**Chapter Three: Methodology**

**3.1 Introduction**

This chapter presents the methodology used for automating school dropout analysis at the University of Nairobi using Power BI. The methodology encompasses the data analysis techniques employed, such as descriptive statistics, correlation analysis, regression analysis, and other relevant methods. Additionally, it covers the Power BI dashboard design process, including the chosen visualizations and interactive features.

**3.2 Data Collection**

The goal was to gather relevant data and information to build a comprehensive and insightful Power BI dashboard for analyzing school dropout patterns and identifying potential risk factors. The objectives of the data collection and fact-finding process were as follows:

* Identify the key variables and factors related to school dropout.
* Gather data on student demographics, academic performance, socio-economic background, and other relevant information.
* Ensure data accuracy and reliability for effective analysis.
* Understand the existing data sources and availability of data for analysis.
* Identify any data gaps or limitations that need to be addressed.

Data Sources:

A variety of data sources were utilized to collect relevant information for the dropout analysis project. These sources included:

* University records and databases: Access to student records, enrollment data, academic performance, and demographic information was obtained from the University's database systems.
* Surveys and questionnaires: Customized surveys and questionnaires were administered to students, teachers, and administrative staff to collect additional data related to dropout factors.
* External data sources: External data sources, such as national education databases, government reports, and research studies, were consulted to supplement the internal data and provide a broader perspective on dropout trends.

Data Collection Process:

The data collection process involved the following steps:

* Defining the variables: A comprehensive list of variables related to school dropout was identified based on previous research and expert consultation. These variables encompassed student characteristics, academic performance, socio-economic factors, and institutional factors.
* Designing surveys/questionnaires: Surveys and questionnaires were designed to collect data on the identified variables. Care was taken to ensure clear and concise questions, and options for categorical data were properly defined.
* Data collection tools: Online survey platforms, paper-based questionnaires, and face-to-face interviews were used as data collection tools, depending on the target audience.
* Data validation and cleaning: After data collection, the collected responses were validated for accuracy and consistency. Data cleaning techniques were applied to remove any inconsistencies, missing values, or outliers.
* Data integration: The collected data from various sources were integrated into a central repository, ensuring that data were properly structured and formatted for analysis.

Fact-Finding Process

The fact-finding process involved gathering relevant information and insights from various stakeholders, including students, teachers, administrative staff, and education experts. This process included:

* Conducting interviews: Face-to-face interviews were conducted with key stakeholders to gather their perspectives on dropout factors, potential interventions, and challenges faced.
* Focus group discussions: Focus group discussions were organized to facilitate open discussions among stakeholders, encouraging the exchange of ideas and experiences related to school dropout.
* Literature review: Existing literature, research studies, and reports on school dropout were reviewed to gain a deeper understanding of the factors and strategies employed in dropout prevention.
* Benchmarking with other institutions: Comparative analysis was conducted by examining dropout analysis methodologies and best practices employed by other educational institutions to identify lessons learned and potential strategies for implementation.

The data for this study was collected from the University of Nairobi's student information system, including demographic information, academic performance records, and dropout data. The dataset covered a specific timeframe and included a representative sample of students. The data collection process involved gathering relevant information from the University's student information system and other appropriate sources. Here is a description of the data collection process:

1. Identification of Data Requirements: The first step was to identify the specific data required for the analysis. This involved determining the variables and attributes that were essential for understanding and predicting school dropout. The data requirements could include demographic information, academic performance records, socioeconomic status, student engagement, and any other factors known to influence dropout rates.
2. Data Sources: Once the data requirements were established, the appropriate data sources were identified. The primary source of data was the University of Nairobi's student information system, which contains comprehensive student records. Additional data sources such as surveys, interviews, or external databases may also be utilized if deemed necessary and available.
3. Data Access and Permissions: To access the required data, necessary permissions and approvals were obtained from relevant authorities. This ensured compliance with data protection and privacy regulations. It is important to ensure that the data collection process adheres to ethical guidelines and safeguards the confidentiality and anonymity of the individuals involved.
4. Data Extraction: After obtaining access to the data sources, the relevant data was extracted using appropriate techniques. This may involve querying databases, exporting data files, or utilizing APIs to collect data programmatically. The extracted data was then stored in a suitable format for further analysis.
5. Data Cleaning and Transformation: The collected data often requires cleaning and transformation to ensure its quality and compatibility for analysis. This step involved identifying and handling missing values, removing duplicate records, standardizing formats, and resolving inconsistencies in the data. Data cleaning techniques such as outlier detection and correction were also applied to enhance the data quality.
6. Data Integration: In some cases, data from multiple sources needed to be integrated to create a comprehensive dataset. This step involved merging or joining different datasets based on common identifiers or key fields. Data integration ensures that all relevant variables are available for analysis.
7. Data Documentation: Proper documentation of the collected data was maintained throughout the process. This documentation included details about data sources, variables, data collection methods, and any modifications or transformations applied. Documentation ensures transparency, reproducibility, and enables others to understand and validate the analysis process.

**3.3 Data Preprocessing**

Before importing the data into Power BI, a preprocessing stage was conducted to clean and transform the data. This involved handling missing values, resolving inconsistencies, and formatting the data to ensure its compatibility with Power BI. Data processing techniques are applied to prepare and transform raw data into a suitable format for analysis. These techniques ensure data quality, enhance compatibility, and enable meaningful insights to be derived from the data. Here are some common data processing techniques used in the context of automating school dropout analysis:

1. Data Cleaning: Data cleaning involves identifying and handling inconsistencies, errors, missing values, and outliers in the dataset. Techniques such as imputation, deletion, or interpolation may be used to address missing values. Inconsistencies and errors are resolved through standardization, correction, or removing erroneous data points. Cleaning the data ensures its integrity and reliability for further analysis.
2. Data Integration: Data integration combines data from multiple sources or tables into a unified dataset. This process involves matching and merging data based on common identifiers or key fields. By integrating data, relationships between different variables can be analyzed, enabling a comprehensive analysis of school dropout factors.
3. Data Transformation: Data transformation involves converting the data into a format that is suitable for analysis. This may include aggregating data, changing data types, creating derived variables, or normalizing data. Transformation techniques such as scaling, logarithmic transformation, or z-score normalization may be applied to ensure data comparability and reduce bias caused by varying scales or distributions.
4. Feature Engineering: Feature engineering involves creating new variables (features) from existing ones that capture relevant information or patterns. For school dropout analysis, feature engineering techniques may include calculating cumulative grades, creating socio-economic status indicators, or deriving variables that measure student engagement. Feature engineering enhances the dataset by incorporating additional information that can contribute to dropout prediction and analysis.
5. Data Reduction: Data reduction techniques are employed to reduce the dimensionality of the dataset without losing important information. Dimensionality reduction methods, such as principal component analysis (PCA) or feature selection algorithms, help identify the most informative variables or combinations of variables. This process simplifies the analysis, improves computational efficiency, and eliminates redundant or irrelevant variables.
6. Data Reshaping: Data reshaping involves reorganizing the data structure to facilitate analysis. This may include pivoting data, transforming wide-format data into long-format (or vice versa), or creating hierarchical structures for efficient representation. Reshaping the data allows for easier manipulation and visualization, enabling better understanding of the relationships between variables.
7. Data Sampling: Data sampling techniques are used to extract a representative subset of the data for analysis. This can help in reducing computational complexity and analyzing large datasets efficiently. Sampling methods, such as random sampling or stratified sampling, ensure that the selected subset maintains the characteristics of the original dataset.

**3.4 Descriptive Statistics**

Descriptive statistics were employed to gain a comprehensive understanding of the dataset. Measures such as mean, median, mode, standard deviation, and quartiles were calculated to summarize various variables. Frequency distributions and histograms were also used to analyze categorical and continuous variables.

Descriptive statistics were employed to gain a comprehensive understanding of the dataset by analyzing the distribution, central tendency, and variability of the variables. Here's a description of the variables and how descriptive statistics can be applied to each of them:

1. Marital Status: This categorical variable indicates the marital status of the student. Descriptive statistics can be used to determine the frequency and proportion of students in each marital status category.
2. Application Mode: This categorical variable represents the method of application used by the student. Descriptive statistics can provide information on the count or percentage of students who applied through different modes.
3. Application Order: This numerical variable indicates the order in which the student applied. Descriptive statistics can be used to calculate measures such as the mean, median, minimum, and maximum to understand the distribution and central tendency of application order.
4. Course: This categorical variable represents the course taken by the student. Descriptive statistics can reveal the frequency or proportion of students in each course category.
5. Daytime/Evening Attendance: This categorical variable indicates whether the student attends classes during the day or in the evening. Descriptive statistics can provide insights into the distribution of students attending classes in each session.
6. Previous Qualification: This categorical variable represents the qualification obtained by the student before enrolling in higher education. Descriptive statistics can reveal the frequency or proportion of students with different qualifications.
7. Nationality: This categorical variable indicates the nationality of the student. Descriptive statistics can be used to determine the count or percentage of students from different countries.
8. Mother's and Father's Qualification: These categorical variables represent the qualifications of the student's mother and father, respectively. Descriptive statistics can reveal the distribution and frequencies of different qualification levels for both parents.
9. Mother's and Father's Occupation: These categorical variables represent the occupations of the student's mother and father, respectively. Descriptive statistics can provide insights into the frequency or proportion of students belonging to different occupational categories for both parents.
10. Displaced: This categorical variable indicates whether the student is a displaced person. Descriptive statistics can reveal the count or percentage of students who are displaced.
11. Educational Special Needs: This categorical variable indicates whether the student has any special educational needs. Descriptive statistics can provide information on the frequency or proportion of students with special needs.
12. Debtor: This categorical variable indicates whether the student is a debtor. Descriptive statistics can reveal the count or percentage of students who have outstanding debts.
13. Tuition Fees Up to Date: This categorical variable indicates whether the student's tuition fees are up to date. Descriptive statistics can provide insights into the distribution of students based on their fee payment status.
14. Gender: This categorical variable represents the gender of the student. Descriptive statistics can be used to determine the count or percentage of students in each gender category.
15. Scholarship Holder: This categorical variable indicates whether the student is a scholarship holder. Descriptive statistics can reveal the frequency or proportion of students who hold scholarships.
16. Age at Enrollment: This numerical variable represents the age of the student at the time of enrollment. Descriptive statistics, such as measures of central tendency (mean, median) and dispersion (standard deviation, range), can help understand the age distribution of enrolled students.
17. International: This categorical variable indicates whether the student is an international student. Descriptive statistics can provide insights into the count or percentage of international students.

**3.5 Correlation Analysis**

Correlation analysis was performed to examine the relationships between different variables. Pearson's correlation coefficient was calculated to assess the strength and direction of linear relationships. Correlation matrices and scatter plots were utilized to visualize the correlations among variables.

**3.6 Regression Analysis**

Regression analysis was conducted to identify the factors contributing to school dropout. Multiple regression models were built, considering variables such as socioeconomic status, academic performance, and student engagement. The significance and strength of each predictor variable were assessed using statistical measures like R-squared, p-values, and coefficient estimates.

3.7 Power BI Dashboard Design

The Power BI dashboard was designed to provide an intuitive and interactive interface for school dropout analysis. The design process involved the following steps:

3.7.1 Data Import and Modeling The preprocessed dataset was imported into Power BI, and appropriate relationships between tables were established. Calculated columns and measures were created to enhance the analysis.

3.7.2 Visualizations Selection Various visualizations were chosen to effectively present the analysis results. The selection included bar charts, line charts and pie charts. These visualizations were carefully selected to represent different aspects of the dropout analysis.

3.7.3 Dashboard Layout and Composition

The dashboard layout was designed to ensure a logical flow of information and ease of navigation. The visualizations were arranged in a cohesive manner, allowing users to explore the data in a structured manner. Filters, slicers, and drill-through actions were added to enhance interactivity and enable users to focus on specific subsets of data.

3.7.4 Interactive Features

The Power BI dashboard incorporated various interactive features to enable dynamic exploration of the data. Users could interact with the visualizations by applying filters, drilling down into details, and highlighting specific data points of interest. Additionally, tooltips and data labels were used to provide additional context and insights.

**3.8 Validation and Testing**

The developed Power BI dashboard was validated and tested to ensure its accuracy and functionality. The analysis results were cross-verified with external sources, and the interactive features were thoroughly tested for usability and responsiveness.

**3.9 Ethical Considerations**

Ethical considerations were taken into account throughout the methodology. Data privacy and confidentiality were maintained by anonymizing and aggregating sensitive information. The study followed ethical guidelines and obtained necessary approvals for data usage.

**3.10 Summary**

This chapter outlined the methodology employed for automating school dropout analysis at the University of Nairobi using Power BI. The data analysis techniques, including descriptive statistics, correlation analysis, regression analysis, and others, were utilized to gain insights into the factors influencing dropout. The Power BI dashboard design process encompassed visualizations selection, interactive features, and careful dashboard layout composition to create a user-friendly and informative analysis tool.