Pytania egzaminacyjne z eksploracji danych i uczenia maszynowego

1. Omów rolę przygotowania danych w procesie eksploracji danych i uczenia maszynowego. Przedstaw, jakie decyzje podejmuje się na etapie importu, walidacji typów i wstępnego czyszczenia danych oraz dlaczego wpływają one na dalsze modelowanie.
2. Jak rozpoznaje się i interpretuje obserwacje odstające oraz braki danych? Omów najważniejsze strategie postępowania z takimi przypadkami i ich konsekwencje dla analizy oraz jakości modelu.
3. Na czym polegają EDA, preprocessing i inżynieria cech w danych tabelarycznych? Omów rolę kodowania zmiennych kategorycznych, skalowania, transformacji oraz konstrukcji nowych cech.
4. Przedstaw regresję liniową wieloraką jako metodę uczenia nadzorowanego. Omów składniki modelu, znaczenie parametrów oraz sens minimalizacji błędu w procesie estymacji.
5. Omów interpretację modeli liniowych oraz ich ograniczenia. Wyjaśnij, kiedy współczynniki modelu są informatywne, a kiedy interpretacja staje się problematyczna.
6. Na czym polega regularyzacja modeli liniowych? Porównaj Ridge, LASSO i Elastic Net z punktu widzenia celu stosowania, wpływu na współczynniki oraz praktycznych zastosowań.
7. Omów regresję logistyczną jako model klasyfikacyjny. Wyjaśnij związek między liniowym predyktorem, prawdopodobieństwem przynależności do klasy i granicą decyzyjną.
8. Porównaj LDA, QDA i regresję logistyczną jako metody klasyfikacji. Zwróć uwagę na różnice w założeniach, sposobie modelowania oraz typowych sytuacjach, w których każda z tych metod może być uzasadniona.
9. Z czego składa się drzewo decyzyjne? Opisz funkcje poszczególnych elementów konstrukcji drzewa oraz sposób, w jaki prowadzą one do decyzji klasyfikacyjnej lub regresyjnej.
10. Jak wyznacza się podziały w drzewach decyzyjnych dla klasyfikacji i regresji? Omów kryteria podziału oraz różnice między pracą na cechach ciągłych i kategorycznych.
11. W jaki sposób kontroluje się złożoność drzewa decyzyjnego? Omów sens reguł stopu, pre-pruningu i post-pruningu oraz ich wpływ na przeuczenie i uogólnianie modelu.
12. Omów zalety i ograniczenia pojedynczych drzew decyzyjnych. W jakich sytuacjach drzewa są szczególnie użyteczne, a kiedy ich własności stają się problemem?
13. Na czym polega bagging i dlaczego bootstrap jest w nim tak istotny? Wyjaśnij, jak zespołowanie wielu drzew wpływa na stabilność i wariancję predykcji.
14. Czym las losowy różni się od zwykłego baggingu drzew? Omów rolę losowania cech, mechanizm budowy modelu oraz skutki tego podejścia dla jakości i różnorodności drzew w zespole.
15. Omów najważniejsze hiperparametry lasów losowych i innych zespołów drzewiastych. Wyjaśnij, jak wpływają one na złożoność modelu, przeuczenie i koszt obliczeniowy.

16. Na czym polega boosting jako model addytywny uczący się na błędach poprzednich modeli? Porównaj ogólną ideę AdaBoost i gradient boostingu.
17. Porównaj XGBoost, LightGBM i CatBoost jako nowoczesne odmiany boostingu drzew. Omów ich najważniejsze pomysły algorytmiczne, sposób pracy z danymi oraz typowe przewagi praktyczne.
18. Wyjaśnij związek między twierdzeniem Bayesa, estymacją metodą największej wiarygodności i estymacją MAP. Omów rolę informacji a priori w uczeniu modeli probabilistycznych.
19. Omów założenia i odmiany naiwnego Bayesa. Porównaj warianty Multinomial, Bernoulli i Gaussian z punktu widzenia typu danych, modelowanych rozkładów i obszarów zastosowań.
20. Przedstaw metodę k-NN jako podejście oparte na podobieństwie. Omów rolę metryki odległości, parametru (k), kompromisu bias-variance oraz główne zalety i wady tej metody.
21. Na czym polega regresja sklejona i czym splajn różni się od zwykłej regresji wielomianowej? Omów ideę aproksymacji kawałkami oraz znaczenie gładkiego łączenia funkcji.
22. Wyjaśnij rolę węzłów, warunków gładkości i elastyczności modelu w metodach splajnowych. Jak dobór tych elementów wpływa na niedouczenie, przeuczenie i interpretację modelu?
23. Przedstaw ideę modeli GAM. Omów ich strukturę addytywną, interpretację efektów cząstkowych oraz najważniejsze zalety i ograniczenia względem prostszych i bardziej sztywnych modeli.
24. Wyjaśnij ideę maksymalnego marginesu w liniowo separowalnym SVM. Omów znaczenie hiperpłaszczyzny decyzyjnej, marginesu oraz wektorów nośnych.
25. Na czym polega miękki margines w SVM? Omów rolę parametru (C) i wyjaśnij, jak wpływa on na kompromis między regularizacją a dopasowaniem do danych treningowych.
26. Na czym polega sztuczka jądrowa w SVM? Porównaj intuicję i zastosowania jąder liniowego, wielomianowego i RBF, a następnie odnieś się do idei regresyjnego wariantu SVR.
27. Przedstaw uczenie ze wzmocnieniem jako odmienny paradygmat uczenia maszynowego. Omów elementy procesu decyzyjnego Markowa oraz różnice między tym podejściem a klasycznym uczeniem nadzorowanym.
28. Omów rolę funkcji wartości, równań Bellmana i aktualizacji w Q-learningu. Wyjaśnij, dlaczego problem eksploracji i eksploatacji jest w uczeniu ze wzmocnieniem kluczowy.
29. Omów analizę koszykową i reguły asocjacyjne. Przedstaw sposób formalizacji danych transakcyjnych oraz ideę wyszukiwania częstych zbiorów i reguł metodami Apriori oraz FP-Growth.
30. Omów problem wykrywania anomalii. Przedstaw typy anomalii, trudności ich formalizacji oraz porównaj intuicję działania metod LOF i Isolation Forest.