A4. Association Rules

## Association Rules for the Diabetes Dataset

#### Association Rule Mining

* Association Rule Mining is a method to discover interesting relationships (associations) among a large set of data items. It is commonly used in market basket analysis to find product purchase patterns.
* Support: The proportion of transactions in the dataset that contain a specific itemset. It indicates how frequently the itemset appears in the dataset.
* Confidence: The likelihood that the consequent (Y) is found in transactions under the condition that the antecedent (X) is present. It is a measure of the rule's reliability.
* Lift: Measures how much more likely the consequent (Y) is to be found in transactions that contain the antecedent (X) compared to a random transaction. A lift value greater than 1 indicates a positive correlation between X and Y.

#### Apriori Algorithm

* The Apriori algorithm is used to identify frequent itemsets and derive association rules. It uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function to extend them.

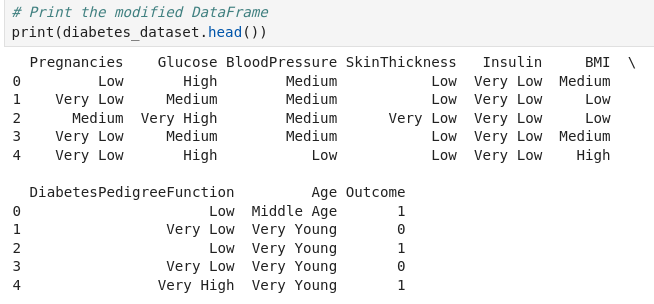
#### Data Preprocessing

* Continuous features are binned into categorical values to apply association rule mining. Binning converts continuous data into intervals, making it suitable for finding patterns in categorical data.

### 2. Detailed Data Analysis

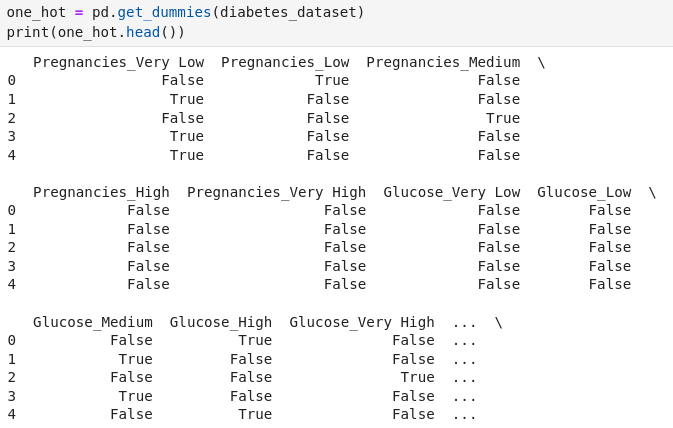
#### Data Preprocessing

* Binning: Continuous features like Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age were binned into categories (e.g., Very Low, Low, Medium, High, Very High) to convert them into categorical data.



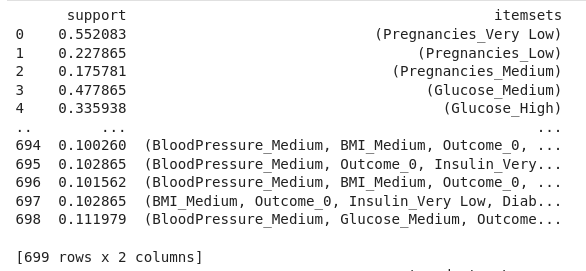
#### One-Hot Encoding

* One-hot encoding: Converts categorical data into a format suitable for the Apriori algorithm by creating binary columns for each category.



#### Applying Apriori Algorithm

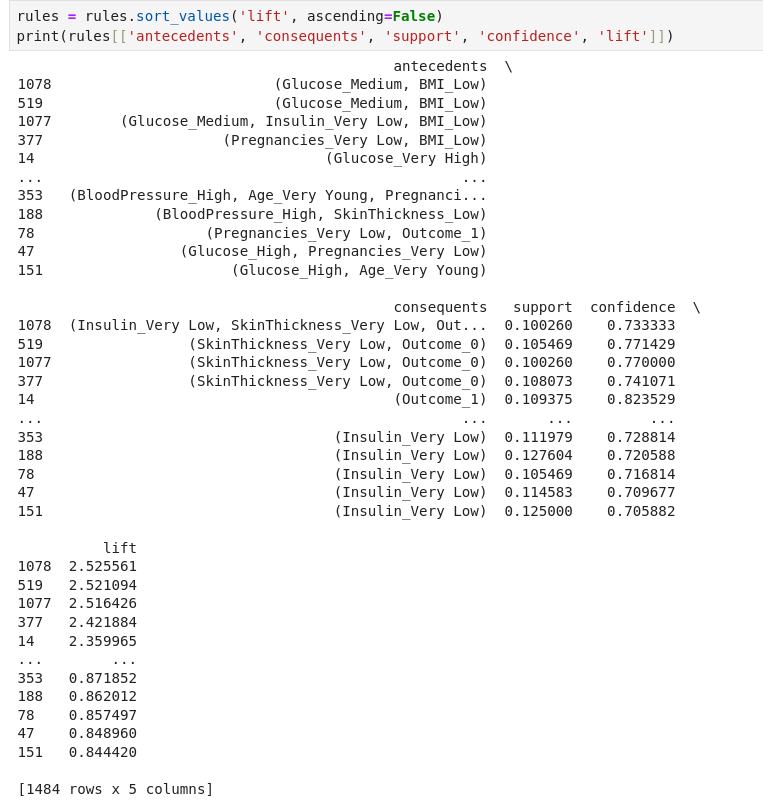
* Frequent Itemsets: Itemsets that appear frequently in the dataset above a minimum support threshold.



* Association Rules: Rules derived from frequent itemsets with confidence and lift measures to identify interesting relationships.

#### Results

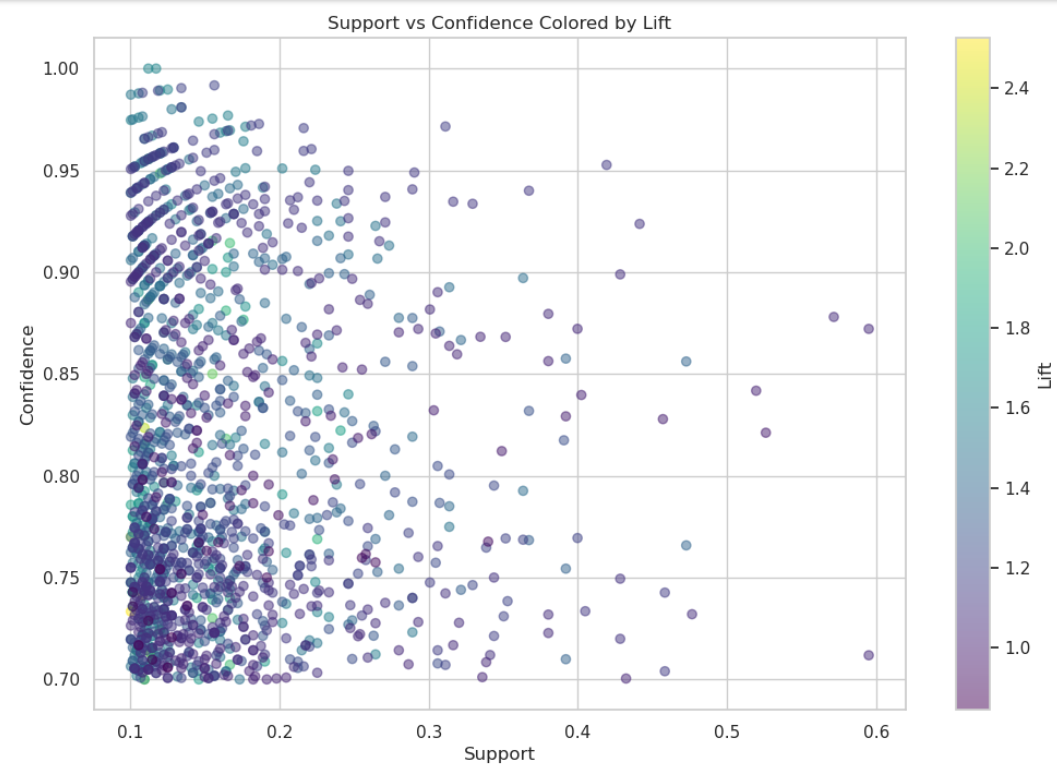
* Top Rules: Sorted by lift, confidence, and support to identify the most significant patterns.



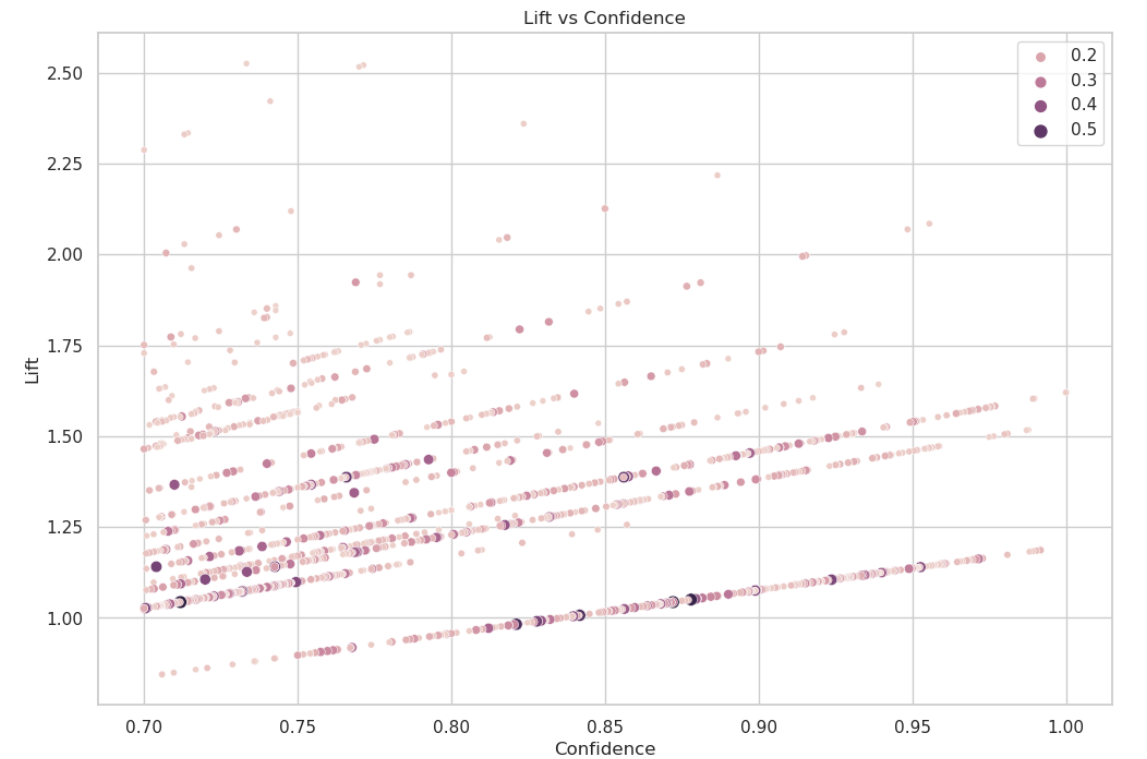
### 3. Interpretation of Results for a Medical Professional

#### Scatter Plots

1. Support vs. Confidence Colored by Lift:
   * Shows the relationship between support and confidence for each rule.
   * Higher lift values indicate stronger associations.
   * Example: A rule with high support, confidence, and lift indicates a strong and frequent relationship in the data.



1. Lift vs. Confidence:
   * Shows how lift varies with confidence.
   * Helps identify rules that are not only reliable (high confidence) but also significant (high lift).



#### Heatmap

* Heatmap of Support for Top 20 Association Rules:
  + Visual representation of support for the most significant rules based on lift.
  + Allows quick identification of the strongest associations.

### 

### 4. Further Explanations

#### Binning Continuous Features

* Purpose: Convert continuous data into categorical intervals to apply association rule mining.
* Reason: Association rule mining algorithms like Apriori are designed for categorical data. Binning helps in transforming continuous features into a format suitable for these algorithms.

#### One-Hot Encoding

* Purpose: Transform categorical data into a binary format required by the Apriori algorithm.
* Reason: The Apriori algorithm works with binary (presence/absence) data, so one-hot encoding is essential for converting categorical data into this format.

#### Filtering and Sorting Rules

* Purpose: Identify the most significant rules that provide meaningful insights.
* Reason: Sorting by lift, confidence, and support helps in prioritizing rules that are both strong (high lift) and reliable (high confidence).

### Complete Process Overview

1. Data Loading and Inspection: Understanding the dataset structure and features.
2. Data Preprocessing: Binning continuous features and one-hot encoding categorical features.
3. Applying Apriori Algorithm: Finding frequent itemsets and generating association rules.
4. Filtering and Sorting Rules: Prioritizing the most significant rules based on lift, confidence, and support.
5. Visualization: Creating scatter plots and heatmaps to visualize and interpret the rules.

By following these steps, the process ensures that the dataset is appropriately prepared for association rule mining, significant patterns are identified, and the results are presented in an interpretable format for medical professionals.

### Data Interpretation

For instance, a rule like {Glucose=Very High} -> {Outcome=1} with high support, confidence, and lift means:

* Support: A significant portion of the dataset has very high glucose levels and diabetes.
* Confidence: Patients with very high glucose levels are highly likely to have diabetes.
* Lift: This relationship is much stronger than what would be expected by random chance.

## Jupyter Notebook source code:

<https://github.com/alexban14/DataMining_Diabetes_DS>

## Resources utilized:

https://www.datacamp.com/tutorial/association-rule-mining-python

<https://www.kaggle.com/code/mervetorkan/association-rules-with-python>