MASTER'S DEGREE EXAMINATION

Study major: Advanced Analytics – Big Data *(inf. 23/24)*

1. Present how to join multiple tables – describe possible methods.
2. Describe the single row functions classification.
3. The database objects - their roles, purposes, methods of using.
4. The views. Why are they created? What are the possible clauses in a statement that create a view?
5. The syntaxes of set statements. What are the set operators and the results of their use?
6. The subqueries. Describe types of subqueries, possible clauses they may be used, possible operators.
7. Describe typical solutions Big Data provides in the area of data storage.
8. Describe the meaning of 3V and 5V in the context of Big Data.
9. Discuss ethical issues related to Big Data.
10. Evaluate capabilities and specific characteristics of analytical environments used in Big Data.
11. Please describe in detail one chosen algorithm used in Big Data analytics.
12. What is MapReduce and how does it work?
13. What is Deep Learning, give an example.
14. What are the typical characteristics of Big Data problems?
15. What is data variability and how to take it into account in data visualization?
16. Discuss examples of pattern recognition techniques used in Big Data.
17. Define and describe distributed computing, in particular, in context of Big Data.
18. Describe a selected methodology describing a method of execution of development process of analytical models.
19. Outline key assumptions that are conditions of application of predictive models in support of decision making processes.
20. Describe how usage of version control systems influences the effectiveness of analytical solution development process.
21. Explain what is meant by the term reproducibility of analytical process and why it is important in business.
22. Describe most important methods of ensuring reproducibility of analytical process.
23. Explain what does the term cutoff threshold mean in classification models and describe what are factors that influence its optimal value in case when such a model is used for supporting decision making.
24. Explain how regularization in used in the process of building of predictive models.
25. Explain the difference between observational, interventional and counterfactual reasoning.
26. Explain Simpson's paradox.
27. List and discuss methods of visualization of spatial data
28. Economic gains from processing data in the cloud.
29. Present serverless computing in gathering and processing data for analytics.
30. Describe storing big data in the cloud.

31. Describe scaling document-oriented databases in the cloud - the case of DynamoDB.
32. Describe scaling analytical processes in the cloud.
33. Present Function as a service - data processing model based on the Lambda architecture.
34. Specify and discuss methods for visualizing proportions.
35. Present creating and managing security of analytical platforms in the cloud for Python and R.
36. Present managing security, users and access rights in the cloud - users, roles, policies and groups.
37. Present managing a relational database in the cloud and applications for data analytics.
38. Present data processing models for the cloud: IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) and SaaS (Software-as-a-Service).
39. Discuss the data properties relevant to the data analysis process.
40. What is the importance of the context in data analysis?
41. What is the uncertainty in data analysis and how can it be influenced?
42. What is the importance of metadata in data analysis?
43. Specify and discuss the coordinate systems used for data visualisation.
44. Specify and discuss methods for visualizing time series.
45. Specify and discuss methods of relationship visualization.
46. What descriptive statistics are robust on outliers?
47. Explain what a distributed version control system is using Git as an example. Propose a typical simple workflow.
48. Discuss a selected data dimension reduction technique, its strong and weak points.
49. Discuss the parallel computation concept and typical problems of parallel computations.
50. What is a robust estimator? Discuss using a selected example.
51. Discuss regularization techniques using a selected example, e.g., LASSO regression.
52. Explain the concepts of structured and unstructured data.
53. Introduce the Lambda and Kappa architectures.
54. Present the key features of learning and prediction in batch (offline learning) and incremental (online learning) modes.
55. Give an example and discuss in what situations it is advisable to use the OLTP processing model.
56. Give an example and discuss in what situations it is advisable to use the OLAP processing model.
57. Explain the concept and business applications of a data warehouse.
58. Describe the problem of time in streaming data processing, what is watermark.
59. Describe the difference between data stream and batch processing.
60. Describe two business applications of real-time data analysis.
61. List and describe methodologies of data mining process.
62. Describe two main groups of data mining methods.
63. Describe the methods of feature selection and sampling for data mining modeling.
64. Data classification methods - present differences and similarities between them.
65. Describe decision tree models.
66. Describe random forest models.

67. Describe models of artificial neural networks.
68. Describe methods of data clustering.
69. Describe methods of transactional data analysis.
70. Discuss the methods for constructing life tables and provide examples of their application.
71. Compare non-parametric and parametric models in duration analysis.
72. Characterise proportional hazard models and provide examples of such models.
73. Characterise accelerated failure time models and provide examples of such models.
74. Characterise semi-parametric models in duration analysis.
75. List the differences between the classical and Bayesian approaches in the context of parameter estimation for duration analysis models.
76. Discuss competing risk models in duration analysis.
77. Discuss the idea of Markov chains Monte Carlo (MCMC) methods in the context of parameter estimation for duration analysis models.
78. Data quality in business analytics. The meaning and assessment techniques.
79. Data imputation. The importance and meaning.
80. Multiple imputation: description of the method, selection of the imputation model and estimation of the parameters.
81. ~~Compare fixed and random effects models. Indicate basic differences and provide examples of applications.~~ *– question deleted 24.11. 2025*
82. Quantile regression: description and applications in business analytics.
83. Adaptive regression: the model, estimation technique and applications in business analytics.
84. K-means method and its application in Customer Lifetime Value CLV models.
85. Name and describe business applications of Customer Lifetime Value CLV models.
86. What descriptive statistics are not affected by outliers?
87. What descriptive statistics should be used for samples taken from populations with a distribution other than the normal?
88. Describe the Information Security triad: Confidentiality, Integrity, and Availability.
89. What is Spear Phishing?
90. Describe basic Cybersecurity principles for SMEs (Small and Mid-size Enterprises).
91. What is the interpretation of a programming language? Give examples of interpreted languages and interpreters.
92. Describe the installation of libraries (packages) in the Python environment. Give examples of popular libraries.
93. Describe iteration techniques using a chosen programming language, e.g., R or Python.
94. Describe the concept of a function and scoping using a chosen programming language, e.g., R or Python.
95. What is a decision engine? List the rules of the credit acceptance process implemented in the decision engine.
96. Discuss the concepts related to data preparation and the modeling event: observation point, data period and observation period, list the most common modeling errors (e.g. taking data from the future) and problems with selecting the length of both periods.
97. Discuss an example of scorecard, how are partial scores calculated, how is its form interpreted?

98. How do we calculate the profitability of the credit acceptance process? What role does the scoring model play in this?
99. What is Reject Inference analysis?
100. Discuss the impact of the human factor on the credit acceptance process. Is it possible to increase sales and reduce credit risk at the same time?

Literature:

1. J. Price, Oracle Database 12c i SQL. Programowanie, Helion 2015;
2. J. Ullman, J. Widom, Podstawowy kurs baz danych Wyd. III, Helion 2011;
3. A. Alapati, D. Kuhn, B. Padfield, Oracle 12c. Problemy i rozwiązania, Helion 2014;
4. https://docs.oracle.com/database/121/SQLRF/toc.htm
5. Mayer-Schönberger V., Cukier K.: Big data: rewolucja, która zmieni nasze myślenie, pracę i życie: efektywna analiza danych; Warszawa: MT Biznes, 2017;
6. Surma J., Cyfryzacja życia w erze Big Data: człowiek, biznes, państwo /Warszawa: Wydawnictwo Naukowe PWN. 2017;
7. Inc, O.M., 2012. Big Data Now: 2012 Edition 2. wyd., O'Reilly Media;
8. Hand D., Mannila H., Smyth P. ,,Eksploracja danych", WNT Wydawnictwa Naukowo-Techniczne, 2005;
9. White T., Hadoop: kompletny przewodnik: analiza i przechowywanie danych /; Gliwice: Helion, cop. 2016;
10. J. Gareth, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2013;
11. B. Kamiński: The Julia Express, http://bogumilkaminski.pl/files/julia_express.pdf;
12. B. Kamiński: Julia DataFrames Tutorial, https://github.com/bkamins/Julia-DataFrames-Tutorial;
13. M. Wittig, A. Wittig. Amazon web services in action, 2nd edition. Manning, 2018;
14. J. Baron, H. Baz, T. Bixler, B. Gaut, K. E. Kelly, S. Senior, J. Stamper. AWS certified solutions architect official study guide: associate exam. John Wiley & Sons, 2016;
15. Amazon (2016) Getting Started with AWS, wersja elektroniczna do pobrania za darmo w sklepie amazon.com;
16. Amazon (2009) The Economics of the AWS Cloud vs. Owned IT Infrastructure, do pobrania ze strony https://aws.amazon.com/whitepapers/;
17. Amazon (2016) Amazon Elastic Compute Cloud (EC2) User Guide for Linux Instances, wersja elektroniczna do pobrania za darmo w sklepie amazon.com;
18. Introduction to AWS Economics, do pobrania ze strony https://aws.amazon.com/whitepapers/;
19. Big Data Analytics Options on AWS, do pobrania ze strony https://aws.amazon.com/whitepapers/;
20. Introduction to High Performance Computing on AWS, do pobrania ze strony https://aws.amazon.com/whitepapers/;
21. Introduction to AWS Security, do pobrania ze strony https://aws.amazon.com/whitepapers/;

22. Kamiński, B., & Szufel, P. (2015). On optimization of simulation execution on Amazon EC2 spot market. Simulation Modelling Practice and Theory, 58, 172-187;
23. D.T. Larose, Data Mining Methods and Models, Wiley, New York 2006;
24. J. Koronacki, J. Ćwik, Statystyczne systemy uczące się, WN-T, Warszawa 2005;
25. M. Lasek, M. Pęczkowski, Enterprise Miner: wykorzystywanie narzędzi Data Mining w systemie SAS, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2013;
26. R. Matignon, Data Mining Using SAS Enterprise Miner, Wiley, Hoboken, NJ 2007;
27. F. Provost, T. Fawcett, Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly, USA 2013;
28. T. Morzy, Eksploracja danych, Metody i algorytmy, PWN, Warszawa 2013;
29. N. Yau, Data points: visualization that means something, Indianapolis, Ind. Wiley, 2013;
30. N.C. Yau, Visualize this the FlowingData guide to design, visualization, and statistics, Indianapolis, Ind. Wiley 2011;
31. J. Maindonald, Data analysis and graphics using R': an example-based approach, Cambridge UK, New York: Cambridge University Press, 2003;
32. Frątczak E. (red.) Zaawansowane Metody Analiz Statystycznych, SGH, Warszawa 2012;
33. Allison P. D., Logistic Regression Using SAS: Theory and Application, Second Edition. Cary, NC: SAS Institute Inc., 2012;
34. Hosmer D. W., Jr., Lemeshow S., Sturdivant R. X., Applied Logistic Regression, Third Edition, John Wiley & Sons, 2013;
35. Kleinbaum D. G., Klein M., Logistic Regression: A Self-Learning Text, Third Edition, Springer, 2010;
36. Stanisz A., Modele regresji logistycznej. Zastosowania w medycynie, naukach przyrodniczych i społecznych. StatSoft Polska, Kraków, 2016;
37. Korczyński A., Screening wariancji jako narzędzie wykrywania zmowy cenowej. Istota i znaczenie imputacji danych, Oficyna wydawnicza SGH, Warszawa, 2018;
38. Frątczak E. red. Zaawansowane Metody Analiz Statystycznych, SGH, Warszawa 2012;
39. Little A, Rubin D., Statistical Analysis with Missing Data. John Wiley & Sons: Hoboken 2002;
40. Malthouse E.C., Segmentation and Lifetime Value Models Using SAS, SAS Institute, 2013;
41. Svolba G., Applying Data Science. Business Case Studies, SAS Institute: Cary, NC, 2017;
42. W. Grzenda, A. Ptak-Chmielewska, K. Przanowski, U. Zwierz. Przetwarzanie danych w SAS, Oficyna Wydawnicza SGH, 2012;
43. SAS programming by example, Ron Cody and Ray Pass, SAS Publishing;
44. Zdzisław Dec, Wprowadzenie do systemu SAS, Wydawnictwo Editio, 2000;
45. Jordan Bakerman, SAS® Programming for R Users. SAS Institute Inc. 2019.Cary, NC: SAS Institute Inc. Copyright © 2019, SAS Institute Inc.;
46. Jóźwiak J., Podgórski J.: Statystyka od podstaw, PWE, Warszawa;
47. Przanowski K., 2014, Credit Scoring w erze Big-Data, Oficyna Wydawnicza SGH;
48. Daniel Kaszyński, Bogumił Kamiński and Tomasz Szapiro, Credit scoring in the context of interpretable machine learning, 2020;
49. Siddiqi N., 2005. Credit risk scorecards: Developing and implementing intelligent credit scoring. Wiley and SAS Business Series.