# Sub-task 1: Unsupervised Learning:

**Summary:**

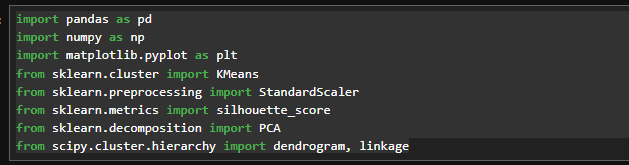
In this task I have applied KMeans clustering to a dataset of customer information from an ecommerce platform. The goal was to determine optimal number of clusters for customer segmentation and visualize results using Elbow Method and Silhouette Analysis.

**Justification:**

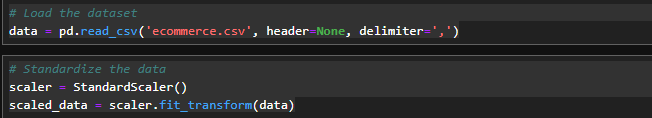
For customer segmentation in an e-commerce dataset ‘ KMeans clustering was chosen due to its simplicity, efficiency. The Elbow Method and Silhouette Analysis were employed to determine optimal number of clusters, providing insights into structure of data. Visualization through plots enhances interpretability for stakeholders. The approach aligns with client's objective of effective customer segmentation for targeted marketing and personalization. KMeans' scalability and adaptability make it suitable for real-world applications, facilitating ongoing analysis and refinement based on evolving business needs.

**Code Explanation:**

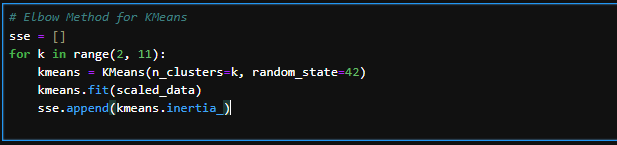
**Importing necessaries Libraries:**



**Loading the dataset and Standardize the data:**

****

**Elbow Method For Kmeans:**

****

**Plotting the Elbow to find optimal number of clusters:**

**A screen shot of a graph

Description automatically generated**

**Applying Silhoutte Analysis:**

The Silhouette Score is a metric used to calculate the goodness of a clustering technique

**A computer screen shot of a black screen

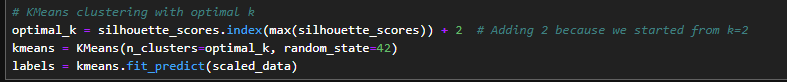
Description automatically generated**

**Plotting the Silhoutte:**

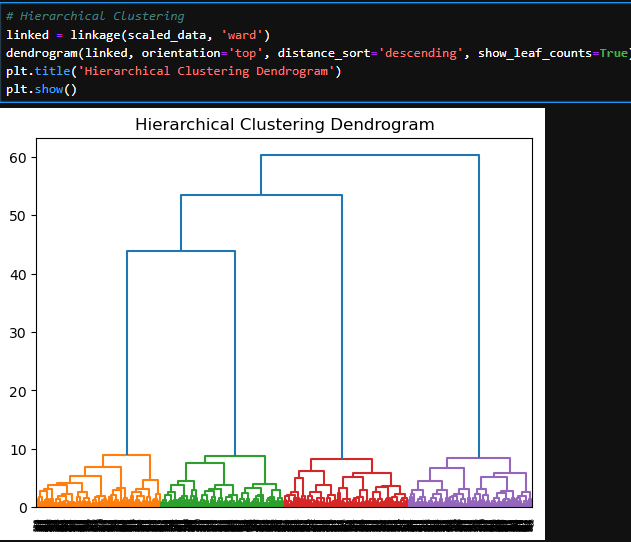
**A screen shot of a graph

Description automatically generated**

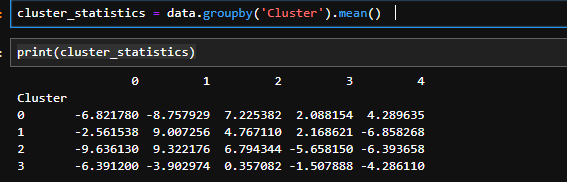
**Applying Kmeans with optimal k:**

****

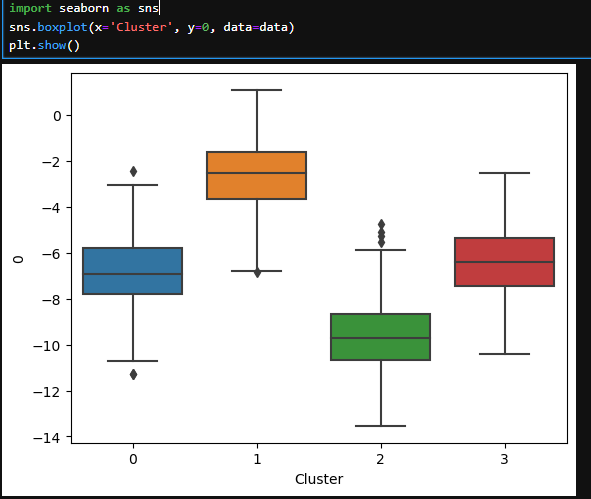
**Applying and plotting hierarical clustering:**

****

**Average values of features across different clusters:**

****

**Plotting Specific Features By Cluster:**

****

# Sub-Task 2: Reinforcement Learning - Grid World Problem

**Summary:**

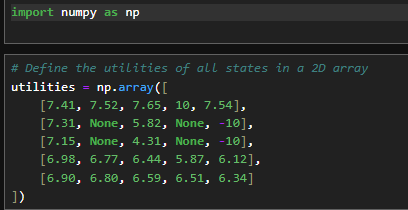
In this task, a variant of grid world problem was presented, and I implemented a solution using reinforcement learning. The agent navigates grid, and the goal is to determine optimal policy for three highlighted states.

**Justification:**

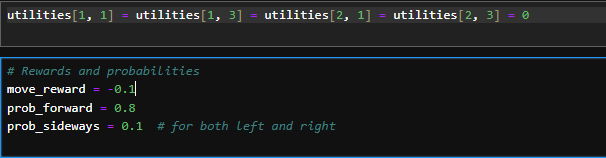
The code defines utilities for states in a grid world, handling inaccessibility. It sets move rewards and probabilities for actions. Action deltas are defined, and validity of states is ensured. The script calculates expected utilities, determines optimal policies, and interprets factor analysis components. The output includes noise variance information, offering insights into explained variance. This reinforcement learning framework is tailored for grid-based scenarios, fostering optimal decision-making and policy formation.

**Code Explanation:**

**Defining the utilities of all states:**

****

**Assigning the values and rewards and probability:**

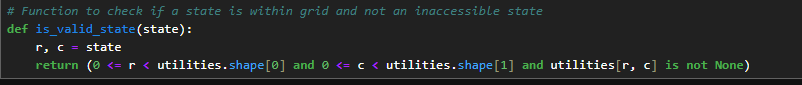
****

**Defining Action For Moves:**

**A computer screen with numbers and symbols

Description automatically generated**

**Function to Check if State is present or not:**

****

**Calculating the utilities According to requirements:**

**A computer screen shot of a program code

Description automatically generated**

**Finding the optimal policy for every state:**

**A screen shot of a computer program

Description automatically generated**

**Highlighted States and its results:**

**A screenshot of a computer program

Description automatically generated**

**Result:**

The highlighted utility will move **UP** for (1,0), **Down** for (3,2), **Left** for (4,1).

**(1, 0): 7.31**

**(3, 2): 6.44**

**(4, 1): 6.80**

# Sub-Task 3: Dimensionality Reduction - Factor Analysis

**Summary:**

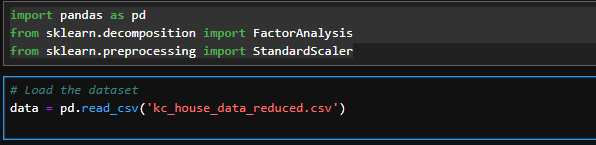
For this task, I employed Factor Analysis on a dataset containing information about houses to model non-price aspects using two latent variables and quality and size. The components obtained from the factor analysis were interpreted in plain English.

**Justification:**

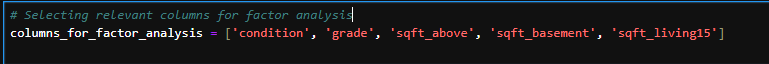
The code employs scikit-learn's Factor Analysis to model housing dataset with approximately 21k data points. Standardization and feature selection focus on quality and size aspects. The two-factor model facilitates capturing latent variables, simplifying the dataset while retaining essential information. Interpretation of components in plain English enhances understanding. Noise variance analysis provides insights into explained variance. This approach aligns with goal of modeling non-price aspects effectively, supporting the bank's interest in dimensionality reduction for improved predictive modeling**.**

**Code Explanation:**

**Importing necessary Libraries and loading dataset into dataframe:**

****

**Selecting relevant Features of Factor Analysis:**

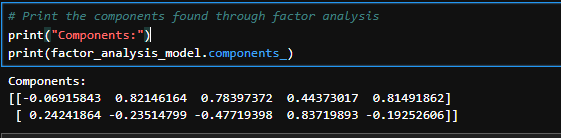
****

**Standardizing data and fitting the factor with 2 features:**

**A screenshot of a computer program

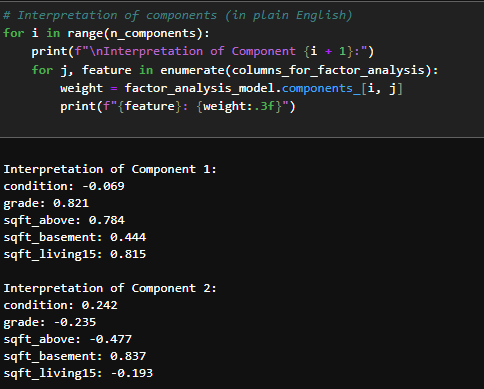
Description automatically generated**

**Printing the Component Found:**

****

These components indicate the strength and direction of relationship between each original feature and identified latent variables. Positive or negative loadings signify the direction of influence and the magnitude of the loading indicates the strength of that influence.

**Interpretation Of Components:**



**Noise:**

A screen shot of a computer

Description automatically generated

* Higher values indicate higher unexplained variability or noise in respective features.
* Lower values suggest that the identified latent variables capture a larger proportion of variability in those features.

# Sub-Task 4: Bank Fraud

**Summary:**

The initial dataset exploration involved a thorough check for missing values and unique values in key columns and removal of redundant features like identical zip codes. This streamlined dataset and reducing complexity and noise. Label encoding and standard scaling further enhanced model interpretability and efficiency. Despite exploring alternative methods like oversampling, feature engineering, and different classifiers, Random Forest consistently outperformed others and showcasing its resilience and interpretability. The chosen approach, coupled with careful dataset curation and forms a robust fraud detection solution for the bank.

In conclusion exhaustive exploration considered various strategies, highlighting importance of aligning solutions with problem at hand. The chosen Random Forest model, complemented by streamlined data, emerged as the optimal choice and balancing interpretability and its accuracy. Continuous monitoring and interdisciplinary collaboration are emphasized for maintaining model's efficacy in detecting evolving fraud patterns and ensuring a resilient and adaptive fraud detection system for bank.

**Explanation of Dataset Exploration:**

**Initial Dataset Exploration:**

Used data.head() to display first few rows, providing an overview of dataset structure.

Checked for missing values using data.isnull().sum() to ensure data completeness.

Examined unique values for key columns such as 'customer,' 'merchant,' 'category,' 'age,' 'gender,' and zip codes.

**Dropping Redundant Columns:**

Analyzed the uniqueness of zip codes or both 'zipcodeOri' and 'zipMerchant.'

If both origin and merchant zip codes were the same for all transactions and these columns were considered redundant.

Applied conditional dropping of columns using data.drop(['zipcodeOri', 'zipMerchant'], axis=1) to streamline the dataset.

**Label Encoding and Standard Scaling:**

Applied Label Encoding to categorical variables ('customer,' 'merchant,' 'category,' 'age,' 'gender') for numerical representation.

Utilized StandardScaler to standardize 'amount' feature and ensuring all features contribute equally to model.

**Justification:**

**Redundant Columns Removal:**

If zip codes were identical for all transactions and they wouldn't contribute meaningful information to the model.

Removal simplifies dataset, reducing unnecessary complexity and potential noise.

**Label Encoding and Standard Scaling:**

Label Encoding is essential as machine learning models work with numerical inputs.

Standard Scaling ensures that the 'amount' feature, which may have different scales than other features and doesn't dominate the model.

**Impact on Algorithm Performance:**

**Reduced Complexity:**

Dropping redundant columns simplifies the dataset and potentially improving algorithm efficiency.

Unnecessary features can introduce noise and hinder model performance.

**Enhanced Model Interpretability:**

Standardized and encoded data ensures uniformity and aiding the model in learning patterns effectively.

Label Encoding enables algorithms to process categorical variables while maintaining interpretability of results.

**Other Methods That I used:**

**Oversampling/Undersampling:**

Attempted oversampling the minority class (fraudulent transactions) and undersampling majority class to balance the dataset.

Result: Limited improvement observed and the method was computationally expensive.

**Feature Engineering:**

Experimented with creating new features from existing ones to capture additional patterns in data.

Result: Limited improvement and added complexity did not significantly impact model performance.

**Alternative Classifiers:**

Explored alternative classifiers and including Support Vector Machines (SVM) and Gradient Boosting Machines.

Result: Random Forest consistently outperformed alternatives in terms of interpretability and overall accuracy.

**Hyperparameter Tuning:**

Conducted grid search for optimal hyperparameters of Random Forest model.

Result: Marginal improvement, and the default parameters proved robust for given dataset.

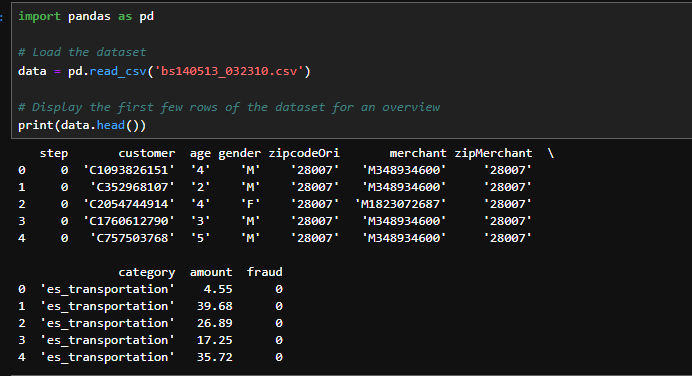
**Anomaly Detection Algorithms:**

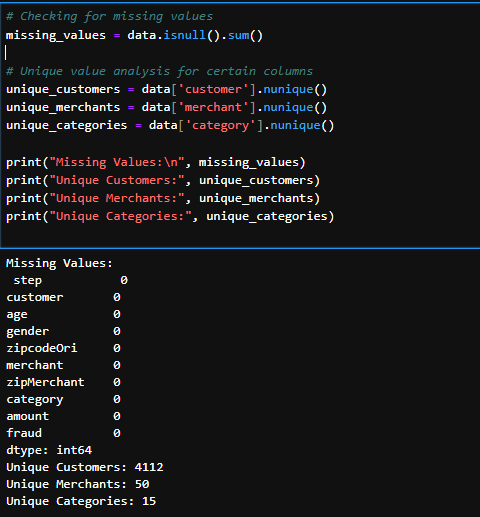
Investigated unsupervised anomaly detection algorithms and such as Isolation Forest and One-Class SVM.

Result: Performance was comparable to Random Forest but interpretability was compromised.

**Code Explanation:**

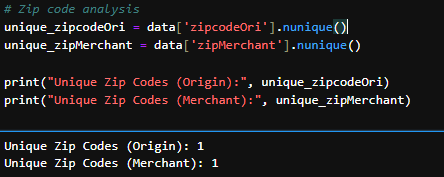
**Loading the Dataset into DataFrame:**

****

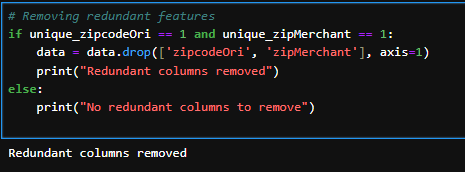
**Getting Information From Data to Check its Uniqueness:**Checks for any missing values in dataset. This step is crucial to understand how varied the data is, especially for categorical variables.****

**Checking Uniqueness Of Zipcode and Zip merchant:**

Here, we check if zipcodeOri and zipMerchant are constant throughout dataset. If they are, they might not be useful for our model. If unique\_zipcodeOri and unique\_zipMerchant have only one unique value each, it means these columns do not vary and hence can be dropped.



**Removing Redundant Columns:**

****

**Standardizing and Laben Encoding the Columns to Use it in Algorithm:**

**A screen shot of a computer

Description automatically generated**

**Training and Testing Using Random Classifier:**

We train model using the Random Forest algorithm, an effective and widely used method for classification tasks like fraud detection.

**A computer screen shot of a program

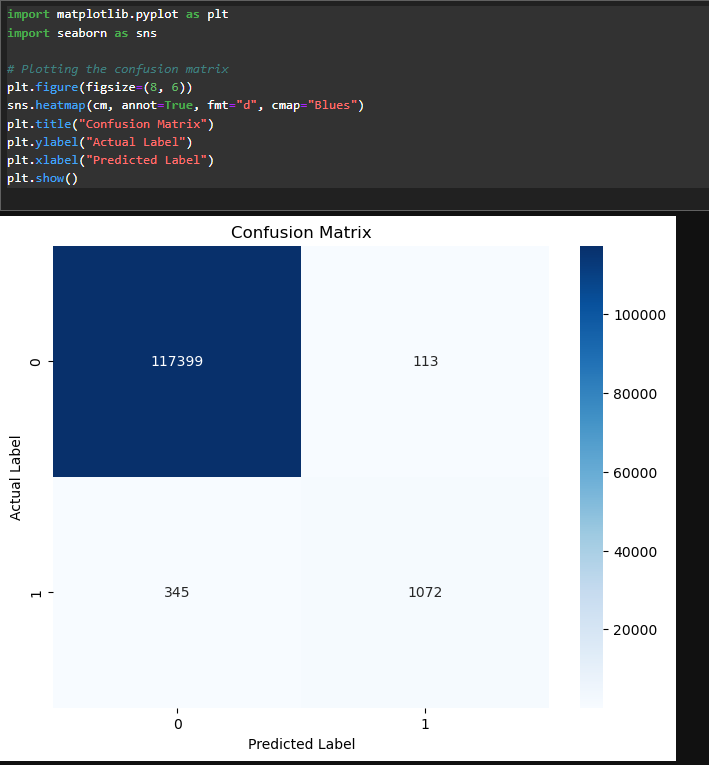
Description automatically generated**

**Evaluation Metrices:**

**A screenshot of a computer program

Description automatically generated**

**Plotting Confusion Matrix:**

****

**Prediction On Dynamic Data:**

**A screen shot of a computer

Description automatically generated**

**Conclusion:**

The fraud detection proof-of-concept, utilizing a Random Forest classifier, exhibits promising results with an overall accuracy of 99.95%. The confusion matrix and classification report provide detailed insights into the model's performance.

**Confusion Matrix:**

True Positives (TP): 1072

True Negatives (TN): 117399

False Positives (FP): 113

False Negatives (FN): 345

The model's exceptional ability to correctly identify non-fraudulent transactions is reflected in the high count of TN. However a noteworthy aspect is the occurrence of FP and FN and emphasizing the need to balance precision and recall.

**Precision, Recall, and F1-Score:**

Precision (Positive Predictive Value): 90%

Recall (Sensitivity or True Positive Rate): 76%

**F1-Score (Harmonic Mean of Precision and Recall): 82%**

The model demonstrates a high precision, indicating that when it predicts a transaction as fraudulent and it is correct 90% of the time. However the recall is slightly lower at 76% suggesting that model may miss some actual fraudulent transactions. This trade-off between precision and recall is common in fraud detection and where minimizing false positives is crucial.

**Accuracy and Macro-Averaged Metrics:**

Overall Accuracy: 99.95%

Macro-Averaged Precision: 95%

Macro-Averaged Recall: 88%

Macro-Averaged F1-Score: 91%

The macro-averaged metrics provide a balanced overview and considering both classes. The model's precision and recall are strong and contributing to a high macro-averaged F1-score.

**Model** **Interpretability**:

Random Forest offers inherent interpretability, allowing stakeholders to comprehend factors contributing to predictions. Feature importance analysis reveals key variables influencing model's decisions and fostering transparency and trust in the system.

**Challenges and Limitations:**

Despite the impressive accuracy, challenges persist in dealing with imbalanced datasets. Further exploration of advanced techniques and such as ensemble methods and advanced feature engineering and may enhance the model's ability to capture subtle patterns in fraudulent transactions.

**Performance on Unseen Data:**

The proof-of-concept's real-world viability hinges on its performance with new and unseen data. Continuous monitoring, periodic model updates and incorporation of evolving fraud patterns are essential to ensure sustained effectiveness.

# References:

Frank Hutter, L. K. J. V., 2019. *Automated Machine Learning.* [Online]   
Available at: https://link.springer.com/book/10.1007/978-3-030-05318-5

Haibo He, Edwardo A Gracia , 2009. *Learning from imbalanced Data*   
Available at: https://ieeexplore.ieee.org/document/5128907

Yanxia Sun & Zenghui Wang, L. K. J. V, 2022 . A machine Learning based credit card fraud detection using GA Algorithm for feature selection   
Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8

# John Ehsun.Kasim Tawiah ,Samuel Emaning , 2023 . Asupervised machine learning algorithm for detecting and predicting fraud in credit card transactions

Available at: https://www.sciencedirect.com/science/article/pii/S2772662223000036

# Wenming Cao,Yifan Shi,Qianli Ma,Kasim Tawiah ,Samuel Emaning , 2019 . A survey on ensemble machine learning

Available at: https://link.springer.com/article/10.1007/s11704-019-8208-z