Top 20 Data Science Research Topics and Areas For the 2020-2030 Decade

2 authors:

Joab O. Odhiambo
University of Nairobi
11 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Stanley Sewe
Pan African University
8 PUBLICATIONS **1** CITATION
SEE PROFILE

Top 20 Data Science Research Topics and Areas For the Decade

*Joab Odhiambo, Bsc., MSc.,PhD (**Actuarial Science**).*

Author's Email: joabodhiambo2022@gmail.com

In the world today, there is no doubt that a data scientist is among the most sought-after experts because of their nature of jobs. The following are the hottest data science topics and areas that any aspiring data scientist should know whether they are data analysts or just business intelligence specialists who aim to advance their data analysis skills and knowledge to enhance their competitiveness.

1. The process of data mining

Data mining is an iterative process, which involves discovering patterns in large data sets commonly known as the **BIG DATA**. It includes methods and techniques such as machine learning, statistics, database systems and etc. The two main data mining objectives are to find out patterns and establish trends and relationship in a dataset in order to solve problems.

The general stages of the data mining process are: problem definition, data exploration, data preparation, modeling, evaluation, and deployment. Core terms related to data mining are classification, predictions, association rules, data reduction, data exploration, supervised and unsupervised learning, datasets organization, sampling from datasets, building a model and etc.

2. Data visualization Data

visualization is defined as the presentation of data especially in a graphical format. Through the process, it enables decision-makers of all levels to see data and analytics presented visually, so they can identify valuable patterns or trends. Data visualization is another broad subject that covers the understanding and use of basic types of graphs (such as line graphs, bar graphs, scatter plots, histograms, box and whisker plots, heatmaps. You cannot go without these graphs. In addition, here you need to learn about multidimensional variables with adding variables and using colors, size, shapes, animations. Manipulation also plays a role during the process of visualization. You should be able to rascal, zoom, filter, aggregate data. Using some specialized visualizations such as map charts and tree maps is a hot skill too.

3. Dimension reduction methods and techniques

Dimension Reduction process involves converting a data set with vast dimensions into a dataset with lesser dimensions ensuring that it provides similar information in short. In other words, dimensionality reduction consists of series of techniques and methods in machine learning and statistics to decrease the number of random variables. There are so many methods and techniques to perform dimension reduction such as Missing Values, Low Variance, Decision Trees, Random Forest, High Correlation, Factor Analysis, Principal Component Analysis, Backward Feature Elimination.

4. Data Classification Methods

Classification is a core data mining technique for assigning categories to a set of data. The aim is to support gathering accurate analysis and predictions from the data. Classification is one of the hottest data science topics too. A data scientist should know how to use classification algorithms to solve different business problems. This includes knowing how to define a classification problem, explore data with univariate and bivariate visualization, extract and prepare data, build classification models, evaluate models, and etc. Linear and non-linear classifiers are some of the key terms here.

5. Simple and multiple linear regression Modeling

Linear regression models are among the basic statistical models for studying relationships between an independent variable X and Y dependent variable. It is a mathematical modeling which allows you to make predictions and prognosis for the value of Y depending on the different values of X. There are two main types of linear regression: simple linear regression models and multiple linear regression models. Key points here are terms such as correlation coefficient, regression line, residual plot, linear regression equation and etc. For the beginning, see some simple linear regression examples.

6. K-nearest neighbor (k-NN)

N-nearest-neighbor is a data classification algorithm that evaluates the likelihood a data point to be a member of one group. It depends on how near the data point is to that group. As one of the key non-parametric method used for regression and classification, k-NN can be classified as one of the best data science topics ever. Determining neighbors, using classification rules, choosing k are a few of the skills a data scientist should have. K-nearest neighbor is also one of the key text mining and anomaly detection algorithms.

7. Naive Bayes Theorem

Naive Bayes is a collection of classification algorithms which are based on the famous Bayes Theorem. The theorem is commonly applied in Machine Learning, Naive Bayes has some crucial applications such as spam detection and document classification. There are different Naive Bayes variations. The most popular of them are the Multinomial Naive Bayes, Bernoulli Naive Bayes, and Binarized Multinomial Naive Bayes.

8. Classification and regression trees (CART)

When it comes to algorithms for predictive modeling machine learning, decision trees algorithms have a vital role. The decision tree is one of the most popular predictive modeling approaches used in data mining, statistics and machine learning that builds classification or regression models in the shape of a tree (that's why they are also known as regression and classification trees). They work for both categorical data and continuous data. Some terms and topics you should master in this field involve CART decision tree methodology, classification trees, regression trees, interactive dihotomiser, C4.5, C5.5, decision stump, conditional decision tree, M5, and etc.

9. Logistic regression Modeling

Logistic regression is one of the oldest data science topics and areas and as the linear regression, it studies the relationship between dependable and independent variable. However, we use logistic regression analysis where the dependent variable is dichotomous (binary). You will face terms such as sigmoid function, S-shaped curve, multiple logistic regression with categorical explanatory variables, multiple binary logistic regression with a combination of categorical and continuous predictors and etc.

10. Neural Networks Systems Neural Networks act as a total hit in the machine learning nowadays. Neural networks (also known as artificial neural networks) are systems of hardware and/or software that mimic the human brain neurons operation. The primary goal of creating a system of artificial neurons is to get systems that can be trained to learn some data patterns and execute functions like classification, regression, prediction and etc.

Neural Networks are a kind of deep learning technologies used for solving complex signal processing and pattern recognition problems. Key terms here relates to concept and structure of Neural Networks, perceptron, Back-propagation, Hopfield Network.

11. Discriminant analysis

Discriminant analysis is a technique, which is used by the researcher when analyzing the research data whenever the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature. This is important in data analysis when conducting parameter estimation criterion.

12. Association rules

Association rule learning is among the important rule-based machine learning method for determining interesting relations in between variables in huge databases such as Big Data. It is intention is to identify strongest rules discovered within databases while using some measures of interestingness applied in the field of data science.

13. Cluster analysis Methods

Cluster analysis or commonly known as clustering is the task of grouping a set of objects in a way that the objects in the similar group are more similar to one other than to those within other group when a comparison is done. 14. Time series A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. 15. Regression-based forecasting It can be highly beneficial for companies to develop a forecast of the future values of some important metrics, such as demand for its product or variables that describe the economic climate. Linear regression is a time-series method that uses basic statistics to project future values for a target variable.

16. Smoothing methods Approaches

Data smoothing is done by using an algorithm to remove noise from a data set. This allows important patterns to stand out. Data smoothing can be used to help predict trends, such as those found in securities prices. There are different methods in which data smoothing can be done. Some of these include the random method, random walk, moving average, simple exponential, linear exponential, and seasonal exponential smoothing. A smoothed moving average places equal weight to both recent prices and historical ones.

17. Time stamps and financial modeling

A timestamp is a sequence of characters or encoded information identifying when a certain event occurred, usually giving date and time of day, sometimes accurate to a small fraction of a second. Financial modeling is the task of building an abstract representation of a real world financial situation. This is a mathematical model designed to represent the performance of a financial asset or portfolio of a business, project, or any other investment. Data science applies the use of financial models to capture the realities that exists when conducting financial modeling.

18. Fraud detection using AI

The machine learning (ML) approach to fraud detection has received a lot of publicity in recent years and shifted industry interest from rule-based fraud detection systems to ML-based solutions. It gives the differences between machine learning and rule-based approaches when detecting fraud in most cases in the financial sectors.

19. Data engineering Techniques – Hadoop, MapReduce, Pregel.

Data engineering is the aspect of data science that focuses on practical applications of data collection and analysis. For all the work that data scientists do to answer questions using large sets of information, there have to be mechanisms for collecting and validating that information. In order for that work to ultimately have any value, there also have to be mechanisms for applying it to real-world operations in some way. Those are both engineering tasks: the application of science to practical, functioning systems.

20. GIS and spatial data

A spatial database is a database that is optimized for storing and querying data that represents objects defined in a geometric space. Most spatial databases allow the representation of simple geometric objects such as points, lines and polygons. Spatial data, also known as geospatial data, is information about a physical object that can be represented by numerical values in a geographic coordinate system. ... Geographic Information Systems (GIS) or other specialized software applications can be used to access, visualize, manipulate and analyze Geospatial data, which is important in data science as a discipline.

Prepared by Joab Odhiambo, *Data Scientist and Actuarial Science Specialist*