


- Tree model accuracy = 94%

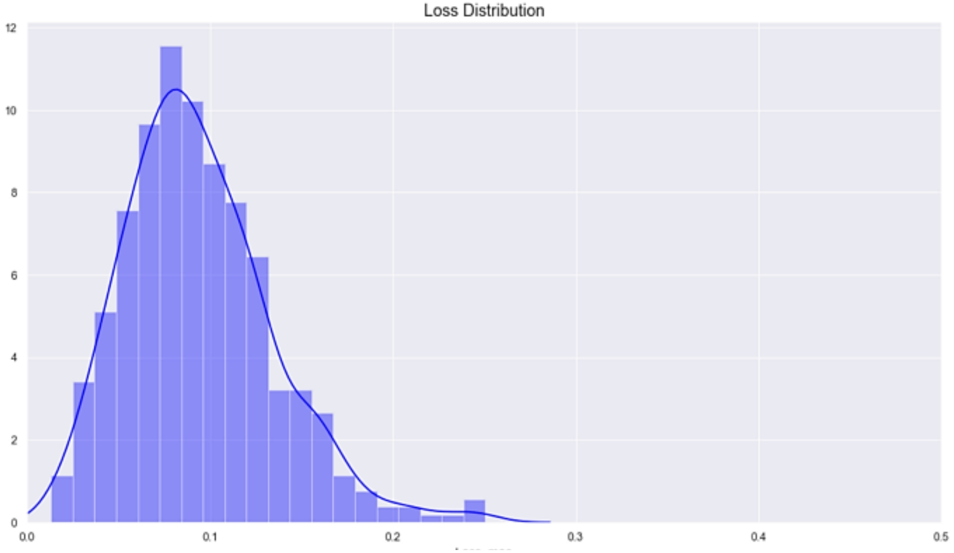
https://se.mathworks.com/help/predmaint/ug/broken-rotor-fault-detection-in-ac-induction-motors-using-vibration-and-electrical-signals.html

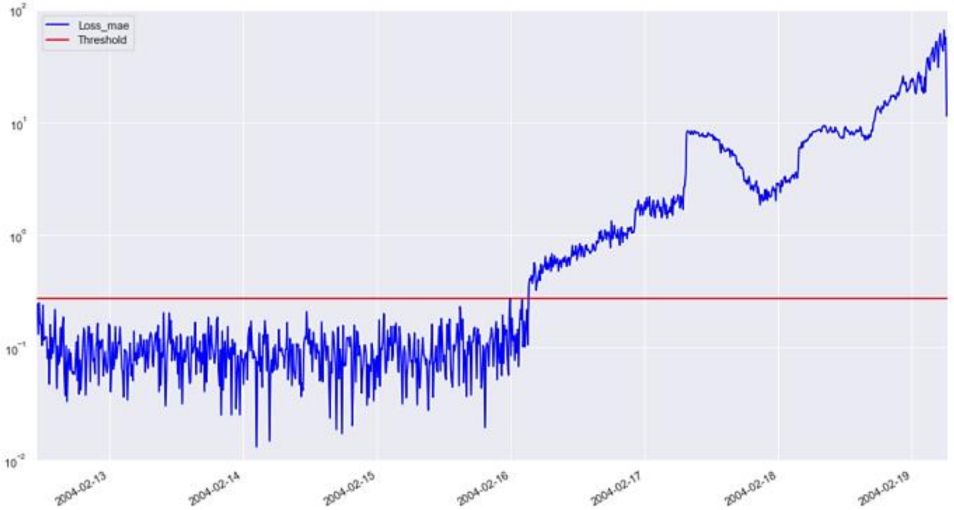
Методы диагностики неисправностей: примеры

Bearing Failure Anomaly Detection (Python Keras and TensorFlow)



	Bearing 1	Bearing 2	Bearing 3	Bearing 4
2004-02-12 10:52:39	0.060236	0.074227	0.083926	0.044443
2004-02-12 11:02:39	0.061455	0.073844	0.084457	0.045081
2004-02-12 11:12:39	0.061361	0.075609	0.082837	0.045118
2004-02-12 11:22:39	0.061665	0.073279	0.084879	0.044172
2004-02-12 11:32:39	0.061944	0.074593	0.082626	0.044659





- LSTM network - RNN который обрабатывает входные данные, перебирая временные шаги и обновляя состояние сети. Содержит информацию за все предыдущие временные шаги. Можно использовать для прогнозирования последующих значений временного ряда, используя предыдущие временные шаги в качестве входных данных.

https://github.com/BLarzalere/LSTM-Autoencoder-for-Anomaly-Detection

Методы диагностики неисправностей: примеры

Vladislav Zaitsev 192296IABM

CFB katla õigeaegne lekke tuvastamine masinaõpe abil

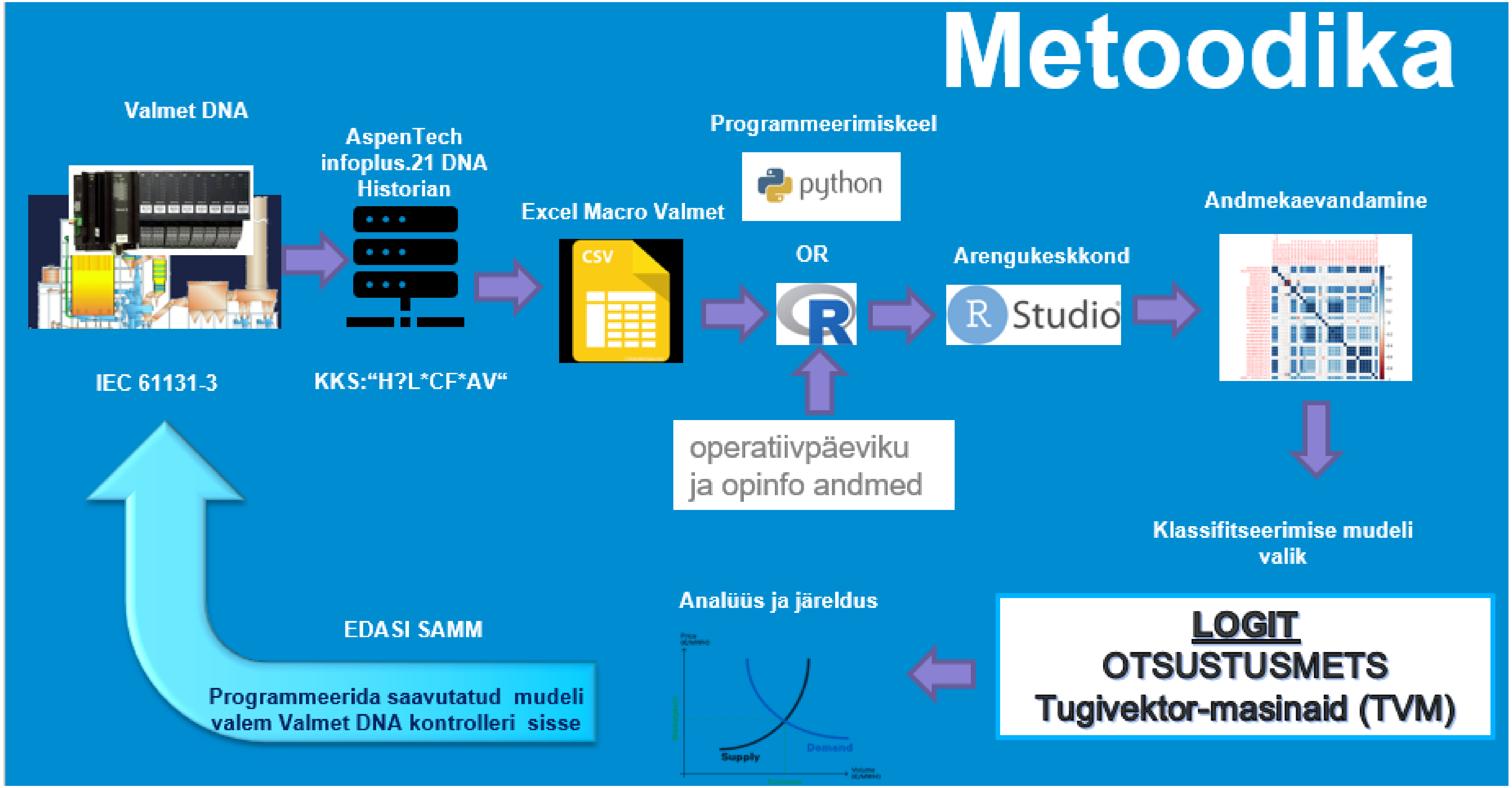
Magistritöö

Juhendaja: Olga Ruban

PhD

Методы диагностики неисправностей: примеры

CFB katla õigeaegne lekke tuvastamine masinaõppe abil



Методы диагностики неисправностей: примеры

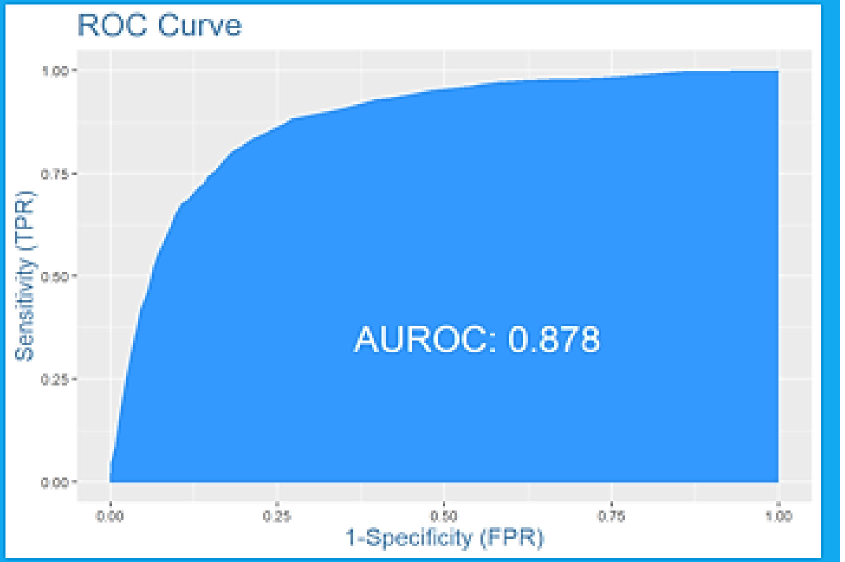
CFB katla õigeaegne lekke tuvastamine masinaõppe abil

Logit mudel

Valem:

H1LEKE = -0.091570342
+ H1LAE50CF201.av * -2.594089826 +
+ H1LAF10CF201.av * -1.01012989 +
+ H0LCJ30CF201.av * 0.267352318 +
+ H2LAF10CF201.av * -3.026367989 +
+ H0LBC55CF201.av * 0.487359665 +
+ H0LBG01CF901 * -0.017021171 +
+ H1LCQ10CF201.av * -1.95875418 +
+ H1LBB10CF901.av * -0.632698247 +
+ H1LAE60CF201.av * 1.052166265 +
+ H0MKA__CE203XQ13.av * 0.00096767 +
+ H0LC+ H10CF201.av * -0.117488671 +
+ H0LCE20CF201.av * 0.495062174 +
+ H1LAB80CF901.av * -0.03446414 +
+ H0LCA40CF901.av * 0.06180533 +
+ H0LAC00CF901XQ12.av * 0.125157848 +
+ H2LBA10CF901.av * 0.028837113 +
+ H1LBA10CF901.av * -0.285148399 +
+ H2LAB80CF901.av * -0.522211265 +
+ H2LBB10CF901.av * -0.237361717 +
+ H2LAE60CF201.av * -1.152462154 +
+ H2LAE50CF201.av * 2.582788892

Logit		Tegelik	
		1	0
Mudeli tulemus	1	1059	230
	0	286	1153
	Täpsus:	0.811	



Методы диагностики неисправностей: примеры

Dmitri Poljakov 192460IABM

Application of Machine Learning Methods to Industrial Equipment Fault Detection

Master's thesis

Supervisor: Olga Dunajeva

PhD

Методы диагностики неисправностей: примеры

Application of Machine Learning Methods to Industrial Equipment Fault Detection

Task:

- Develop a methodology for detecting faults and technology for its implementation on the example of the enterprise Enefit Power AS.
- The oil plant Enefit 280 turbine was chosen as the object of the research.

The samples contain the sensor measurements as:

- turbine speed
- control signal to control valve 1
- control signal to control valve 2
- control signal to control valve 4
- feedback signal from valve 1
- feedback signal from valve 2
- feedback signal from valve 3
- output signal of the turbine controller
- electrical load of the turbine
- turbine power regulator mismatch
- turbine speed regulator mismatch

Методы диагностики неисправностей: примеры

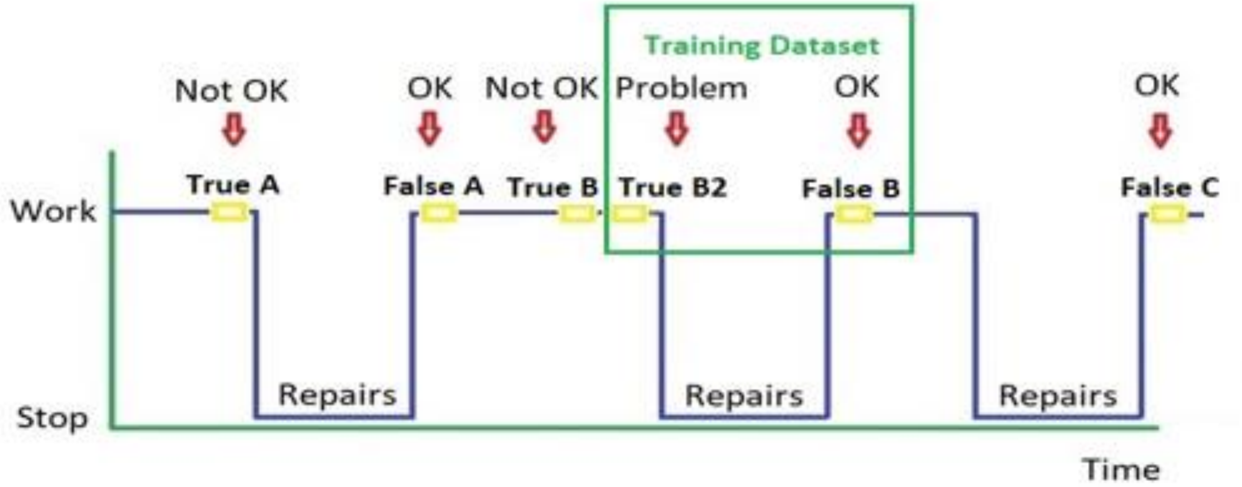
Application of Machine Learning Methods to Industrial Equipment Fault Detection

Datasets creation

The samples were collected for the period:

- from 26/11/2015 + week, 12 variables and 10003 lines
- from 08/01/2016 + week, 12 variables and 10080 lines
- from 06/07/2017 + week, 12 variables and 10080 lines
- from 19/07/2017 + week, 12 variables and 8641 lines
- from 23/12/2017 + week, 12 variables and 10080 lines
- from 20/01/2019 + week, 12 variables and 10080 lines

Wear-out equipment state
True (1) Class
Mechanical probleem,
True (1) Class
Healthy equipment state,
False (0) Class



Методы диагностики неисправностей: примеры

Application of Machine Learning Methods to Industrial Equipment Fault Detection

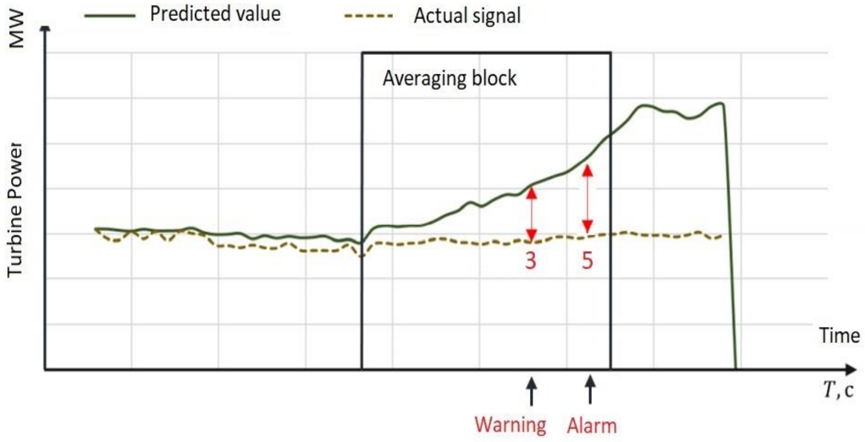
- **Logistic regression model results**

Logistic regression model metrics			
	Training dataset	Test dataset 20/01/2019 + on week	Test dataset 06/07/2017 + one week
Correctly classified values	99.98%	100.00%	88.71%
Incorrectly classified values	0.01%	0.00%	11.28%
Precision	1	1	1
Recall	1	1	0.88
F-Measure	1	1	0.94
Confusion matrix	a b <-- classified as 10079 1 | a = False 1 8640 | b = True	a b <-- classified as 10080 0 | a = False 0 0| b = True	a b <-- classified as 0 0 | a = False 1138 8942| b = True

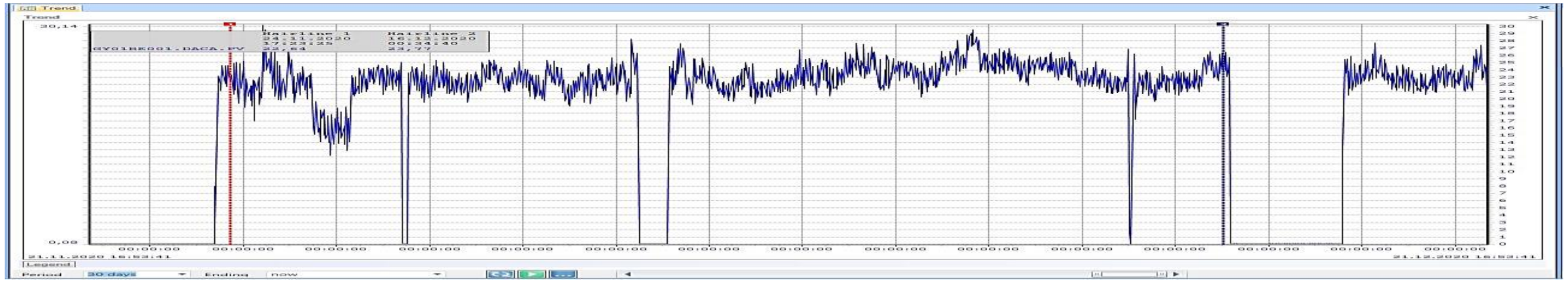
Методы диагностики неисправностей: примеры

Method for detecting anomalies on equipment

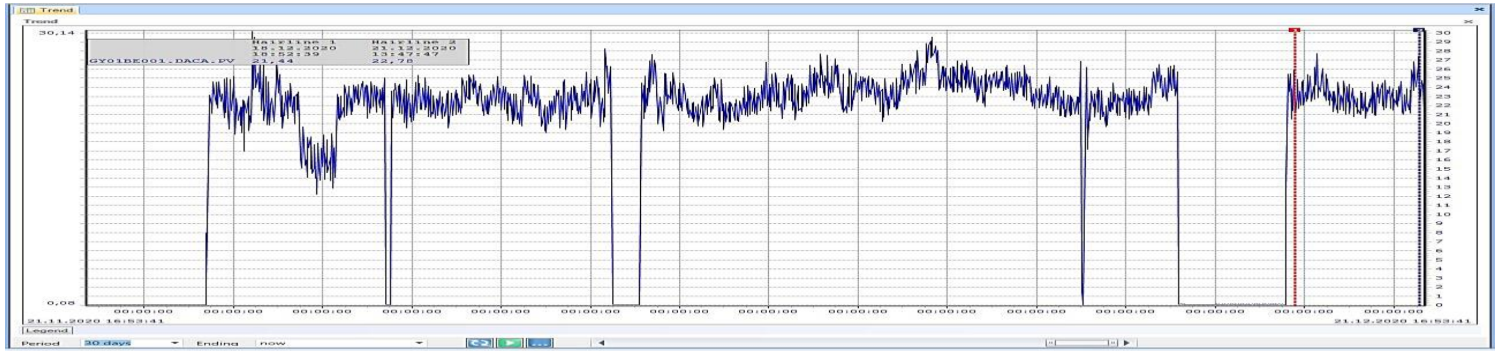
- The main goal is to create such a ML model that will predict the good condition of the equipment. Moreover, any deviation of the predicted value from the normal will be perceived as a malfunction.
- All models predict the turbine load value Y (GX01BV001) based on the mechanical measurements (bearings temperature, rotor and generator vibration, lubrication oil temperature, cooling air temperature, etc.).



- training dataset - data for 3 weeks



- test dataset - data for 4 days



Методы диагностики неисправностей: примеры

Method for detecting anomalies on equipment

Machine learning models

Basic linear regression model after the cross-validation function	Test dataset from 18/12/2020 to 21/12/2020	0.64	89.64%
Additionally, trained linear regression model.	Test dataset from 18/12/2020 to 21/12/2020	0.57	90.79%
Model Random Forest	Test dataset from 18/12/2020 to 21/12/2020	0.85	86%
Neural Network	Test dataset from 18/12/2020 to 21/12/2020	1.24	87%
Nonlinear regression model	Test dataset from 18/12/2020 to 21/12/2020	0.72	91%
Model MARS	Test dataset from 18/12/2020 to 21/12/2020	0.64	89.65%
Model PPR	Test dataset from 18/12/2020 to 21/12/2020	0.52	94.7%

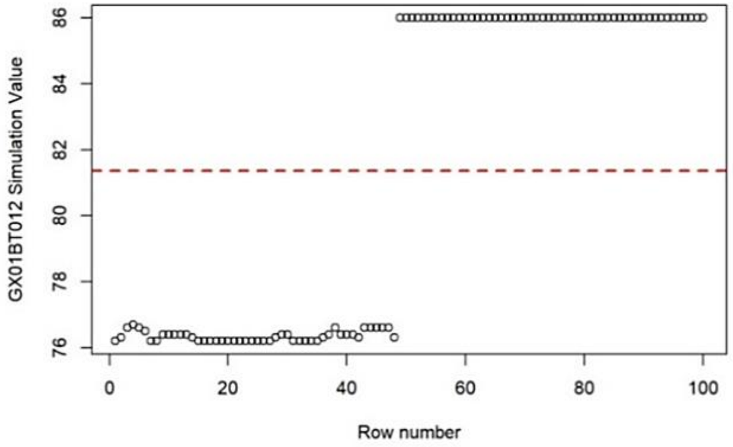
Методы диагностики неисправностей: примеры

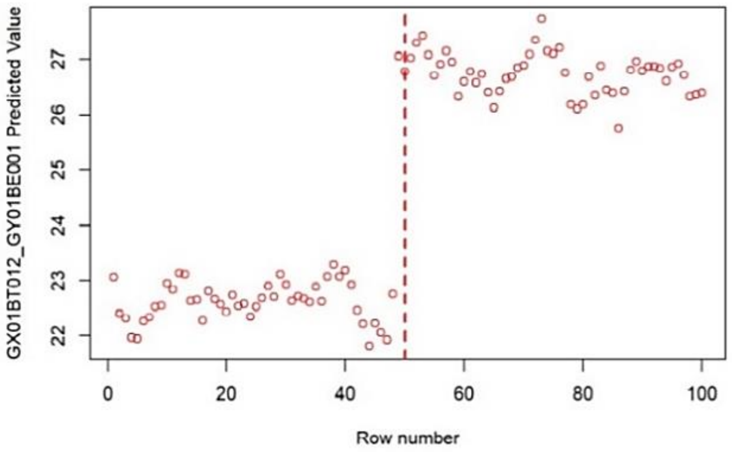
Method for detecting anomalies on equipment

Process equipment fault simulation

Several experiments were performed to activate the malfunctions occurrence on equipment by simulating some attributes in the test data.

- The model predicted values reacted to the bearing temperature change from 76 to 86 degrees by increasing the load from 22.7 MW to 26.7 MW (**4.36** MW).
- The model predicted values responded to the vibration level change from 0.3 to 2 mm/s by increasing the load from 23.2 MW to 25.3 MW (**2.67** MW).
- The increase in bearing temperature joined with increased vibration levels, increased the predicted load from 22.8 MW to 29.2 MW (**6.85** MW). The purpose of this experiment was to test the hypothesis of the cumulative effect of two or more attributes on the model's predicted values.

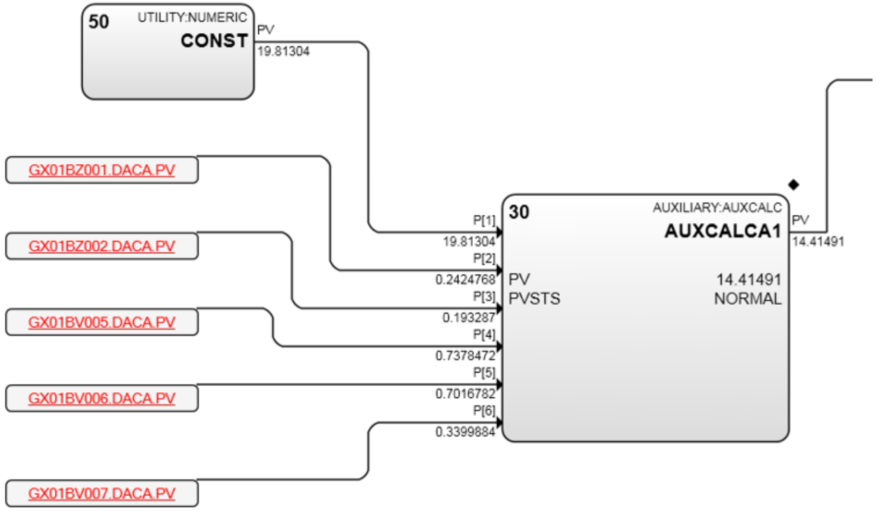




Методы диагностики неисправностей: примеры

Linear regression model integration into the enterprise management system

- At the oil plant, the industrial automated system Honeywell is used to control the process.
- For integration into the enterprise automation system, an additionally trained linear regression model was selected
- AUXCALC mathematical modules were used for calculations. This module has restrictions on the number of inputs, so the general formula was divided into several fragments.



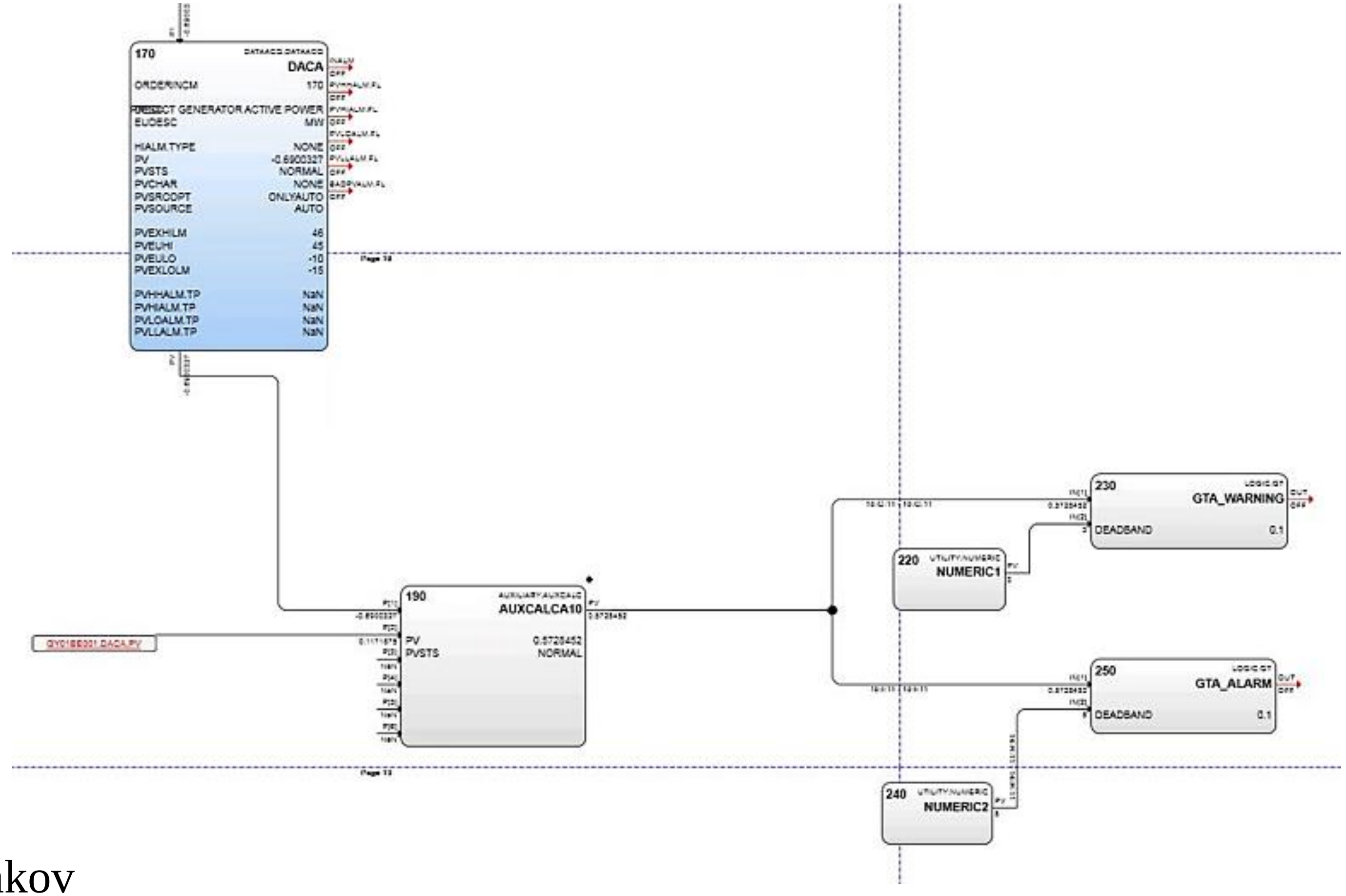
The formula for the AUXCALCA1 module is as follows:

P[1] + P[2]*(12.170114) + P[3]*(-11.290135) + P[4]*(-3.028353) + P[5]*(-0.745362) + P[6]*(-10.027987)

Методы диагностики неисправностей: примеры

Linear regression model integration into the enterprise management system

At the output of the module, was formed absolute deviation between predicted and actual values. Two comparison modules were organized to generate a warning (3 MW) and alarm signals (6 MW) for the enterprise operating personnel.



Методы диагностики неисправностей: примеры

Linear regression model integration into the enterprise management system

After the calculation, the predicted values of the turbine load were recorded on the enterprise information server. Comparative graphs of the actual (blue graph) and predicted (green graph) turbine load.