**The Data Science Self-Efficacy Survey**

This document introduces a self-efficacy survey for data science to evaluate and measure an individual's confidence in utilizing data science concepts. The elicit form comprises 48 questions and is categorized into eight sections, collectively identified by computer science, business, and statistics experts. These categories encompass the critical skills deemed essential for constructing data science solutions. Image 1 visually represents the eight proposed stages that the survey follows.

A diagram of a model

Description automatically generated with medium confidence

**Fig. 1 Data Science Self-Efficacy Categories**

**Domain Knowledge and Research Design**

1. I am confident in creating a plan and designing an effective strategy to develop necessary solutions in a data science project.
2. I am confident in establishing realistic timelines and defining achievable milestones using the data science life cycle.
3. I am confident in exploring a domain to acquire the necessary knowledge for a specific data science project.
4. I am confident in exploring trends and preparing reviewed literature and other scholarly justification from the data science project.
5. I am confident in my ability to formulate investigative questions that align with the nature of the problem.
6. I am confident in my ability to consider ethical implications related to data privacy, bias, and fairness throughout the process.
7. I am confident in creating clear documentation for code, models, and any essential insights made during the project.

**Data Planning and Data Collection**

1. I am confident in articulating the investigated problem and identifying suitable and trustworthy data sources to help derive insights.
2. I am confident in my ability to design an efficient data collection method while identifying challenges that might arise in the collection process.
3. I am confident in my ability to [iteratively] adapt modifications to the data collection and cleaning process in response to new findings.
4. I am confident in my ability to identify and use suitable tools for data collection.
5. I am confident in my ability to effectively handle the collection of both big and small, structured, unstructured, numerical, quantitative, and qualitative data.

**Data Cleaning - Data Wrangling - Feature Engineering**

1. I am confident in understanding the structure and characteristics of diverse datasets.
2. I am confident in merging or joining datasets from different sources to create a unified dataset.
3. I am confident in using appropriate tools to visualize data distributions of missing values, duplicate values, inconsistency types, and outliers.
4. I am confident in my ability to inform decisions to standardize or normalize values as needed, depending on project requirements.
5. I am confident in making informed decisions on handling invalid data. Based on the visualized data distributions and stakeholders ' requirements,
6. I am confident in my ability to validate and ensure data quality after cleaning to determine whether the data is cleaned, structured, and ready for feature extraction.
7. I am confident in my ability to identify when there is a need to create subsets based on project requirements.

**Feature Selection**

1. I am confident in my ability to understand the meaning of each feature and the relationships between features by communicating with domain experts, ensuring a comprehensive understanding of the features.
2. I am confident in applying exploratory data analysis to understand the dataset better using basic statistics (Central Tendency Descriptive Summary), Principal Component Analysis (PCA), or Self-Organizing Map (SOM)
3. I confidently use descriptive statistics and machine learning measures to rank features based on their relationship with the target variable.
4. I am confident in my ability to filter the features using selection techniques like Forward Selection, Backward Elimination, Recursive Feature Elimination, or Akaike information criterion (AIC), Schwarz or Bayesian Information Criterion (SIC), and Likelihood results selection.
5. I am confident in my ability to experiment with multiple techniques to find the most effective approach for a specific model.
6. I am confident in creating new features by transforming existing ones to enhance the model outcome.
7. If necessary, I am confident in applying transformations to variables, such as transforming values from categorical to numerical data, to strengthen model efficiency.
8. I am confident in my ability in removing redundancies and selecting features to improve model efficiency.
9. I am confident in my ability to identify trends and patterns detecting anomalies or novel patterns in the data.

**Models Design**

1. I am confident in my ability to develop a model-building and validation plan.
2. I am confident in choosing the appropriate tools suited for model development.
3. I am confident in evaluating trade-offs between model complexity, interpretability, and performance.
4. I am confident in determining when to use statistical inference, simulation, classification, regression, or clustering methods.
5. I am confident in customizing my dataset to match the suitable learning algorithm (supervised, unsupervised).
6. I am confident in my ability to choose suitable machine learning or statistical models based on the nature of the problem that can minimize the loss function.
7. I am confident in identifying when sampling is needed and selecting appropriate sampling methods.

**Model Evaluation**

1. I am confident that I can scale the model to handle larger datasets.
2. I am confident in performing hyperparameter tuning and addressing potential biases or imbalances during model building.
3. I am confident in performing validation techniques (e.g., cross-validation) to assess the model's generalization ability.
4. I am confident in defining metrics for evaluating model performance, such as accuracy, precision, and recall metrics.
5. I am confident in performing diverse analyses on the developed model and its outcome, such as hypothesis testing, estimation, prediction intervals, and determining the significance of relationships.
6. I am confident in generating appropriate data visualizations for model outcomes.
7. I am confident in using the model's outcomes to inform insight.
8. I am confident in my ability to provide explanations for model outcomes.
9. I am confident in interpreting my result to the lowest denomination so that non-academic readers understand it.

**Communication and Proposed Action**

1. I am confident in connecting my results to exciting trends and literature to draw inferences when applicable.
2. I am confident in combining complex visualized structures, encompassing multidimensional and hierarchical data, to create a non-complex, meaningful, and insightful representation of our results through data storytelling.
3. I am confident in my ability to tailor visualizations to the specific needs and understanding of different audiences, including non-technical stakeholders.
4. I am confident in my ability to follow best practices for data visualization, including appropriate chart selection, color usage, and labeling.

Q1-Q7: Domain Knowledge and Research design,

Q8-Q12: Data Planning and Data collection,

Q13-Q19: Data cleaning, wrangling and Feature Engineering,

An

Q20-Q28: Feature Selection,

Q29-Q35 Model design,

Q36-Q40 Model evaluation and

Q41-Q48 Communicate and propose action.

Version II is the final version that passed A diagram of a model

Description automatically generated with medium confidence

**Demographic Information**

* 1. Personal Information
     1. Gender:

- Male

- Female

- Prefer not to say

- Other (please specify – short answer): \_\_\_\_\_\_\_\_\_\_\_

* + 1. Year of Birth (please specify – short answer): \_\_
    2. Ethnicity:

- Asian

- Black

- African American

- Hispanic/Latino

- White/Caucasian

- Native American

- Arab

- Other (please specify – short answer): \_\_\_\_\_\_

* + 1. Location:

-. Country: (please specify – short answer): \_\_

- State/Province: \_(please specify – short answer): \_\_\_\_\_\_\_\_\_\_\_

- City: (please specify – short answer): \_\_\_\_\_\_\_\_\_\_\_\_\_\_

* 1. Education:

1. Highest level of education completed:

- High school

- Some college

- Bachelor's degree

- Master's degree

- Doctorate

- Other (please specify): \_\_

1. if currently a student, What year in?

Freshman/level 1

Sophomore/level 2

Junior/level 3

Senior/level 4 \_\_\_\_\_\_\_

1. if currently a student, what is your major?

Computer Science

Business Analytics

Statistic

Data Analytics

Others (please specify): \_\_\_\_\_\_\_

1. if currently a student, provide your university name?

(Please specify): \_\_\_\_\_\_\_\_\_\_\_\_

* 1. Occupation:

a. Employment Status:

- Employed

- Unemployed

- Student

- Retired

- Other (please specify): \_\_\_\_\_\_\_\_\_\_\_\_

b. Anticipated employment Industry:

- Technology

- Healthcare

- Education

- Finance

- Manufacturing

- Other (please specify): \_\_\_\_\_\_\_\_\_\_\_\_

**Data Science Background knowledge**

1. I took statistics course before

Never

At least 1 course

More than one course

Part of a course/s

1. I took research design course before

Never

At least 1

More than one

Part of a course/s

1. I took coding course before

Never

At least 1

More than one

Part of a course/s

1. I took machine learning /AI course before

Never

At least 1

More than one

Part of a course/s

1. I took business analytics course before

Never

At least 1

More than one

Part of a course/s

1. I took data analytic course before

Never

At least 1 course

More than one course

Part of a course/s

1. Check the concept or the tool that you are familiar with or can use

Excel

Matplotlib

Pandas

NumPy

Sci-Kit Learn

Tableau

Power BI

D3.js

QlikView

Apache Flink

MongoDB

SQL

Logistic Regression

Naive Bayes

Decision Tree

Linear Regression

Neural Networks

Random Forest

Support Vector Machine (SVM)

K-Nearest Neighbors (KNN)

Reinforcement Learning

BigML

Rapidminer

KNIME

Weka

DataRobot

R Programming

Matlab

Julia

NLTK

Python

Jupyter

TensorFlow

Scrapy

SAS

ANOVA

Hypothesis Testing

Analytics Platforms

Google Analytics

Minitab

Clustering

K-Means Clustering

Principal Component Analysis (PCA)

Unsupervised Learning

Probability

Imputation

Coefficients

Central Tendency Descriptive Summary

Class Imbalance

Self-Organizing Map (SOM)

Likelihood Results Selection

Sampling

Information Criteria

Akaike Information Criterion (AIC)

Schwarz or Bayesian Information Criterion (SIC)

Variations

High Cardinality

Partitioning