MAGISTERSKI EGZAMIN DYPLOMOWY

Kierunek: Analiza danych – Big Data *(inf. 23/24)*

1. Omów mechanizmy łączenia danych z wielu tabel.
2. Omów klasyfikację funkcji działających na pojedynczych wierszach.
3. W jakim celu buduje się perspektywy? Omów możliwe klauzule polecenia do tworzenia perspektyw.
4. Operacje na zbiorach – omów składnię poleceń i znaczenie uzyskanych wyników.
5. Przedstaw podzapytania – typy, klauzule, w których mogą wystąpić, operatory.
6. Omów typowe rozwiązania Big Data w obszarze baz/repozytoriów danych.
7. Przedstaw specyfikę środowisk analitycznych stosowanych w Big Data.
8. Omów wybrany algorytm stosowany w analityce Big Data.
9. Na czym polega MapReduce?
10. Co to jest Deep Learning, podaj przykład.
11. Jakimi cechami charakteryzują się typowe problemy Big Data?
12. Omów przykładowe techniki stosowane w rozpoznawaniu wzorców.
13. Na czym polega przetwarzanie rozproszone?
14. Omów wybraną metodykę opisującą sposób realizacji procesu wytwórczego modelu analitycznego.
15. Wymień kluczowe założenia będące warunkami zastosowania modeli predykcyjnych do wspomagania procesów decyzyjnych.
16. Jak mierzymy jakość modelu prognostycznego?
17. Omów w jaki sposób wykorzystanie systemu kontroli wersji wpływa na efektywność procesu wytwórczego rozwiązań analitycznych.
18. Wyjaśnij co to jest reprodukowalność procesu analitycznego i dlaczego jest ona ważna w praktyce gospodarczej.
19. Omów podstawowe sposoby zapewnienia reprodukowalności procesu analitycznego.
20. 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.
21. Wyjaśnij do czego wykorzystywana jest regularyzacja w procesie budowy modeli predykcyjnych.
22. Wyjaśnij różnicę pomiędzy wnioskowaniem obserwacyjnym, interwencyjnym i kontrfaktycznym.
23. Wyjaśnij na czym polega paradoks Simpsona.
24. Przedstaw korzyści ekonomiczne z przetwarzania danych w chmurze.
25. Omów technologie serveless w gromadzeniu i przetwarzaniu danych na potrzeby procesów analitycznych.
26. Przedstaw metody przechowywania danych dużych rozmiarów w chmurze.
27. Omów skalowanie dokumentowych baz danych typu noSQL w chmurze na przykładzie DynamoDB.
28. Omów skalowanie procesów analitycznych w chmurze.
29. Omów Function as a service - model przetwarzania oparty o architekturę Lambda.

30. Omów tworzenie i zarządzanie bezpieczeństwem środowisk analitycznych dla języków Python i R w chmurze.
31. Omów zarządzanie bezpieczeństwem, użytkownikami i prawami dostępu w chmurze - użytkownicy, role, polityki i grupy.
32. Przedstaw systemy zarządzania relacyjną bazą danych w chmurze i ich zastosowania w analityce danych.
33. Przedstaw modele przetwarzania danych w chmurze: IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) oraz SaaS (Software-as-a-Service).
34. Omów kwestie etyczne związane z Big Data.
35. Omów cechy danych istotne w procesie analizy danych.
36. Przedstaw, na czym polega zmienność danych i jak ją uwzględnić w wizualizacji danych.
37. Przedstaw, na czym polega niepewność w analizie danych i jak można wpływać na jej wielkość.
38. Jakie znaczenie mają metadane w analizie danych.
39. Wymień i omów układy współrzędnych stosowane przy wizualizacji danych.
40. Wymień i omów metody wizualizacji proporcji.
41. Wymień i omów metody wizualizacji relacji.
42. Wymień i omów metody wizualizacji danych geolokalizacyjnych.
43. Wymień obiekty bazy danych i omów ich przeznaczenie.
44. Wymień i omów metody wizualizacji szeregów czasowych.
45. Przedstaw, na czym polega uwzględnienie kontekstu w analizie danych.
46. Wyjaśnij co to jest system kontroli wersji na przykładzie systemu Git i zaproponuj typowy workflow.
47. Omów wybraną technikę redukcji wymiaru danych, jej zalety i wady.
48. Omów pojęcie obliczeń równoległych i podstawowe problemy, które pojawiają się przy obliczeniach równoległych.
49. Omów pojęcie estymatora odpornego na wybranym przykładzie.
50. Omów technikę regularyzacji na wybranym przykładzie, np. regresji LASSO.
51. Co oznacza określenie 3V oraz 5V w kontekście problematyki Big Data?
52. Wyjaśnij pojęcia danych ustrukturyzowanych i nieustrukturyzowanych.
53. Przedstaw architektury: Lambda i Kappa.
54. Przedstaw kluczowe cechy uczenia i predykcji w trybie wsadowym (offline learning) i przyrostowym (online learning).
55. Podaj przykład i omów w jakich sytuacjach wskazane jest zastosowanie modelu przetwarzania OLTP.
56. Podaj przykład i omów w jakich sytuacjach wskazane jest zastosowanie modelu przetwarzania OLAP.
57. Wyjaśnij pojęcie i zastosowania biznesowe hurtowni danych.
58. Omów problem czasu w strumieniowym przetwarzaniu danych, czym jest Watermark.
59. Przedstaw różnicę pomiędzy wsadowym i strumieniowym sposobem przetwarzania danych.
60. Opisz dwa biznesowe zastosowania analizy danych w czasie rzeczywistym.
61. Wymień i omów metodyki procesu eksploracji danych.
62. Omów dwie główne grupy metod eksploracji danych.
63. Omów metody selekcji zmiennych i obserwacji do modelowania data mining.

64. Metody klasyfikacji danych - przedstaw różnice i podobieństwa pomiędzy nimi.
65. Przedstaw model drzewa decyzyjnego.
66. Omów modele lasów losowych.
67. Przedstaw modele sztucznych sieci neuronowych.
68. Omów metody grupowania danych.
69. Omów metody analizy danych transakcyjnych.
70. Omów metody konstrukcji tablic trwania życia oraz podaj przykłady ich wykorzystania.
71. Porównaj modele nieparametryczne i modele parametryczne analizy czasu trwania.
72. Scharakteryzuj modele proporcjonalnych hazardów oraz podaj przykłady takich modeli.
73. Scharakteryzuj modele przyspieszonej porażki oraz podaj przykłady takich modeli.
74. Scharakteryzuj modele semiparametryczne analizy czasu trwania.
75. Wymień różnice pomiędzy podejściem klasycznym a bayesowskim w kontekście estymacji parametrów modeli analizy czasu trwania.
76. Omów modele ryzyk konkurencyjnych w analizie czasu trwania.
77. Omów ideę metod Monte Carlo bazujących na łańcuchach Markowa (MCMC) w kontekście estymacji parametrów modeli analizy czasu trwania.
78. Jakość danych w analizach biznesowych. Znaczenie i metody oceny.
79. Imputacja danych. Istota i znaczenie.
80. Imputacja wielokrotna: opis metody, wybór modelu do imputacji oraz estymacja parametrów.
81. Porównaj modele z efektami stałymi oraz modele z efektami losowymi. Przedstaw podstawowe różnice i zastosowania obu typów modeli.
82. Regresja kwantylowa: opis i zastosowania w analityce biznesowej.
83. Regresja adaptacyjna: model, opis estymacji i zastosowania w analityce biznesowej.
84. Metoda k-średnich i jej zastosowanie w ocenie wartości klienta w czasie CLV.
85. Wymień i omów zastosowania biznesowe modeli oceny wartości klienta w czasie CLV.
86. Jakie statystyki opisowe są odporne na wartości nietypowe?
87. Jakie statystyki opisowe należy stosować w przypadku prób pobranych z populacji o rozkładzie innym niż rozkład normalny?
88. Omów trzy elementy bezpieczeństwa informacji: Poufność, Integralność, Dostępność.
89. Co to jest Spear Phishing (ukierunkowany atak na osobę lub organizację)?
90. Przedstaw podstawowe zasady cyberbezpieczeństwa dla MŚP (Małe i Średnie Przedsiębiorstwa).
91. Na czym polega interpretacja języka programowania, podaj przykłady języków interpretowalnych oraz kilka różnych interpreterów.
92. Omów sposób instalacji i zarządzania bibliotekami (pakietami) w środowisku języka Python, podaj przykłady popularnych bibliotek.
93. Omów techniki iterowania na przykładzie wybranego języka programowania, np. R, Python.
94. Omów koncepcję funkcji oraz zakresu ważności zmiennych na przykładzie wybranego języka programowania, np. R, Python.
95. Co to jest silnik decyzyjny? Wymień reguły procesu akceptacji kredytowej implementowane w silniku decyzyjnym.

96. Omów pojęcia związane z przygotowaniem danych i zdarzeniem modelowym: punkt obserwacji, okres danych i okres obserwacji, wymień najczęstsze błędy modelowania (np. wzięcie danych z przyszłości) i problemy z doborem długości obu okresów.
97. Omów przykładową kartę scoringową. Jak wyznacza się punkty cząstkowe? Jak interpretuje się kartę scoringową?
98. W jaki sposób obliczana jest opłacalność procesu akceptacji kredytowej? Jaką rolę w tym odgrywa model scoringowy?
99. Co to jest analiza wpływu wniosków odrzuconych (Reject Inference)?
100. Omów wpływ ludzkiego czynnika na proces akceptacji kredytowej, czy można zwiększyć sprzedaż i zmniejszyć ryzyko kredytowe jednocześnie?

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. 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.