Data Scientist problem statement

🧠 Problem Statement

You are provided with a dataset containing various data points, each described by a set of features. Your task is to apply unsupervised learning techniques to cluster these data points into 'n' distinct clusters. Additionally, you need to identify the most important features and their corresponding values that contribute significantly to the grouping of data points within these clusters.

Dataset Description:
About Dataset: Breast Cancer Gene Expression Profiles (METABRIC). The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database is a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples. Clinical and genomic data was downloaded from cBioPortal.

Dataframe Name: METABRIC_RNA_Mutation.csv (v1 - 8.39MB)
Dataframe Link - https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric
Paper Link - https://www.nature.com/articles/s41523-018-0056-8

The dataset consists of 1904 data points, each described by 692 features. The features may include numerical, categorical, or a combination of both. The target variable is not available since this is an unsupervised learning problem.

Objective:

1. **Clustering**: Implement an unsupervised learning algorithm to partition the data points into 'n' clusters. You should select an appropriate clustering algorithm based on the characteristics of the dataset and the problem requirements.
2. **Feature Importance**: Identify the most important features that influence the formation of clusters. Determine the relevance and significance of each feature in grouping data points together. This analysis will help you understand which features are driving the cluster assignments.
3. **Feature Values**: Determine the specific values or ranges of values for the identified important features that are associated with each cluster. In other words, find the feature-value combinations that differentiate one cluster from another.
   Note: You get Bonus Points ⭐ , if you find correct feature(s) and their value(s) with showing the evidence and validated approach.

👩‍🏫 Tasks:

1. Data Preprocessing:
   - Handle missing values, if any.
   - Normalise or standardise the features, if necessary.
   - Encode categorical variables, if applicable.
2. Clustering:
   - Choose an appropriate clustering algorithm (e.g., K-means, DBSCAN, hierarchical clustering).
   - Determine the optimal number of clusters 'n' based on a suitable method or criterion (e.g., elbow method, silhouette score).
   - Apply the chosen algorithm to cluster the data points into 'n' clusters.
3. Feature Importance Analysis:
   - Employ techniques such as feature importance scores, dimensionality reduction, or visualization methods (e.g., PCA, t-SNE) to identify the most important features.
   - Determine the relevance and contribution of each feature to the cluster assignments.
4. Feature-Cluster Associations:
   - For each identified important feature, analyse its values or value ranges that are prevalent in each cluster.
   - Create a summary or report that presents these feature-value associations for each cluster.
5. Evaluation and Documentation:

- Evaluate the quality of the clustering results using appropriate metrics (e.g., silhouette score, Davies-Bouldin index).
- Document your findings, including the cluster assignments, important features, and their associated values for each cluster.

🚚 Deliverables:

1. A report or presentation summarising your findings, including the cluster assignments, important features, and their values associated with each cluster.
2. Code and scripts used for data preprocessing, clustering, and feature importance analysis.
3. Visualisations, if applicable, to support your analysis and interpretation of the results.

Remember that the choice of clustering algorithm, feature importance analysis technique, and other specific details should be determined based on the nature and characteristics of the dataset, and the problem statement's objectives.