

Data Science in Human Health: Advanced Applications

Data Science · Answer Key · 17 Questions

1. Which data science technique is primarily employed to identify complex gene-disease associations from large-scale genomic datasets, often revealing epistatic interactions?

- A) Linear Regression
- B) K-Means Clustering
- C) Ensemble Methods (e.g., Random Forests, Gradient Boosting)**
- D) Principal Component Analysis (PCA)

2. In the context of analyzing electronic health records (EHRs), what type of data science model is most effective for predicting patient readmission risk by capturing temporal dependencies and sequential patterns of care?

- A) Support Vector Machines (SVMs)
- B) Hidden Markov Models (HMMs)**
- C) Naive Bayes Classifiers
- D) Decision Trees

3. The application of convolutional neural networks (CNNs) in medical imaging, such as detecting diabetic retinopathy from fundus photographs, relies on their ability to learn hierarchical features. What is a key characteristic of these learned features?

- A) Global pixel intensity distributions
- B) Low-level features like edges and corners, progressing to more complex shapes and textures**
- C) Simple statistical moments of image noise
- D) Average color saturation across the entire image

4. When modeling drug response variability using pharmacogenomic data, which data science approach is most suitable for discovering interactions between multiple genetic variants and their combined effect on drug efficacy or toxicity?

- A) Simple Logistic Regression
- B) Hierarchical Clustering
- C) Factor Analysis
- D) Multi-locus Interaction Analysis (e.g., multifactor dimensionality reduction - MDR)**

5. What data science methodology is crucial for imputing missing values in large-scale microbiome datasets, which are often characterized by sparsity and high dimensionality, to enable robust downstream analysis?

- A) Mean Imputation
- B) k-Nearest Neighbors (k-NN) Imputation
- C) Multiple Imputation by Chained Equations (MICE)**
- D) Hot-deck Imputation

6. In analyzing the protein-protein interaction networks that underpin cellular signaling pathways, what data science concept is used to identify functionally important proteins that act as hubs or bottlenecks?

- A) Graph partitioning
- B) Centrality measures (e.g., degree centrality, betweenness centrality)**
- C) Community detection algorithms
- D) Network motif discovery

7. Which data science approach is essential for identifying subtle patterns in wearable sensor data (e.g., ECG, accelerometer) to detect early signs of cardiac arrhythmias, considering the time-series nature and noise inherent in such signals?

- A) Independent Component Analysis (ICA)
- B) Recurrent Neural Networks (RNNs) / Long Short-Term Memory (LSTM) networks**
- C) Latent Dirichlet Allocation (LDA)
- D) Non-negative Matrix Factorization (NMF)

8. The development of personalized cancer vaccines often involves analyzing neoantigens derived from tumor sequencing data. What data science task is critical for predicting the immunogenicity of these potential neoantigens?

- A) Dimensionality reduction
- B) Anomaly detection
- C) Predictive modeling using machine learning algorithms trained on known T-cell epitopes**
- D) Topic modeling

9. When analyzing the complex interplay of environmental factors and genetic predispositions to chronic diseases, what data science technique can reveal non-linear interactions and complex causal pathways?

- A) ANOVA (Analysis of Variance)
- B) Partial Least Squares Regression (PLSR)
- C) Bayesian Networks**
- D) Simple Linear Regression

10. In the field of metabolomics, identifying biomarkers for disease states often involves dealing with high-dimensional and correlated data. Which data science method is commonly used for feature selection and dimensionality reduction while preserving variance?

- A) Factor Analysis of Mixed Data (FAMD)
- B) Sparse Principal Component Analysis (Sparse PCA)**
- C) Linear Discriminant Analysis (LDA)
- D) Independent Component Analysis (ICA)

11. What data science approach is employed to model the spread of infectious diseases, incorporating factors like population density, mobility patterns, and intervention strategies, to predict epidemic trajectories?

- A) Time Series Forecasting (e.g., ARIMA)
- B) Agent-Based Modeling (ABM)**
- C) Survival Analysis
- D) Clustering algorithms

12. The analysis of single-cell RNA sequencing (scRNA-seq) data to identify distinct cell populations and their developmental trajectories utilizes which core data science concept for dimensionality reduction and visualization?

- A) Independent Component Analysis (ICA)
- B) t-distributed Stochastic Neighbor Embedding (t-SNE)**
- C) Canonical Correlation Analysis (CCA)
- D) Factor Analysis

13. When applying natural language processing (NLP) to extract clinical information from unstructured physician notes, what data science task is crucial for identifying and categorizing medical concepts like diseases, symptoms, and medications?

- A) Sentiment Analysis
- B) Named Entity Recognition (NER)**
- C) Topic Modeling
- D) Text Summarization

14. In the context of understanding the gut-brain axis, what data science technique can be used to uncover complex relationships between gut microbiome composition and neurological health indicators derived from patient data?

- A) K-means Clustering
- B) Regularized Regression (e.g., Lasso, Ridge)**
- C) Principal Component Analysis (PCA)
- D) Association Rule Mining

15. Which data science framework is essential for building robust models that predict patient adherence to medication regimens, considering multiple influencing factors such as socioeconomic status, treatment complexity, and social support?

- A) Clustering algorithms
- B) Time series analysis
- C) Survival analysis (e.g., Cox proportional hazards)**
- D) Association rule mining

16. When analyzing functional magnetic resonance imaging (fMRI) data to identify brain regions associated with specific cognitive tasks, what data science technique is used to map signal changes over time and across the brain volume?

- A) Principal Component Analysis (PCA)
- B) Independent Component Analysis (ICA)**
- C) Graph signal processing
- D) Support Vector Machines (SVMs)

17. The identification of potential drug repurposing candidates from existing drug databases and disease knowledge graphs often relies on which data science approach to infer novel therapeutic links?

- A) Clustering
- B) Dimensionality Reduction
- C) Link Prediction (e.g., using graph embeddings and similarity measures)**
- D) Anomaly Detection