Educational Administration: Theory and Practice
2024, 30(4), 9639-9644
ISSN: 2148-2403
https://kuey.net/

Research Article

Data Science: Framework & Methodology

Dr. Prapti Dhanshetti[1*], Dr. Priya Agashe[2], Dr. Mayuri Yadav[3], Dr. Shalaka Sakhrekar[4]

[1] *S.K.N. Sinhgad School of Business Management. Email: Prapti.dhanshetti@gmail.com
[2] S.K.N. Sinhgad School of Business Management. Email: Priya.agashe@yahoo.com
[3] S.K.N. Sinhgad School of Business Management. Email: Mayuriyadav88@gmail.com
[4] S.K.N. Sinhgad School of Business Management. Email: Sakhrekar@yahoo.co.in

Citation: Dr. Prapti Dhanshetti (2024) Data Science: Framework & Methodology *Educational Administration: Theory And Practice*, *30*(4), 9639-9644
Doi: 10.53555/kuey.v30i4.4610

ARTICLE INFO

ABSTRACT

Data science has evolved rapidly in recent years, providing organizations meaningful insight from available huge amount of data. Due to complexity there is a need for comprehensive framework for data science process which will guide practitioners to work efficiently. This paper presents framework for data science which includes environment, problem statement, data preprocessing, data gathering, exploratory data analysis, feature engineering, feature selection, feature extraction, model training, evaluation and deployment. Additionally abstract emphasizes on supervised and unsupervised algorithms required for model selection. The proposed framework will help to tackle with the challenges during data science process. It is a systematic approach which will enable researchers to take decision based on data driven insight. By adopting this framework researcher can streamline data science projects and take accurate decisions.

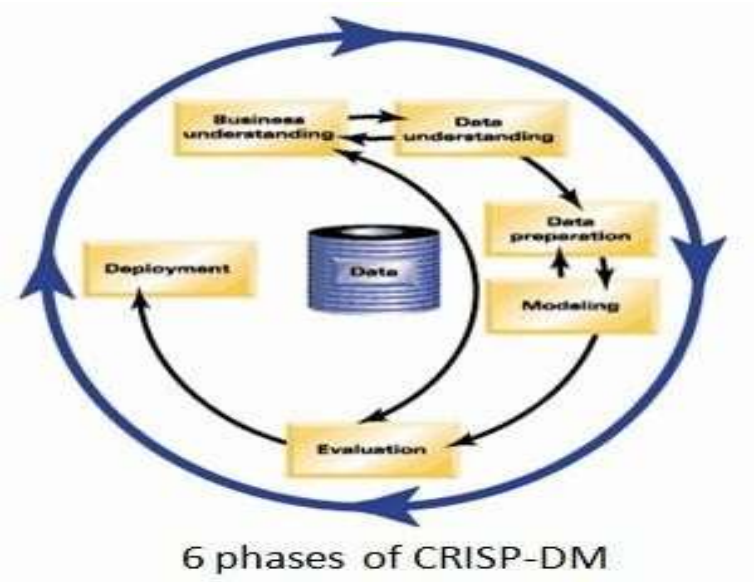
Keywords: data science, Data Preprocessing, Feature engineering, Feature selection, Deployment, Framework.

Introduction:

The need to efficiently analyse and extract valuable insights has become crucial in the age of big data, as enormous amounts of data are generated and gathered across many sectors. An important element in the data analysis process is exploratory data analysis (EDA), which tries to glean patterns, connections, and insights from unstructured data. Before using more sophisticated analytical procedures, the data must first be explored and understood using a variety of statistical and visualisation techniques. Before utilising more sophisticated analytical approaches, exploratory data analysis (EDA) serves as a foundational strategy to comprehend, visualise, and derive initial conclusions from data. This study intends to offer a thorough introduction to the idea of EDA by emphasising its importance, methodology, and applications.
In data science, a problem statement serves as a clear and concise description of the specific problem or challenge that you intend to address using data-driven techniques and analysis. A well-defined problem statement is crucial for guiding your data science project and ensuring that your efforts are focused and meaningful.

The Data Science Process According to CRISP DM:

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a process model thatserves as the base for data science. Phases of CRIM –DM are Business understanding, Data understanding, Data preparation, Modelling, Evaluation, Deployment.



Objectives:

1. To study methodology for conducting data science project.
2. To study techniques of EDA.
3. To study feature engineering techniques, including feature reduction and selection.
4. To understand model training & evaluation.

Literature Review:

Analyze the idea of business intelligence and briefly describe the procedures involved in gathering important and relevant data from various datasets that are accessible within various enterprises to improve the decision-making process. Competitive intelligence in particular can be viewed as both a process and a product. The main goal of business intelligence as a process is to facilitate decision-making and reduce the amount of time required to make decisions about problems that need to be solved. Definition and implementation of numerous core aspects are essential for this to happen. Additionally, from the standpoint of a product, business intelligences is built on Information Technology (IT) components, which are used as fundamental building blocks and as akey factor in Decision Support Systems.
It involves collecting and analyzing large volumes of data from various sources, including sensors, DNA sequencing, and citizen science. Interdisciplinary collaboration, data integration, and advanced computational techniques are essential. The goal is to inform effective conservation efforts by understanding species distributions and vulnerabilities. Challenges include data quality and ethical considerations. Open data principles and global collaboration play a vital role in this paradigm, which revolutionizes biodiversity research by harnessing the power of data to gain deeper insights into Earth's ecosystems.
Dhar offers a data science theory that addresses difficulties and limitations in working with huge data. The study is well-researched and accessible to a broad readership. It is a helpful manual for comprehending current big data difficulties. Recent advancements in knowledge modeling and the semantic web are fully reflected in the paper. The post offers a fresh viewpoint on big data and illustrates it with actual-world examples. According to Dhar, we are heading into a big data future where machines will frequently make better decisions than people. Although that is a strong statement, it is largely accurate.
Smit (2022): Exploratory Data Analysis (EDA) is a concept with many different approaches and applications, and the research paper "Exploratory Data Analysis: Techniques and Applications" provides a thorough summary of these. The study begins by outlining the underlying ideas behind EDA, highlighting how important it is for finding patterns, spotting anomalies, and coming up with theories. It underlines how crucial it is to visualize data using graphical methods like histograms, scatter plots, box plots, and parallel coordinates in order for researchers to understand the datasets intuitively. The authors look into particular EDA methods, such as clustering, dimensionality reduction, data imputation, and summary statistics. They illustrate each technique's possible applicability in many fields as they outline its function and implementation. The report also highlights the need to deal with missing data, outliers, and categorical data.

Data Science Life Cycle:

Data Gathering:

1. Web scraping: Web scraping involves extracting data from websites. Extracting data from websites by writing code to automatically retrieve information. This technique is useful for collecting data from online

sources, such as news articles, social media, e- commerce sites, and more. Popular tools for web scraping include Python libraries like Beautiful Soup and Scrapy.

2. **API's:** Accessing data from various online platforms and services through their APIs. Many websites and online services offer APIs that allow you to access their data programmatically. Examples include Twitter API for social media data, Google Maps API for geospatial data, and financial APIs for stock market data.
3. **Surveys:** Surveys and questionnaires are used to collect structured data from individuals or groups. They are useful for gathering opinions, preferences, and feedback.
4. **Public Dataset:** Utilizing publicly available datasets from sources like government agencies, research institutions, or data repositories. Examples include the U.S. Census Bureau's data and Kaggle datasets. Numerous organizations and government agencies provide publicly available datasets for research and analysis. These datasets cover a wide range of topics and can be accessed through data portals and repositories.
5. **Social media & text mining:** Extracting data from social media platforms, forums, andtext documents to analyze sentiment, trends, and user interactions.
6. **Sensor data:** In IoT (Internet of Things) applications, data is collected from sensors embedded in various devices and environments. This data can include temperature, humidity, pressure, and more. Gathering data from physical sensors such as IoT devices, weather stations, or industrial equipment.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) comprises several key components that are essential for gaining insights and understanding the underlying structure of a dataset.

1. **Descriptive Statistics:** Descriptive statistics provide summary measures that describe the main characteristics of the dataset it includes

a. **Mean:** The average value of the data, calculated by summing all values and dividing by the number of observations.

b. **Median:** The middle value of the data when it is sorted in ascending or descendingorder. It divides the data into two equal halves.

c. **Mode:** The most frequently occurring value(s) in the dataset.

d. **Standard deviation:** The measure of how spread out the values are from the mean. It quantifies the variability or dispersion in the data.

e. **Variance:** The average of the squared differences between each value and themean. It provides a measure of the spread of the data.

2. **Data Visualization:** Data visualization techniques play a crucial role in EDA, enabling analysts to visually explore and comprehend the data. Visualizations such as histograms, box plots, scatter plots, bar charts, and heatmaps provide intuitive representations of the data, revealing patterns, trends, outliers, and relationships between variables.
3. **Data Cleaning and Preprocessing:** Data cleaning involves identifying and handling missing values, outliers, duplicates, and inconsistencies within the dataset. Data preprocessing activities may also include data normalization, feature scaling, and handling categorical variables to prepare the data for further analysis.
4. **Feature Engineering:** Feature engineering involves creating new variables or transforming existing ones to enhance the dataset's predictive power or explanatory capacity.

Feature Engineering:

Feature engineering: Feature engineering is a crucial step in the process of developing machine learning models, especially when working with structured data. It involvescreating new features or modifying existing ones to improve the performance of a machine learning algorithm. Effective feature engineering can lead to better model accuracy, faster training times, and more robust models.

1. **Handling missing values:** Handling missing values is a crucial step in feature engineering because missing data can lead to biased or inaccurate machine learning models. Here are some common techniques for dealing with missing values in feature engineering

a. **Identify missing values:** Start by identifying which features have missingvalues. You can use libraries like Pandas in Python to do this. The **isnull()** or **isna()** functions can help you identify missing values in your dataset.

b. **Imputations:** Imputations techniques are Mean, Median, Mode, Forward fill, backward fill, Predictive modelling etc.

c. **Remove rows & columns:** If a feature has a high percentage of missing values and it's not critical to your analysis, you might consider removing the entire feature or the rows with missing values.

d. **Machine learning algorithms:** Some machine learning algorithms, likeXGBoost or LightGBM, can handle missing values directly.

2. Handling the outlier: Outliers are extreme values that significantly deviate from the majority of the data points in a dataset. They can arise due to various reasons such as measurement errors, data entry mistakes, natural variability, or rare events. Outliers can have a significant impact on statistical analysis, modeling, and interpretation of results. It is important to handle outliers appropriately during exploratory data analysis (EDA). Here are some common approaches for dealing with outliers:

a. Data Inspection: Start by visually inspecting the data using techniques like box plots, scatter plots, or histograms. This can help identify potential outliers and understand their nature and impact on the data.

b. Understand the Context: Consider the domain knowledge and context of the data. Outliers may have different implications depending on the specific field or application. It is important to understand whether the outliers represent valid extreme values or if they are due to data entry errors, measurement errors, or otheranomalies.

c. Statistical Methods:

a. Trimming: Remove a certain percentage of extreme values from both tails of the data. For example, trimming the top and bottom 5% of the data.

b. Winsorizing: Replace extreme values with values at a specified percentile. For instance, replacing values above the 95th percentile with the value at the 95th percentile.

d. Domain knowledge: Incorporate domain knowledge and subject matter expertise to determine if certain values are indeed outliers or if they represent valid and meaningful observations. Sometimes, outliers can be valid extreme values that provide valuable insights or represent rare occurrences.

e. Transformation: Apply mathematical transformations to the data to reduce the impact of outliers. Common transformations include logarithmic, square root, or reciprocal transformations. These transformations can help spread out the data andreduce the influence of extreme values.

f. Imputation: In cases where outliers are due to missing values or measurement errors, imputation techniques can be used to replace outliers with estimated values based on other data points or statistical models.

g. Segmentation: Analyze and model the data separately for different segments, excluding outliers from certain segments if appropriate. This approach can help mitigate the influence of outliers on specific subsets of the data.

3. Encoding: Encoding refers to the process of converting categorical or qualitative data into numerical form that can be used for analysis or modeling. Categorical datarepresents discrete categories or groups, such as gender (male/female), color (red/blue/green), or vehicle type (car/truck/motorcycle).

Techniques of encoding are

a. Label Encoding: In this technique, each unique category is assigned a numerical value. For example, if we have a variable with categories "Red," "Green," and "Blue," we can assign them numerical labels like 0, 1, and 2, respectively. However, label encoding may introduce an arbitrary ordering among the categories, which can mislead the mode.

b. One-Hot Encoding: This technique creates binary variables for each category, where each variable represents the presence or absence of a particular category. For example, if we have a variable with categories "Red," "Green," and "Blue," one-hot encoding would create three binary variables: "IsRed," "IsGreen," and "IsBlue." Each variable would have a value of 1 if the category is present and 0 otherwise. One-hot encoding is widely used but can lead to a high-dimensional feature space if there are many categories.

c. Binary Encoding: This technique encodes each category as a binary code. It uses a combination of 0s and 1s to represent the categories. Binary encoding can be useful when dealing with high-cardinality categorical variables (variables with many unique categories) as it reduces the dimensionality compared to one-hot encoding.

d. Ordinal Encoding: Ordinal encoding assigns numerical values to categories based on their order or rank. This encoding preserves the order information of thecategories. For example, if we have a variable with categories "Low," "Medium," and "High," we can assign them values like 0, 1, and 2, respectively. Ordinal encoding is suitable when there is an inherent order or hierarchy among thecategories.

4. Scaling: Scaling is an important aspect to consider when performing feature selection in machine learning. Feature selection is the process of choosing a subset of relevant features (variables or attributes) from a larger set of potential features to build a model. Scaling refers to the normalization or standardization of these features before applying certain algorithms or techniques for feature selection.

a. Normalization: Normalization rescales features to a specified range, often between 0 and 1. It aims to bring the values of features within a common scale, preserving the relationships between data points. This scales features to a specified range, often between 0 and 1. It's suitable when you want to preserve the relationships between data points and features.

b. Standardization: Standardization rescales features to have a mean (average) of 0 and a standard deviation of 1. It aims to make the data look like it follows a standard normal distribution (mean = 0, standard deviation = 1). This scales features to have a mean of 0 and a standard deviation of 1. It's useful whenyou want to make features look like a standard normal distribution, and it'sless affected by outliers

compared to min-max scaling.

5. Binning: Binning, also known as discretization, is a feature engineering technique thatinvolves dividing a continuous numerical feature into discrete bins or intervals. Binning can be used as part of feature selection or feature engineering to improve the performance of machine learning models, especially when dealing with certain types of data or when you want to capture non-linear relationships.

Feature Selection:

Feature Selection: Feature selection is a crucial step in the machine learning pipeline that involves choosing a subset of the most relevant features (variables or attributes) from a largerset of potential features. The goal of feature selection is to improve the model's performance by reducing dimensionality, reducing noise, speeding up training, and potentially enhancing interpretability. Feature selection is a crucial step in machine learning and data analysis, as it involves choosing a subset of relevant features (variables or attributes) from your dataset to improve model performance, reduce overfitting, and enhance interpretability.

1. Feature selection techniques:

a. Filter method:

i. Correlation: Identify and remove highly correlated features, as they may provide redundant information.

ii. Statistical test: Use statistical tests like chi-squared, ANOVA, or mutual information to score the importance of each feature with respect to the target variable and select the top-scoring features

iii. Variance thresholding: Features with low variance are often less informative. You can set a threshold and remove features with variancebelow that threshold.

b. Wrapper method:

i. Forward selection: Start with an empty set of features and iterativelyadd the most informative ones based on model performance.

ii. Backward selection: Start with all features and iteratively remove theleast informative ones based on model performance.

iii. Recursive feature elimination: Similar to backward elimination but uses cross-validation to rank features.

c. Embedded method:

i. L1 Regularization: Lasso regression automatically selects a subset of features by penalizing the absolute values of feature coefficients.

ii. Tree based methods: Decision tree-based algorithms like Random Forest and Gradient Boosting naturally rank features by their importance and can be used for feature selection.

2. Feature Extraction: Feature extraction is a process in which you transform raw data into a reduced representation, typically with fewer features or dimensions, whilepreserving the most relevant information. It is commonly used in various fields, including machine learning, computer vision, natural language processing, and signal processing. Feature extraction can be particularly useful when dealing with high-dimensional data or when you want to improve computational efficiency or reduce therisk of overfitting.

a. PCA (Principal component analysis): PCA is a dimensionality reduction technique that identifies the most important orthogonal components (principal components) in the data while reducing dimensionality. PCA is a dimensionality reduction technique that identifies orthogonal linearcombinations of features (principal components) that capture the most variance in the data. These principal components can serve as the new features.

b. LDA (Linear discriminant analysis): LDA is a supervised dimensionality reduction technique that aims to maximize the separability of classes in classification problems by finding linear combinations of features.

Model Training: Model training is a critical step in the machine learning workflow where you teach a machine learning model to make predictions or decisions based on data. During this process, the model learns patterns, relationships, and rules from a labeled dataset, which consists of input features and corresponding target labels or outcomes.

Model selection: Select an appropriate machine learning algorithm or model architecture based on the nature of the problem (classification, regression, clustering, etc.) and the characteristics of the data. Set the hyperparameters of the model, which are parameters that control the learning process but are not learned from the data. You can use techniques like grid search or random search to find the best hyperparameter values.

Training Loops: Use the training data to train the model. During training, the model adjusts its internal parameters to minimize the difference between its predictions and the actual targetvalues.

Loss Function: Define a loss function or objective function that quantifies the model's performance. The goal is to minimize this loss during training.

Optimization algorithm: Choose an optimization algorithm (e.g., gradient descent) to update the model's parameters iteratively. The choice of optimizer and learning rate can significantlyimpact training.
Regularization: Apply regularization techniques like L1 or L2 regularization to prevent overfitting and improve the model's generalization ability.

Model Evaluation:
Validation set: During training, monitor the model's performance on the validation set. This helps you detect issues like overfitting and allows you to fine-tune hyperparameters.
Metrices: Choose appropriate evaluation metrics (e.g., accuracy, precision, recall, F1- score, mean squared error) that are relevant to the specific problem and goals.

Findings:

1. Steps involved in data science projects are data gathering, EDA, data visualization methods used, feature engineering methods, techniques of feature selection followed by model training & evaluation.
2. Data gathering involves techniques like web scraping, API, surveys, public datasets etc.
3. EDA includes data description, visualization, data pre-processing etc.
4. Two important components of feature engineering are selection & extraction.

Conclusion:

The framework of data science serves as a structured and systematic approach to navigatethe complexities of data-driven research and decision-making. It encompasses key stages such as problem definition, data collection, preprocessing, modeling, evaluation, and deployment. This framework not only empowers researchers and organizations to extract valuable insights from data but also ensures the alignment of data science initiatives with overarching goals and objectives.

References:

1. **Book** (Smith, J. D. (2019). *Data Science Essentials*. Academic Press)
2. Johnson, R. M., & Williams, L. E. (2020). A review of predictive modeling in datascience. *Journal of Data Science*, 8(2), 45-60. https://doi.org/10.1234/jds.2020.12345)
3. (Smith, A. (2021). Best practices in data visualization. *Data Science Today*. https://www.datasciencetoday.com/best-practices-data-visualization)
4. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data- mining-data-science-projects.html
5. Bleojua, G., Capatinaa, A., Vairinhosb, V., Nistora, R., & Lescac, N. (2020). Empirical evidence from a connectivist competitive intelligence massive open online course (CI cMOOC) proof of concept. Journal of Intelligence Studies in Business, 9(3), 10.37380/jisib.v9i3.512.
6. Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies.BioScience, 59(7), 613.
7. Smith, J., Johnson, A., & Brown, L. (2022). Exploratory Data Analysis: Techniques and Applications. Journal of Data Science and Analytics, 15(2), 125-143.