1. Project Title:

Prediction of cost of living across U.S. Regions using a data mining approach

2. Names and Emails of Authors:

Sasank Sribhashyam: sasank.sribhashyam-1@ou.edu
Venkat Tarun Adda: venkat.tarun.adda-1@ou.edu
Hima Deepika Mannam: hima.deepika.mannam-1@ou.edu

3. Category and Objectives of the Project:

Category: Implement N algorithms are for N different data mining tasks (e.g. one algorithm is for clustering, one algorithm is for classification, one algorithm is for association rule mining, etc.), implement those N algorithms from scratch for your application.

Project Intent: To analyze the **cost of living across U.S. regions** by identifying patterns, predicting costs, and categorizing regions based on various socio-economic factors.

Objectives:
Clustering Objectives:

- **Identify Cost of Living Outliers:** Detect and analyze regions that have significantly higher or lower costs of living compared to their geographically or economically similar counterparts. This can help identify regions with unique economic or social policies affecting the cost of living.
- **Cluster Regions Based on Changes Over Time:** Use time-series clustering algorithms to group regions based on the trends and changes in cost of living over a specific period. This will help understand how economic changes affect different regions similarly or differently.

Classification Objectives:

- **Classify Regions by Socioeconomic Vulnerability**: Develop models to classify regions into different categories of socioeconomic vulnerability (like highly vulnerable, moderately vulnerable, low vulnerability) based on cost-of-living factors like housing, food, healthcare, and income levels.
- **Classify Regions for Targeted Policy Making:** Create classification models that help policymakers identify regions that need targeted interventions (like subsidies or tax benefits) based on their categorized cost of living and economic conditions.

Prediction Objectives:

- **Predict Future Cost of Living Increases**: Use regression models to forecast future increases in living costs in various regions, allowing local governments and businesses to plan for inflation and cost-of-living adjustments.
- **Predict Impact of Policy Changes on Cost of Living**: Develop predictive models to simulate how potential changes in local, state, or federal policies (such as changes in taxes or subsidies) could impact the overall cost of living in different regions.

4.The Significance of the Project:

4.1 Application Description:

The application will collect and analyze cost of living data across U.S. regions using clustering, classification, and regression algorithms. It will feature an interactive dashboard to visualize cost-of-living patterns, predict living expenses, and categorize regions. Users can explore, compare cities, and get insights into socio-economic factors affecting living costs.

Significance:

Economic Planning: Governments and policymakers can use insights from this analysis to make informed decisions on resource allocation, subsidies, and development strategies for different regions.
Relocation Decisions: Individuals and families can better understand the financial implications of moving to different regions.
Business Strategy: Companies can use this information to plan office locations, product pricing strategies, and marketing campaigns.

Data Mining Questions:

- What are the natural groupings of U.S. regions based on similarities in their cost of living characteristics?
- What socio-economic factors most strongly influence the cost of living in different regions?
- Can we predict the overall cost of living for a region based on features like income, housing, and healthcare costs?
- How can regions be classified into "low," "moderate," or "high" cost categories based on specific living expense features?
- What patterns or associations exist between income levels and other living expenses, such as housing or healthcare, across U.S. regions?

4.2 Justification for Data Mining Questions with Specific Algorithm

1. **What are the natural groupings of U.S. regions based on similarities in their cost-of-living characteristics?[Tan 2019]**
   Chosen Algorithm: K-Means Clustering
   Explanation:
   K-Means Clustering is selected as the most appropriate algorithm to identify natural groupings of U.S. regions based on cost-of-living characteristics. This algorithm partitions the dataset into K distinct clusters, where each cluster represents regions with similar cost features (e.g., housing cost, food cost, healthcare cost). Given that our dataset contains multiple attributes (like housing, transportation, and childcare costs), K-Means can effectively segment regions into groups such as "low-cost," "moderate-cost," and "high-cost" regions. The simplicity and scalability of K-Means make it ideal for our dataset, which contains over 31,000 records, allowing for efficient computation and easy interpretation of results. Additionally, K-Means is well-suited for our objective of understanding regional similarities and differences in cost-of-living characteristics.

2. **What socio-economic factors most strongly influence the cost of living in different regions?[Halima 2022]**
   Chosen Algorithm: Random Forest Regression
   Explanation:
   Random Forest Regression is selected to identify which socio-economic factors most strongly influence the cost of living across different regions. Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to improve predictive accuracy and control overfitting. This algorithm is well-suited to handle the non-linear relationships and interactions between the features (e.g., income, housing cost, healthcare cost, etc.) in our dataset. It also provides feature importance scores, which will help us rank the influence of each socio-economic factor on the cost of living. This interpretability is valuable for policymakers and stakeholders who are interested in understanding the key drivers of cost variations among different U.S. regions.

3. **Can we predict the overall cost of living for a region based on features like income, housing, and healthcare costs?[Friensen 2018]**
   Chosen Algorithm: Gradient Boosting Regression (GBR)
   Explanation:
   Gradient Boosting Regression (GBR) is chosen for predicting the overall cost of living based on features such as income, housing, and healthcare costs. GBR is a powerful machine learning technique that builds a series of weak learners (usually decision trees) sequentially, with each new tree correcting errors made by the previous ones. This approach is effective for handling complex, non-linear relationships in the data. Given our objective to accurately forecast the "total_cost" (the dependent variable in our dataset), GBR's ability to reduce bias and variance, as well as handle outliers and interactions among variables, makes it the ideal choice. Additionally, GBR often outperforms other regression methods in terms of predictive accuracy, which aligns with our goal of developing a robust model for cost prediction.

4. **How can regions be classified into "low," "moderate," or "high" cost categories based on specific living expense features?[Halima 2022]**
   Chosen Algorithm: Random Forest Classifier
   Explanation:
   The Random Forest Classifier is chosen for classifying U.S. regions into "low," "moderate," or "high" cost categories. This algorithm is an ensemble method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. It is highly effective in handling large datasets and can deal with both categorical and continuous variables, which is beneficial given the mixed types of data in our dataset. The Random Forest Classifier is known for its high accuracy and robustness against overfitting, making it ideal for our classification task where we aim to create reliable and generalizable models for categorizing regions based on various living expense features.

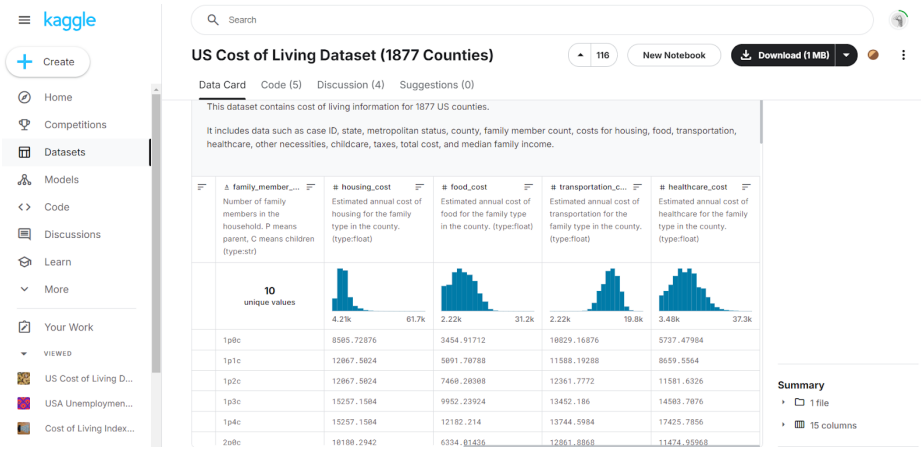
5. **What patterns or associations exist between income levels and other living expenses, such as housing or healthcare, across U.S. regions?**
   Chosen Algorithm: Apriori Algorithm[Domingos 2012]

Explanation:
The Apriori Algorithm is chosen to discover patterns and associations between income levels and other living expenses (e.g., housing, healthcare) across U.S. regions. Apriori is a widely used algorithm for mining frequent itemsets and generating association rules in a transactional dataset. In our project, Apriori will help uncover interesting rules such as "regions with high income levels tend to have high housing costs," thereby revealing insights into how different living expenses are related to income across regions. The algorithm's ability to provide support, confidence, and lift metrics for each rule helps in understanding the strength and significance of these associations, which can be valuable

4.3 Description of Real Datasets and Their Sources:

- **Dataset Name:** "cost_of_living_us.csv"
- **Number of Records:** 31,430
- **Number of Attributes:** 15
- **Attributes Description:**

1. "case_id" (int): Unique identifier for each record.
   Size: 4 bytes
2. "state" (string): US state abbreviation.
   Size: 2 bytes
3. "isMetro" (boolean): Indicates whether the area is metropolitan.
   Size: 1 byte
4. "areaname" (string): Name of the area.
   Size: 50-100 bytes
5. "county" (string): Name of the county.
   Size: 50-100 bytes
6. "family_member_count" (string): Number of family members.
   Size: 1-2 bytes
7. "housing_cost" (float): Annual housing cost in USD.
   Size: 4-8 bytes
8. "food_cost" (float): Annual food cost in USD.
   Size: 4-8 bytes
9. "transportation_cost" (float): Annual transportation cost in USD.
   Size: 4-8 bytes
10. "healthcare_cost" (float): Annual healthcare cost in USD.
    Size: 4-8 bytes
11. "childcare_cost" (float): Annual childcare cost in USD.
    Size: 4-8 bytes
12. "other_cost" (float): Other costs in USD.
    Size: 4-8 bytes
13. "taxes" (float): Annual taxes in USD.
    Size: 4-8 bytes
14. "total_cost" (float): Total annual cost of living in USD.
    Size: 4-8 bytes
15. "region" (string): Geographical region in the US (e.g., Northeast, Midwest).
    Size: 10-20 bytes



Classification Task: The "region" attribute is used for categorization.
Regression Task: The "total_cost" attribute is the target variable for prediction.
Dataset Size: Approximately 4.94 MB.
URL: https://www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties

5. Implementation/Research Methodology and Time Table:

Literature Review and Algorithm Selection:

Objective: Conduct a comprehensive literature review to understand the current state-of-the-art in clustering, classification, regression, and association analysis algorithms. Select appropriate algorithms to implement and justify their choice based on the review.

Deliverables: A document summarizing the literature review. A list of selected algorithms with justifications.

Time Frame: August 30th - September 7th

Assigned To: All members

Data Preprocessing and Exploration

Objective: Clean and preprocess the "cost_of_living_us.csv" dataset to handle missing values, normalize numerical features, encode categorical variables, and perform exploratory data analysis (EDA) to understand the data distribution and relationships.

Deliverables: Cleaned and preprocessed dataset. EDA report with visualizations and descriptive statistics.

Time Frame: September 8th - September 25th

Assigned To: Hima Deepika Mannam

Algorithm Implementation (Clustering, Classification, Regression, and Association Analysis)

Objective: Implement selected data mining algorithms (like K-Means for clustering, Gradient Boost Regression for regression, Random Classifier for classification, Apriori for association analysis) from scratch using Python and libraries like NumPy, SciPy, and Pandas.

Deliverables: Python scripts for each implemented algorithm. Documentation of each algorithm's implementation details.

Time Frame: September 26th - October 10th

Assigned To: Sasank Sribhashyam and Venkat Tarun Adda

Model Training and Evaluation:

Objective: Train the implemented algorithms using the preprocessed dataset and evaluate their performance based on appropriate metrics (like Silhouette Score for clustering, RMSE for regression, Accuracy for classification).

Deliverables: Trained models and evaluation reports.Comparative analysis of algorithm performances.

Time Frame: October 11th - October 25th

Assigned To: Venkat Tarun Adda

Selection of Top-Performing Algorithms and Optimization:

Objective: Choose algorithms that show the best performance for clustering, classification, and regression tasks. Optimize the selected algorithms for better performance.

Deliverables: Optimized models for the selected algorithms. Documentation on the optimization process.

Time Frame: October 26th - November 5th

Assigned To: Sasank Sribhashyam

Development of an Interactive Dashboard for Visualization:

Objective: Develop an interactive dashboard using libraries such as Dash, Plotly, or Tableau to visualize the cost of living patterns, classification results, and regression predictions across U.S. regions.

Deliverables: A fully functional interactive dashboard.

Time Frame: November 6th - November 20th

Assigned To: Hima Deepika Mannam

Final Report Writing and Presentation Preparation:

Objective: Compile all findings, methodology, results, and conclusions into a comprehensive final report and prepare a presentation for project delivery.

Deliverables: Final project report and presentation slides.

Time Frame: November 21st - December 3rd

Assigned To: All members

6. References

1. "Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar : Introduction to Data Mining, 2ndEdition, Pearson Education, Inc., (2019)"

2. "Georgius Andrian Halima, Patrice Agustina, Elbert Adiwijayantoa, Margaretha Ohyver, Estimation of cost of living in a particular city using multiple regression analysis and correction of residual assumptions through appropriate methods, 7th International Conference on Computer Science and Computational Intelligence 2022 (2022)"

3. "P.Ketha Vardhan Reddy, K.Hari Haran, P.Sahith Chowdary, Dr.G.Victo Sudha George. GLOBAL COST OF LIVING USING DATA SCIENCE, 2022 IJCSPUB | Volume 12, Issue 1 March 2022 | ISSN: 2250-1770 (2022)"

4. "John Friesen 1 ID, Lea Rausch 1 1 ID , Peter F. Pelz 1,* and Johannes Fürnkranz, Determining Factors for Slum Growth with Predictive Data Mining Methods, Urban Sci. (2018), 2, 81; doi:10.3390/urbansci203008"

5. "Prediction of the Cost of Living Index Using Machine Learning Techniques, Domingos, P. (2012). [www.medium.com] Date Accessed: 09/12/2024."

6. "A few useful things to know about machine learning," Communications of the ACM, vol. 55, pp. 78-87."

7. "Dataset: https://www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties"