Awarding Body: Arden University
Programme Name: MSc Data Analytics and Information Systems Management
Module Name (and Part if applicable): DAT7001: Data Handling and Decision Making- Part 2: Report
Assessment Title: Data Handling and Decision Making
Student Number: stu180056
Tutor Name: Zingsho Vashum
Word Count:

Data Handling and Decision Making
Part 2: Report

Task 2.1: Discuss data preparation process

Data Collection:

Data collection is the foundational step in any data analysis project, as it involves gathering relevant data from various sources. In our case, the primary dataset used for analysis is the Drug Review Dataset, which aggregates patient reviews sourced from online pharmaceutical review platforms. This dataset serves as a rich source of information, providing insights into patient experiences with various drugs and medical conditions.

Sources of Data:

The Drug Review Dataset is compiled from multiple online platforms where patients share their experiences with different medications. These platforms may include dedicated healthcare forums, social media platforms, and specialized review websites. By accessing data from diverse sources, we can capture a wide range of patient perspectives and experiences, contributing to the richness and comprehensiveness of our dataset.

Data Collection Process:

The data collection process involves systematically retrieving data from these online platforms while adhering to ethical and legal considerations. This may entail web scraping techniques to extract information from web pages or accessing publicly available datasets from reputable sources. Care must be taken to ensure compliance with data privacy regulations and terms of service for the platforms from which the data is collected.

Data Filtering:

Data filtering is a critical step in the data preparation process, aimed at refining the dataset to include only relevant and high-quality information for analysis. In the context of our project, data filtering involves applying criteria to exclude or retain specific data points based on their relevance and reliability.

Importance of Data Filtering:

Effective data filtering is essential for ensuring the accuracy, reliability, and validity of the analysis results. By eliminating irrelevant or erroneous data points, we can enhance the representativeness of the dataset and reduce the risk of bias in the analysis. Moreover, filtering helps streamline the data processing pipeline, allowing for more efficient analysis and interpretation of the findings.

Challenges and Considerations:

Despite its importance, data filtering can present challenges, such as balancing the need for data inclusivity with the desire for data quality. Additionally, subjective judgments may be required when determining the relevance or quality of certain data points, highlighting the importance of transparency and documentation in the filtering process. Overall, careful consideration and systematic approaches are essential to effectively filter data and prepare a robust dataset for analysis.

Data Integration:

Data integration involves combining data from different sources to create a unified dataset. In our case, the Drug Review Dataset served as the primary dataset. However, we may augment this dataset with additional sources if needed to enrich our analysis. Integration ensures that all relevant data are available for analysis, providing a comprehensive view of the subject matter.

Data Representativeness:

Analyzing data representativeness is crucial to ensure that the dataset accurately reflects the underlying population or phenomenon. In our analysis of patient sentiments towards various drugs and medical conditions, representativeness entails having a diverse and unbiased sample of patient reviews. We need to ensure that the dataset covers a wide range of drugs, medical conditions, and patient demographics to draw meaningful conclusions.

Generalizability and Limitations:

Generalizability refers to the extent to which the findings from our analysis can be applied to a broader population or context. While our dataset provides valuable insights into patient sentiments within the pharmaceutical domain, it's essential to acknowledge its limitations. For example, the dataset may not be fully representative of the entire patient population due to sampling biases inherent in online reviews. Additionally, the dataset may lack information on certain drugs or medical conditions, limiting the generalizability of our findings.

Furthermore, the dataset may contain inherent biases or inaccuracies, such as subjective patient reviews or incomplete information. These limitations need to be carefully considered when interpreting the results of our analysis and drawing conclusions.

In conclusion, the data preparation process is a critical step in any data analysis project. By collecting, filtering, and integrating data effectively, we can ensure that our analysis is based on high-quality, representative data. However, it's essential to acknowledge the limitations of the dataset and exercise caution when generalizing the findings to broader populations or contexts.

Task 2.2: Perform data modeling

1- Justify selection of inferential and/or machine learning models relevant to objectives.

Objective 1: Analyze Drug Reviews and Ratings, the goal is to understand the sentiment and effectiveness of different drugs based on patient reviews and ratings. Here are more details on the data modeling justification for this objective:

Inferential Models:

1. **Hypothesis Testing**: Hypothesis testing can be utilized to compare the average ratings of different drugs and identify statistically significant differences. By formulating hypotheses such as "Drug A has a higher average rating than Drug B," statistical tests such as t-tests or ANOVA can be conducted to evaluate these hypotheses and determine whether observed differences in ratings are likely due to chance or represent true differences in drug effectiveness as perceived by patients.

```
75%    173826.750000      10.000000      36.000000
max    232291.000000      10.000000    1291.000000
```

In [22]:

```
# Hypothesis testing: Compare ratings of different drugs
# Example: Compare ratings of Levonorgestrel and Etonogestrel
levonorgestrel_ratings = df_cleaned[df_cleaned['drugName'] == 'Levonorgestrel']['rating']
etonogestrel_ratings = df_cleaned[df_cleaned['drugName'] == 'Etonogestrel']['rating']
```

In [23]:

```
# Perform t-test
t_stat, p_value = ttest_ind(levonorgestrel_ratings, etonogestrel_ratings)
print("\nHypothesis Testing Results:")
print("T-Statistic:", t_stat)
print("P-Value:", p_value)

if p_value < 0.05:
    print("The difference in ratings between Levonorgestrel and Etonogestrel is statistically significant.")
else:
    print("There is no statistically significant difference in ratings between Levonorgestrel and Etonogestrel.")
```

```
Hypothesis Testing Results:
T-Statistic: 20.689647449791543
P-Value: 2.4885341119723292e-92
The difference in ratings between Levonorgestrel and Etonogestrel is statistically significant.
```

2. **Correlation Analysis**: Correlation analysis can explore the relationships between drug ratings and other variables such as the condition being treated. By examining the correlation between drug ratings and conditions, insights can be gained into which drugs are perceived as more effective for treating specific conditions, and whether certain conditions are associated with higher or lower ratings overall.

In [24]:
```
# Correlation Analysis: Correlation between drug ratings and conditions
# Convert condition labels to numeric values for correlation analysis
df_cleaned['condition_code'] = df_cleaned['condition'].astype('category').cat.codes

# Calculate the correlation between drug ratings and condition codes
rating_condition_corr = df_cleaned['rating'].corr(df_cleaned['condition_code'])

print("\nCorrelation between Drug Ratings and Condition Codes:", rating_condition_corr)
```

```
Correlation between Drug Ratings and Condition Codes: 0.05103031419545069
```

Machine Learning Models:

1. **Sentiment Analysis with Supervised Learning**: Supervised learning models such as Support Vector Machines (SVM), Naive Bayes, or deep learning architectures like Recurrent Neural Networks (RNNs) can be trained to classify reviews into positive, negative, or neutral sentiments based on their textual content. By analyzing the sentiment distribution for each drug, healthcare providers can gain insights into patient perceptions and sentiment towards different drugs.
2. **Topic Modeling**: Topic modeling techniques like Latent Dirichlet Allocation (LDA) can uncover latent topics or themes within drug reviews, allowing for the identification of common issues, concerns, or benefits associated with specific drugs. By associating drugs with prevalent topics, healthcare providers can better understand patient perspectives on drug effectiveness and side effects across different conditions.
3. **Regression Analysis**: Regression models can explore the relationship between drug ratings and other predictors such as review length, frequency of use, or demographic factors. By examining how these factors influence drug ratings, healthcare providers can identify potential drivers of patient satisfaction and tailor interventions accordingly.

Apply statistical tools and report initial outcomes:

1: Data Preparation

- Import the dataset containing drug reviews and ratings into Jupyter notebook.
- Use Pandas to read the data and ensure it is clean and properly formatted.

2: Descriptive Statistics

- Use Pandas to calculate basic descriptive statistics for the drug ratings, such as mean, median, mode, standard deviation, minimum, and maximum values.
- Visualize the distribution of ratings using Matplotlib or Seaborn.

3: Comparison of Drug Ratings

- Compare the ratings of different drugs using Pandas to identify those with the highest and lowest average ratings.
- Conduct statistical tests using libraries like SciPy to determine if there are significant differences in ratings between different drugs or drug categories.

4: Reporting Initial Outcomes

- Report the descriptive statistics of drug ratings, including key findings such as the average rating, variability, and distribution of ratings across drugs.
- Summarize any significant differences in ratings between drugs or drug categories, along with corresponding p-values from statistical tests.
- Provide visualizations such as histograms or box plots to illustrate the distribution of ratings and differences between drugs.

Example Output:

- Average Rating: 4.2 (SD = 0.8)
- Drug A: Mean Rating = 4.5, Drug B: Mean Rating = 3.8 ($p < 0.05$)
- Histogram showing the distribution of ratings, indicating a skew towards higher ratings.

Explanation what the decision in question should be, based on these outcomes.

Objective 2: most relevant inferential and/or machine learning models to Identify Trends and Patterns in Patient Reviews

Selection of Models:

1. **Time Series Analysis**: Time series analysis is highly relevant for identifying trends and patterns in patient reviews over time. Techniques such as decomposition, smoothing, and forecasting can help uncover temporal variations, seasonality, and long-term trends in patient sentiment and satisfaction.
2. **Topic Modeling**: Topic modeling techniques like Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF) can be used to identify latent topics or themes within patient reviews. This allows for the discovery of common issues, concerns, or experiences shared by patients, helping to uncover underlying trends and patterns.

Justification of Models:

1. **Time Series Analysis**: Patient reviews are often collected over time, making time series analysis a natural choice for identifying temporal trends and patterns. By analyzing how patient sentiments evolve over time, healthcare providers can detect emerging issues, track the impact of interventions, and identify seasonal variations in patient experiences.

2. **Topic Modeling**: Patient reviews may cover a wide range of topics and concerns related to healthcare experiences. Topic modeling allows for the automatic discovery of themes or topics within the reviews, enabling healthcare providers to understand the most common issues faced by patients and identify patterns in patient feedback across different conditions, treatments, or healthcare providers.

Modeling Process:

1. **Data Preprocessing**: Clean and preprocess the patient review data, including text normalization, removing stop words, and performing lemmatization or stemming.
2. **Time Series Analysis**: Apply time series analysis techniques to the patient review data, such as decomposition to separate trends, seasonality, and noise. Use smoothing techniques to visualize temporal patterns and trends in patient sentiment over time. Finally, use forecasting methods to predict future trends in patient reviews.
3. **Topic Modeling**: Apply topic modeling techniques to identify latent topics or themes within the patient review data. Analyze the distribution of topics over time to uncover temporal trends in patient concerns or experiences. Visualize the results to identify patterns and insights.

Evaluation and Validation:

1. **Time Series Analysis**: Evaluate the performance of time series models using metrics such as mean absolute error (MAE) or root mean square error (RMSE) for forecasting accuracy. Validate the models using techniques such as cross-validation to ensure robustness.
2. **Topic Modeling**: Evaluate the coherence and interpretability of the identified topics using metrics such as topic coherence scores. Validate the models by assessing the relevance of the identified topics to patient experiences and healthcare domains.

Objective 3: a comprehensive set of modeling techniques for predicting drug efficacy and patient satisfaction:

Inferential Models:

1. **Time Series Analysis**: Time series analysis can be applied to examine temporal trends in drug reviews and ratings over time. Techniques such as decomposition, smoothing, and forecasting can help identify patterns such as seasonality, trends, and cycles in the data. This analysis can provide insights into how the volume of drug reviews has changed over the years and whether there are any recurring patterns or trends.
2. **Regression Analysis**: Regression models can be used to predict patient outcomes based on various predictors such as treatment plans, demographics, and medical history. For example, linear regression can predict continuous outcomes like blood pressure changes, while logistic regression can predict binary outcomes like treatment success or failure.

3. **Survival Analysis**: Survival analysis techniques like Cox proportional hazards model can be employed to analyze time-to-event data, such as time until relapse or time until medication discontinuation. This can help identify factors influencing patient adherence and treatment persistence.

Machine Learning Models:

1. **Predictive Analytics for Temporal Trends**: Machine learning models, such as autoregressive integrated moving average (ARIMA) or Long Short-Term Memory networks (LSTM), can be used to forecast future trends in the volume of drug reviews. By training these models on historical review data, they can predict changes in review volume over time, allowing healthcare providers to anticipate fluctuations and allocate resources accordingly.
2. **Clustering Analysis for Seasonal Variations**: Clustering algorithms can group drug reviews based on temporal patterns, allowing for the identification of seasonal variations in review volume. By clustering reviews based on temporal features such as month or season, healthcare providers can identify periods of increased or decreased review activity and tailor interventions accordingly.
3. **Decision Trees and Random Forests**: Decision tree-based models can predict patient outcomes by partitioning the data based on features such as treatment plans, medication adherence, and patient demographics. Random Forests can improve predictive accuracy by combining multiple decision trees.
4. **Gradient Boosting Machines (GBM)**: GBM models sequentially build decision trees to minimize prediction errors, making them suitable for predicting patient outcomes based on complex interactions between predictors.
5. **Survival Analysis with Machine Learning**: Machine learning techniques like Random Survival Forests or Deep Learning for Survival Analysis can be used to analyze time-to-event data while capturing complex relationships between predictors and survival outcomes.

Model Evaluation and Validation:

- Techniques such as concordance index, calibration curves, and AUC-ROC can assess the performance of predictive models in predicting patient outcomes over time.

Interpretability and Explainability:

- Methods such as feature importance analysis and partial dependence plots can provide insights into the factors influencing treatment effectiveness and adherence.

Objective 4: most relevant Selection and justification of the inferential and machine learning models, for Enhance Data Analytics and Decision-Making Processes

Selection of Models:

1. **Regression Analysis**: Regression models can be utilized to analyze the relationships between variables and make predictions about future outcomes. They are useful for understanding how different factors influence various aspects of healthcare data, such as patient outcomes, resource utilization, or treatment effectiveness.
2. **Classification Models**: Classification models, such as logistic regression or decision trees, can be employed to categorize data into different classes or groups. They are valuable for tasks like patient risk stratification, disease diagnosis, or predicting treatment response based on patient characteristics.
3. **Clustering Analysis**: Clustering algorithms, like k-means or hierarchical clustering, can group similar data points together based on their characteristics. This approach is beneficial for identifying patterns or subgroups within healthcare data, such as patient cohorts with similar clinical profiles or response to treatment.
4. **Time Series Forecasting**: Time series forecasting models, including ARIMA or LSTM networks, can predict future trends or values based on historical data. They are essential for anticipating changes in patient outcomes, resource demands, or healthcare utilization over time.

Justification of Models:

1. **Regression Analysis**: Regression models are well-suited for analyzing healthcare data due to their ability to quantify relationships between variables. They enable healthcare providers to understand the impact of different factors on patient outcomes or operational metrics, facilitating evidence-based decision-making.
2. **Classification Models**: Classification models allow for the categorization of healthcare data into meaningful groups, aiding in risk assessment, disease diagnosis, or treatment planning. By accurately classifying patients or conditions, healthcare providers can tailor interventions to individual needs and improve patient outcomes.
3. **Clustering Analysis**: Clustering algorithms help uncover hidden structures within healthcare data, enabling the identification of patient subgroups or patterns that may not be apparent through manual inspection. This information can guide personalized treatment strategies, resource allocation, or intervention planning.
4. **Time Series Forecasting**: Time series forecasting models provide valuable insights into future trends and patterns in healthcare data, allowing healthcare providers to anticipate changes and proactively adjust strategies or interventions. This proactive approach enhances decision-making and improves the efficiency of healthcare delivery.

Modeling Process:

1. **Data Preprocessing**: Clean and preprocess the healthcare data, including handling missing values, encoding categorical variables, and scaling numerical features.
2. **Model Training**: Train the selected inferential and/or machine learning models using historical healthcare data. Tune model hyperparameters and optimize performance using appropriate evaluation metrics.
3. **Model Evaluation**: Evaluate the performance of the trained models using relevant metrics such as accuracy, precision, recall, F1-score, or mean absolute error (MAE), depending on the task and model type.

4. **Model Interpretation**: Interpret the results of the trained models to gain insights into the relationships between variables, identify important features or predictors, and understand the implications for decision-making processes.

Task 2.3: Present further outcomes

Visualize outcomes with charts and tables showing identified and analyzed relationships.

Interpret demonstrated results.

Task 2.4: Propose recommendations

Recommend implementation, acceptance, and assessment of the decision.

Discuss how the decision contributes to strategic management of the chosen organization or project.