MAGISTERSKI EGZAMIN DYPLOMOWY
Kierunek: Analiza danych – Big data

1. Przedstaw sposoby agregacji danych.
2. Omów mechanizmy łączenia danych z wielu tabel.
3. Kiedy należy stosować funkcje działające na pojedynczych wierszach, a kiedy funkcje grupowe? Na jakich typach danych działają?
4. Omów klasyfikację funkcji działających na pojedynczych wierszach.
5. Jakie znasz polecenia zmieniające zawartość tabeli? Jakie są ich skutki oraz zakres oddziaływania?
6. Jaką rolę pełni Data Dictionary (Słownik Danych) i jak się nim posługiwać?
7. W jakim celu buduje się perspektywy? Omów możliwe klauzule polecenia do tworzenia perspektyw.
8. Operacje na zbiorach – omów składnię poleceń i znaczenie uzyskanych wyników.
9. Przedstaw podzapytania – typy, klauzule, w których mogą wystąpić, operatory.
10. Omów typowe rozwiązania Big Data w obszarze baz/repozytoriów danych.
11. Przedstaw specyfikę środowisk analitycznych stosowanych w Big Data.
12. Omów wybrany algorytm stosowany w analityce Big Data.
13. Na czym polega MapReduce?
14. Co to jest Deep Learning, podaj przykład.
15. Jakimi cechami charakteryzują się typowe problemy Big Data?
16. Dokonaj oceny mocy predykcyjnej modelu regresji logistycznej.
17. Omów przykładowe techniki stosowane w rozpoznawaniu wzorców.
18. Na czym polega przetwarzanie rozproszone?
19. Omów wybraną metodykę opisującą sposób realizacji procesu wytwórczego modelu analitycznego.
20. Wymień kluczowe założenia będące warunkami zastosowania modeli predykcyjnych do wspomagania procesów decyzyjnych.
21. Jak mierzymy jakość modelu prognostycznego?
22. Omów w jaki sposób wykorzystanie systemu kontroli wersji wpływa na efektywność procesu wytwórczego rozwiązań analitycznych.
23. Omów model proporcjonalnych szans
24. Wyjaśnij co to jest reprodukowalność procesu analitycznego i dlaczego jest ona ważna w praktyce gospodarczej.
25. Omów podstawowe sposoby zapewnienia reprodukowalności procesu analitycznego.
26. Wyjaśnij co to jest próg odcięcia w modelach klasyfikacyjnych oraz omów od czego zależy jego optymalna wartość w przypadku wykorzystania takiego modelu do wspomagania podejmowania decyzji.
27. Wyjaśnij do czego wykorzystywana jest regularyzacja w procesie budowy modeli predykcyjnych.
28. Wyjaśnij różnicę, pomiędzy wnioskowaniem obserwacyjnym, interwencyjnym i kontrfaktycznym.
29. Wyjaśnij na czym polega paradoks Simpsona.
30. Przedstaw korzyści ekonomiczne z przetwarzania danych w chmurze.
31. Omów technologie serveless w gromadzeniu i przetwarzaniu danych na potrzeby procesów analitycznych.
32. Przedstaw metody przechowywania danych dużych rozmiarów w chmurze.

33. Omów skalowanie dokumentowych baz danych typu noSQL w chmurze na przykładzie DynamoDB.
34. Omów skalowanie procesów analitycznych w chmurze.
35. Omów Function as a service - model przetwarzania oparty o architekturę Lambda.
36. omów tworzenie i zarządzanie bezpieczeństwem środowisk analitycznych dla języków Python i R w chmurze.
37. Omów zarządzanie bezpieczeństwem, użytkownikami i prawami dostępu w chmurze - użykownicy, role, polityki i grupy.
38. Przedstaw systemy zarządzania relacyjną bazą danych w chmurze i ich zastosowania w analityce danych.
39. Przedstaw modele sztucznych sieci neuronowych.
40. Przedstaw modele przetwarzania danych w chmurze: IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) oraz SaaS (Software-as-a-Service).
41. Wymień i omów metodyki procesu eksploracji danych.
42. Omów dwie główne grupy metod eksploracji danych.
43. Omów metody selekcji zmiennych i obserwacji do modelowania data mining.
44. Przedstaw metody klasyfikacji danych – przedstaw różnice i podobieństwa pomiędzy nimi.
45. Omów kwestie etyczne związane z Big Data.
46. Przedstaw model drzewa decyzyjnego.
47. Omów modele lasów losowych.
48. Dokonaj oceny mocy predykcyjnej modeli klasyfikacyjnych.
49. Przedstaw metody grupowania danych.
50. Omów metody analizy danych transakcyjnych.
51. Omów cechy danych istotne w procesie analizy danych.
52. Przedstaw, na czym polega zmienność danych i jak ją uwzględnić w wizualizacji danych.
53. Przedstaw, na czym polega niepewność w analizie danych i jak można wpływać na jej wielkość.
54. Jakie znaczenie mają metadane w analizie danych.
55. Wymień i omów układy współrzędnych stosowane przy wizualizacji danych.
56. Wymień i omów metody wizualizacji proporcji.
57. Wymień i omów metody wizualizacji relacji.
58. Wymień i omów metody wizualizacji danych geolokalizacyjnych.
59. Czy modele regresji logistycznej należą do klasy uogólnionych modeli liniowych? Odpowiedź uzasadnij.
60. Wymień obiekty bazy danych i omów ich przeznaczenie.
61. Dokonaj estymacji modeli regresji logistycznej.
62. Wykonaj interpretację dopasowanego modelu regresji logistycznej.
63. Przedstaw wnioskowanie statystyczne w regresji logistycznej.
64. Wymień i omów metody wizualizacji szeregów czasowych.
65. Dokonaj oceny dobroci dopasowania modelu regresji logistycznej.
66. Dokonaj diagnostyki modelu regresji logistycznej.
67. Przedstaw, na czym polega uwzględnienie kontekstu a analizie danych.
68. Omów wielomianową regresję logistyczną.
69. Porównaj model proporcjonalnych szans z modelem wielomianowej regresji logistycznej.

70. Korzyści ekonomiczne, biznesowe z analityki z wykorzystaniem modeli AHZ (modeli analizy przeżycia).
71. Przedstaw istotę: filozofia – statystyka – matematyka (w tym podstawowe pojęcia) modelu pojedynczego epizodu AHZ.
72. Omów modele AHZ o czasie dyskretnym versus modele AHZ o czasie ciągłym (w tym zasady konstrukcji baz danych do obu typów modeli).
73. Tradycyjne modele regresji versus modele regresji AHZ – przedstaw podobieństwa i różnice w teorii, diagnostyce i obszarach zastosowań.
74. Przedstaw podstawowe procedury do estymacji modeli AHZ w SAS i dostępnym oprogramowaniu OPEN SOURCE.
75. Przedstaw podstawy teoretyczne – aplikacje – diagnostyka i interpretacja wyników modeli nieparametrycznych.
76. Przedstaw podstawy teoretyczne – aplikacje – diagnostyka i interpretacja wyników modeli parametrycznych.
77. Przedstaw podstawy teoretyczne – aplikacje – diagnostyka i interpretacja wyników modeli semiparametrycznych.
78. Omów semiparametryczne modele ryzyk konkurencyjnych (porównanie modeli: Cox'a i Fine-Gray'a).
79. Omów zaawansowane modele AHZ & modele CLTV. Predykcja na bazie modeli AHZ.
80. Jakość danych w analizach biznesowych – omów znaczenie i metody oceny.
81. Imputacja danych. Omów istotę i znaczenie.
82. Zdefiniuj i opisz proces modelowania predykcyjnego.
83. Podaj i omów wybraną miarę mocy predykcyjnej modelu statystycznego.
84. Wyjaśnij co to jest system kontroli wersji na przykładzie systemu Git i zaproponuj typowy workflow.
85. Omów wybraną technikę redukcji wymiaru danych, jej zalety i wady.
86. Omów modele sztucznych sieci neuronowych na przykładzie wybranej topologii sieci neuronowej.
87. Omów pojęcie obliczeń równoległych i podstawowe problemy, które pojawiają się przy obliczeniach równoległych.
88. Omów pojęcie estymatora odpornego na wybranych przykładzie.
89. Omów technikę regularyzacji na wybranym przykładzie, np. regresji LASSO.
90. Przedstaw metody łączenia tabel w SAS i SQL.
91. Przedstaw plusy i minusy przetwarzania danych w SAS i SQL.
92. Na czym polega makroprogramowaie w SAS?
93. Co to jest biblioteka w systemie SAS?
94. Przedstaw przykłady procedur Base SAS i SAS/STAT.
95. Jakie statystyki opisowe są odporne na wartości nietypowe?
96. Które statystyki opisowe są właściwsze dla rozkładów, które nie są normalne?
97. W jaki sposób można zweryfikować, czy dany rozkład jest zgodny z rozkładem normalnym?
98. Przedstaw plusy i minusy struktur danych: analitycznej i transakcyjnej.
99. Co to jest PDV i sekwencyjne przetwarzanie danych w SAS?
100. Co oznacza określenie 3V oraz 5V w kontekście problematyki Big Data?

Literatura:

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. Frątczak E., ,U. Sienkiewicz, H. Babiker, Analiza historii zdarzeń. Teoria, przykłady zastosowań z wykorzystaniem programów: SAS, TDA, STATA. SGH, Warszawa wyd. 2017;
38. Borucka J., Analiza i modelowanie ryzyka zachorowalności. Parametryczne i semiparametryczne modele przeżycia, SGH, 2017;
39. Allison P. , Survival Analysis Using SAS: A Practical Guide, Second Edition, 2010;
40. Broström G. Event History Analysis with R, Series: Chapman & Hall/CRC The R Series, CRC Press, 2012;
41. Crowder M., Multivariate Survival Analysis and Competing Risks. Chapman & Hall/CRC Texts in Statistical Science, 2012;
42. Elasthoff R. M., Gang Li, Ning Li. Joint Modelling of Longitudinal and Time to Event Data. Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 2016;
43. Xian Liu, Survival Analysis. Models and Applications. Wiley, 2013;
44. Korczyński A., Screening wariancji jako narzędzie wykrywania zmowy cenowej. Istota i znaczenie imputacji danych, Oficyna wydawnicza SGH, Warszawa, 2018;
45. Frątczak E. red. Zaawansowane Metody Analiz Statystycznych, SGH, Warszawa 2012;
46. Little A, Rubin D., Statistical Analysis with Missing Data. John Wiley & Sons: Hoboken 2002;
47. Malthouse E.C., Segmentation and Lifetime Value Models Using SAS, SAS Institute, 2013;
48. Svolba G., Applying Data Science. Business Case Studies, SAS Institute: Cary, NC, 2017;
49. W. Grzenda, A. Ptak-Chmielewska, K. Przanowski, U. Zwierz. Przetwarzanie danych w SAS, Oficyna Wydawnicza SGH, 2012;
50. SAS programming by example, Ron Cody and Ray Pass, SAS Publishing;
51. Zdzisław Dec, Wprowadzenie do systemu SAS, Wydawnictwo Editio, 2000;
52. Jordan Bakerman, SAS® Programming for R Users. SAS Institute Inc. 2019.Cary, NC: SAS Institute Inc. Copyright © 2019, SAS Institute Inc.;
53. Jóźwiak J., Podgórski J.: Statystyka od podstaw, PWE, Warszawa.