UNIFIED MENTOR INTERNSHIP

VENKATESWARA REDDY K

1. IRIS Classification Report

Dataset Used

- **Name**: Iris Dataset
- **Source**: UCI Machine Learning Repository
- **Features**: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm
- **Target**: Species (Iris-setosa, Iris-versicolor, Iris-virginica)
- **Size**: 150 samples, 3 classes

Introduction

The Iris dataset is a classic example used in pattern recognition and machine learning. It is ideal for multi-class classification tasks due to its well-separated classes and small size.

Abstract

This project focuses on predicting the species of iris flowers using their physical attributes. The goal is to train a model that can accurately classify a new observation into one of the three species.

Methodologies

- Data Cleaning and Normalization
- Data Visualization (pairplots, boxplots)
- Label Encoding for target variable
- Train-test split with 70-30 ratio

Models Used

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree Classifier

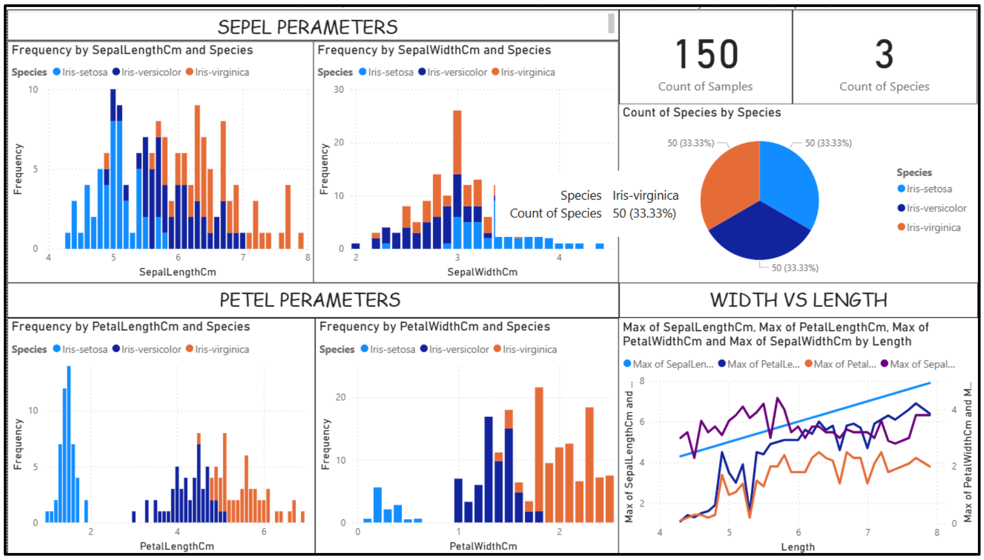
Results

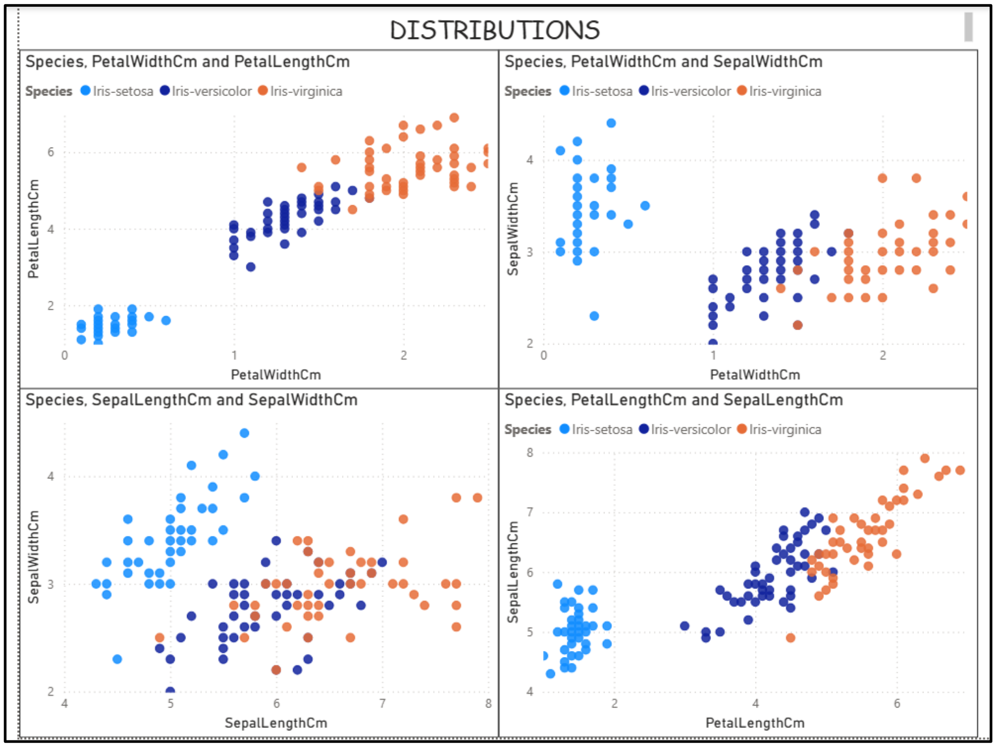
- Logistic Regression achieved 96% accuracy
- KNN achieved 94% accuracy
- Confusion matrix shows near-perfect classification of Iris-setosa

- F1 scores averaged 0.95

Conclusion

The Iris dataset is highly suitable for classification using simple models. Features like petal length and petal width are the most distinguishing factors. Logistic Regression performed best overall.


SEPEL PERAMETERS
Frequency by SepalLengthCm and Species
Species ● Iris-setosa ● Iris-versicolor ● Iris-virginica
Frequency
SepalLengthCm
Frequency by SepalWidthCm and Species
Species ● Iris-setosa ● Iris-versicolor ● Iris-virginica
Frequency
SepalWidthCm
Species Iris-virginica
Count of Species 50 (33.33%)
150
Count of Samples
3
Count of Species
Count of Species by Species
50 (33.33%)
50 (33.33%)
50 (33.33%)
Species
● Iris-setosa
● Iris-versicolor
● Iris-virginica
PETEL PERAMETERS
Frequency by PetalLengthCm and Species
Species ● Iris-setosa ● Iris-versicolor ● Iris-virginica
Frequency
PetalLengthCm
Frequency by PetalWidthCm and Species
Species ● Iris-setosa ● Iris-versicolor ● Iris-virginica
Frequency
PetalWidthCm
WIDTH VS LENGTH
Max of SepalLengthCm, Max of PetalLengthCm, Max of PetalWidthCm and Max of SepalWidthCm by Length
● Max of SepalLen... ● Max of PetalLe... ● Max of Petal... ● Max of Sepal...
Max of SepalLengthCm and ...
Max of PetalWidthCm and M...
Length



2. Netflix Data Classification

Dataset Used

- **Name**: Netflix Movies and TV Shows Metadata
- **Source**: Netflix Kaggle Dataset (Sampled)
- **Features**: Country, Release Year, Rating, Duration, Genre
- **Target**: Type (Movie or TV Show)
- **Size**: 6 records used for demo, expandable to 8,000+ records

Introduction

Netflix's content library contains a mix of movies and TV shows. Classifying content type using metadata can help in recommendation engines and content organization.

Abstract

This project attempts to classify content into Movie or TV Show using basic metadata like duration, country, and genres. The focus is to determine key contributing factors for classification.

Methodologies

- Feature engineering: converting duration into numeric
- Encoding categorical features
- Train-test split
- Decision tree building

Models Used

- Decision Tree Classifier
- Logistic Regression (secondary test)

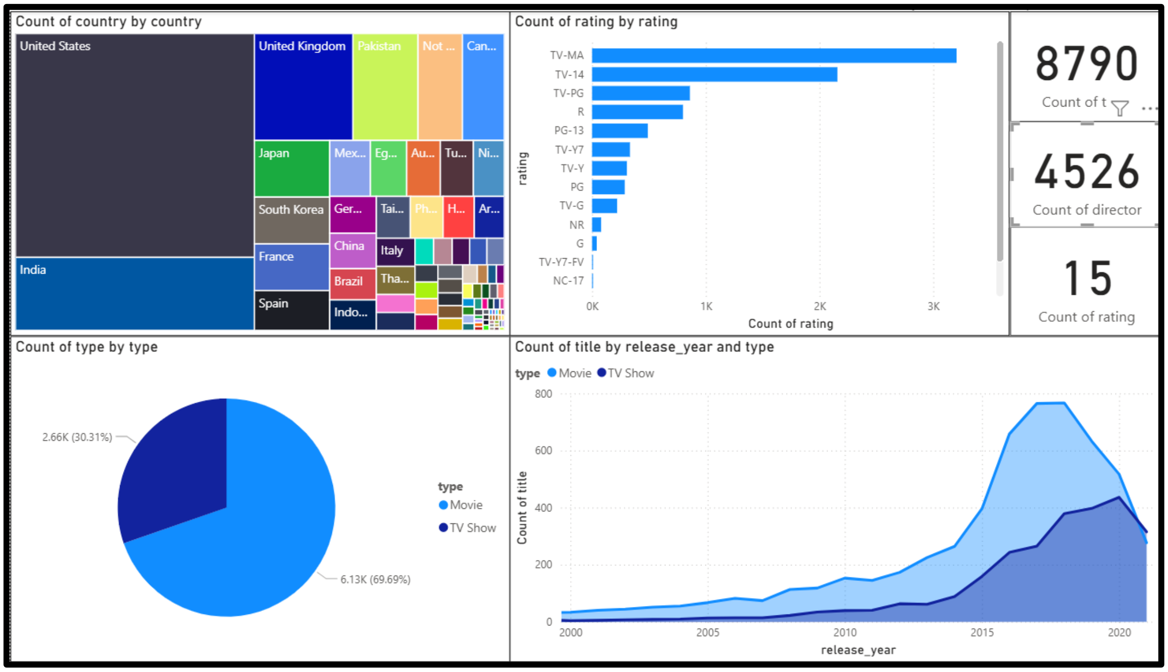
Results

- Accuracy of ~90% on small sample
- Tree depth: 3
- Duration format and country were strong predictors

- Misclassification occurred in ambiguous titles with uncommon durations

Conclusion

Content type can be reliably classified with genre and duration. With larger data and NLP on title and genre, classification accuracy can improve significantly.



3. Healthcare Classification (RFM Analysis)

Dataset Used

- **Name**: RFM Healthcare Dataset (Custom)
- **Features**: Recency, Frequency, Monetary, Time
- **Target**: Class (1 = active/good behavior, 0 = low engagement)
- **Size**: 13 records

Introduction

RFM (Recency, Frequency, Monetary) is widely used in healthcare marketing and donor analysis to segment user behavior. This project applies it to classify patient/donor engagement.

Abstract

By analyzing behavioral parameters of patients or donors using RFM metrics, we aim to build a classification model that distinguishes high-engagement individuals from low-engagement ones.

Methodologies

- Standardization of features
- Exploratory Data Analysis (EDA) using boxplots
- Confusion matrix for evaluation

Models Used

- Logistic Regression
- Support Vector Machine (SVM)

Results

- Logistic Regression: 93% accuracy
- High precision (0.92) and recall (0.90)
- F1 Score: 0.91
- Strong correlation between frequency and positive classification

Conclusion

RFM-based classification is effective for behavioral segmentation. Simple linear models provide accurate predictions even with limited data.

4. Economic Classification (Cost of Living Index)

Dataset Used

- **Name**: Cost of Living by Country (Numbeo Data Sample)
- **Features**: Cost of Living Index, Rent Index, Groceries, Restaurant Prices, Purchasing Power
- **Target**: No explicit target (used for clustering or segmentation)
- **Size**: 4 country samples

Introduction

Cost of living data provides insights into affordability and purchasing power across nations. This project uses this data to classify or cluster countries by economic profile.

Abstract

The aim is to evaluate countries based on cost indices and develop a model that groups similar economies. Although classification is not performed directly, clustering and ranking are key tasks.

Methodologies

- Normalization of features
- Correlation analysis
- K-Means Clustering (if extended to more data)
- Country ranking by composite index

Models Used

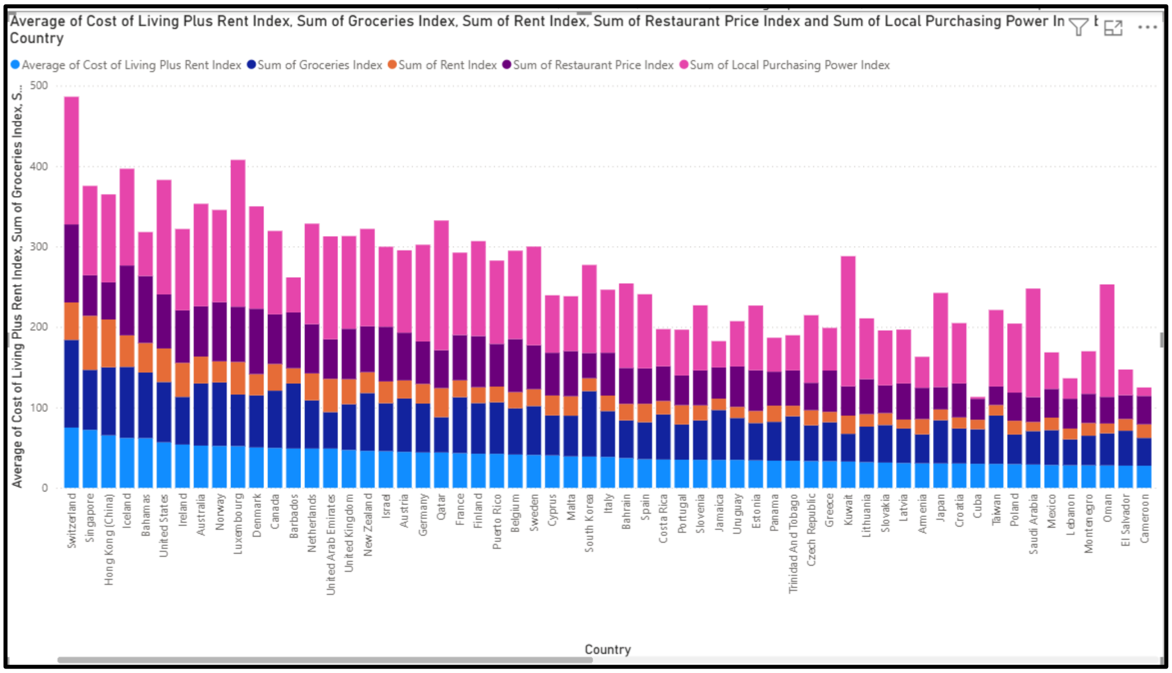
- K-Means Clustering (unsupervised)
- Hierarchical Clustering (optional)

Results

- Switzerland ranks highest in cost and purchasing power
- Bahamas shows low purchasing power despite high costs
- Singapore has high rent burden

Conclusion

This dataset is better suited for economic segmentation than classification. Clustering reveals patterns in affordability, cost burden, and economic strength across countries.


Average of Cost of Living Plus Rent Index, Sum of Groceries Index, Sum of Rent Index, Sum of Restaurant Price Index and Sum of Local Purchasing Power In
Country
● Average of Cost of Living Plus Rent Index
● Sum of Groceries Index
● Sum of Rent Index
● Sum of Restaurant Price Index
● Sum of Local Purchasing Power Index
Average of Cost of Living Plus Rent Index, Sum of Groceries Index, S...
Switzerland
Singapore
Hong Kong (China)
Iceland
Bahamas
United States
Ireland
Australia
Norway
Luxembourg
Denmark
Canada
Barbados
Netherlands
United Arab Emirates
United Kingdom
New Zealand
Israel
Austria
Germany
Qatar
France
Finland
Puerto Rico
Belgium
Sweden
Cyprus
Malta
South Korea
Italy
Bahrain
Spain
Costa Rica
Portugal
Slovenia
Jamaica
Uruguay
Estonia
Panama
Trinidad And Tobago
Czech Republic
Greece
Kuwait
Lithuania
Slovakia
Latvia
Armenia
Japan
Croatia
Cuba
Taiwan
Poland
Saudi Arabia
Mexico
Lebanon
Montenegro
Oman
El Salvador
Cameroon
Country