**Methodology**

To generate explanations for the notebook content following the CRISP-DM (Cross-Industry Standard Process for Data Mining) steps, I'll explain the process as observed from the provided content, maintaining simple academic language.

### 1. Business Understanding

This stage is about understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan to achieve the objectives.

In the context of the provided notebook, the objective seems to be analysing data from a health survey (specifically focusing on HIV services) to understand service satisfaction levels, identify gaps in service provision, and improve health services based on client feedback. The business goal could be enhancing the quality of healthcare services, ensuring services meet the needs of the community, and increasing overall satisfaction among service recipients.

### 2. Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities to get familiar with the data, identify data quality issues, discover first insights into the data, or detect interesting subsets to form hypotheses for hidden information.

The notebook indicates the use of a dataset named `imonitor\_1703.csv`, which includes responses from a health survey with various attributes like service satisfaction, awareness of services, challenges in accessing services, etc. The dataset initially contains 46,549 rows and 85 columns, but after cleaning and preprocessing, the shape changes, indicating data quality and structure adjustments.

### 3. Data Preparation

Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modelling tools.

The notebook shows extensive data preparation steps:

Removal of Unnecessary Columns: This step is crucial to streamline the analysis by eliminating data that might not contribute to the model's performance or could potentially introduce noise. For example, free-text fields requiring manual interpretation or fields with high cardinality can complicate the modelling process without adding value.

Renaming of Columns: Improving the readability and consistency of the dataset makes it easier for analysts and stakeholders to understand the data. Clear column names reduce the risk of misunderstandings or errors in data handling.

Dealing with Missing Data: High levels of missing data in certain columns can significantly impact the quality of the analysis. Removing or imputing these values ensures that the models are built on reliable and complete information, thereby improving their accuracy and reliability.

Data Type Conversion: Correcting data types, such as converting strings to datetime objects or categorizing rating scales, ensures that mathematical and logical operations during the modelling phase are applicable and meaningful. It also helps in using certain types of models that require numerical input.

Encoding of Categorical Variables: Machine learning models require numerical input, so categorical variables are encoded into a format that can be interpreted by these models. This step is essential for leveraging the predictive power of categorical data.

### 4. Modelling

In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

The notebook demonstrates the application of multiple machine learning models (CatBoost, LightGBM, XGBClassifier, RandomForestClassifier, LogisticRegression, and SVC) to predict service satisfaction levels. Trying out different algorithms helps in identifying the most effective model based on the problem at hand. Each algorithm has its strengths and weaknesses depending on the nature of the data and the specific task.

Performance Evaluation: The use of metrics like ROC AUC, accuracy, precision, recall, and F1 score allows for a comprehensive assessment of each model's performance. It helps in understanding not just the model's accuracy but also its ability to balance false positives and false negatives, which is crucial in healthcare applications.

Model Selection: Based on the comparative analysis, the best-performing model is selected. This ensures that the deployment phase uses the most effective tool available for making predictions, thereby maximizing the positive impact of the project.

### 5. Evaluation

At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

The evaluation in the notebook involves assessing model performance on a test set and examining metrics like the classification report, ROC curves, Precision-Recall curves, and confusion matrices. This helps in understanding the model's generalizability and identifying any potential issues before deployment.

Assessment on Test Set: Evaluating the model on a separate test set provides an unbiased estimate of its performance on unseen data, which is a strong indicator of its generalizability.

Use of Various Metrics: Different metrics can provide insights into different aspects of model performance. For instance, precision and recall are particularly important in healthcare settings where the cost of false negatives might be very high.

ROC and Precision-Recall Curves: These allow for the evaluation of model performance across different threshold settings, offering a more nuanced view than single-number metrics.

Confusion Matrix: Provides a clear visualization of the model's predictions compared to the actual labels, helping identify patterns in misclassifications that could guide further refinement of the model.

### 6. Deployment

In the deployment phase, the cleaned and pre-processed dataset was exported for further use, such as exploratory data analysis (EDA) in tools like Power BI. This step is significant because EDA can uncover patterns, anomalies, and relationships in the data that might not be immediately apparent, providing valuable insights for business decisions and strategy formulation.

Additionally, the best-performing model and its evaluation report were integrated into a Power BI dashboard. This inclusion allows stakeholders to interact with the findings dynamically, understand the predictive performance of the model, and explore how different variables impact service satisfaction. This holistic approach ensures that the insights derived from the model are accessible and actionable, supporting informed decision-making. The dashboard serves as a powerful tool for presenting complex analytical outcomes in an intuitive and understandable manner, facilitating the practical application of these insights in improving health services.

**System Architecture and Implementation**

This chapter describes the architecture and implementation of a system designed to clean a dataset, run a machine learning model for classification prediction, and visualize the results. The system comprises three main components: a Deepnote notebook, a Supabase Postgres database, and a Power BI dashboard.

### System Overview

The system was architected to streamline the process of analysing healthcare survey data, from initial cleaning and processing through to predictive modelling and visualization. The end goal was to enable efficient insights into service satisfaction and identify areas for improvement.

### System Architecture

The system architecture was designed with modularity, scalability, and ease of use in mind. It consists of three primary components:

1. \*\*Deepnote Notebook:\*\* Utilized for data cleaning and execution of the machine learning model.

2. \*\*Supabase Postgres Database:\*\* Served as the storage mechanism for both the cleaned dataset and the outputs of the machine learning model.

3. \*\*Power BI Dashboard:\*\* Employed to visualize the cleaned data, model outputs, and conduct exploratory data analysis (EDA).

#### [Diagram Placeholder: System Architecture Diagram] Insert a diagram that visually represents the system architecture, indicating the flow of data between the Deepnote notebook, Supabase Postgres database, and Power BI dashboard.

### System Implementation Process

#### Deepnote Notebook

The Deepnote notebook was implemented to perform initial data cleaning tasks, including the removal of irrelevant columns, handling missing values, and encoding categorical variables. Following the cleaning process, the notebook executed a machine learning model to predict service satisfaction levels based on survey responses. The model selection process involved evaluating several classifiers, with the best-performing model based on accuracy, precision, recall, and F1 score being chosen for deployment.

Key steps in the notebook implementation included:

- Importing the dataset.

- Performing exploratory data analysis to understand the data's characteristics.

- Cleaning the data to prepare it for modelling.

- Splitting the data into training and testing sets.

- Training multiple machine learning models and selecting the best performer.

- Exporting the cleaned data and model predictions to the Supabase Postgres database.

#### Supabase Postgres Database

The Supabase Postgres database was configured to store the cleaned dataset and the outputs from the machine learning model. Two primary tables were created: one for the cleaned dataset and another for the model predictions. Database triggers and views were not utilized in this implementation; however, they were considered for future work to automate updates and enhance data management.

#### Power BI Dashboard

The Power BI dashboard was developed to provide interactive visualizations of the cleaned dataset and the results of the machine learning model. The dashboard included various charts and filters to enable stakeholders to explore the data from multiple angles, including service satisfaction levels, respondent demographics, and predictors of satisfaction.

Key features of the dashboard included:

- Interactive filters for selecting specific subsets of data.

- Visualizations for model accuracy and performance metrics.

- Charts depicting key findings from the exploratory data analysis.

### Integration Process

The integration process involved exporting the cleaned dataset and model predictions from the Deepnote notebook to the Supabase Postgres database using a direct API call. Once stored in the database, the data was then connected to Power BI through its Postgres database connector. This setup allowed for real-time updates to the dashboard as new data was processed and analyzed in the notebook.

### Challenges Encountered in the Implementation

During the implementation of the system, several significant challenges were faced, most notably related to data interoperability and real-time data processing. A major obstacle encountered was the lack of interoperability between the iMonitor system, which houses the healthcare survey data, and the implemented system comprising the Deepnote notebook, Supabase Postgres database, and Power BI dashboard. The iMonitor system did not offer an API or direct integration feature for automatically exporting the data. This limitation hindered the possibility of creating a real-time application that could directly ingest data from the survey application, necessitating manual data export and import processes. To overcome this challenge, a manual process for exporting data from the iMonitor system and importing it into the Deepnote notebook was established. While this approach did not allow for real-time data updates, it ensured data availability for cleaning, analysis, and modelling. To mitigate the impact of this manual step and make the process as efficient as possible, detailed documentation was created, outlining the steps for data export and import, ensuring consistency and reducing the potential for errors.

**Results and Discussion**

### Machine learning Results

### Comparison between Homa Bay and Kilifi

Prevalence rates link: https://nsdcc.go.ke/kasf-areas-of-focus/

### Conclusion

**Conclusions, Recommendations, and Future Work**

### Conclusion

The system was successfully implemented, achieving the objective of providing a streamlined process for data analysis, from cleaning and modelling in a Deepnote notebook, through storage in a Supabase Postgres database, to visualization in a Power BI dashboard. This implementation not only facilitated a deeper understanding of the healthcare survey data but also laid the groundwork for ongoing improvements in service satisfaction analysis.

The challenges encountered during the implementation of this system underscored the importance of flexibility and adaptability in the face of technical limitations. By developing effective solutions to these challenges, the project successfully leveraged the available tools and technologies to achieve its objectives, albeit with some compromises on real-time data processing capabilities. Future enhancements will focus on exploring alternative solutions to achieve greater interoperability and automation, further streamlining the data analysis process.