Project Documentation: Predicting Lifestyle Choices

In our project, we worked on a dataset of 475,675 records and 29 features. Occupying approximately 112.5+ MB of memory. It includes 6 categorical and 23 numerical features.
The dataset includes a variety of features such as:
Demographic: Age, Gender, and Location.
Financial: financial wellness indices and environmental awareness ratings.
Personal routine: Average Weekly Exercise Hours, Average Daily Screen Time.
Our target is "Lifestyle Choice" which is one of the categorical variables in our dataset that raises the potential of our prediction model being a classification task.

Project Design

Objectives

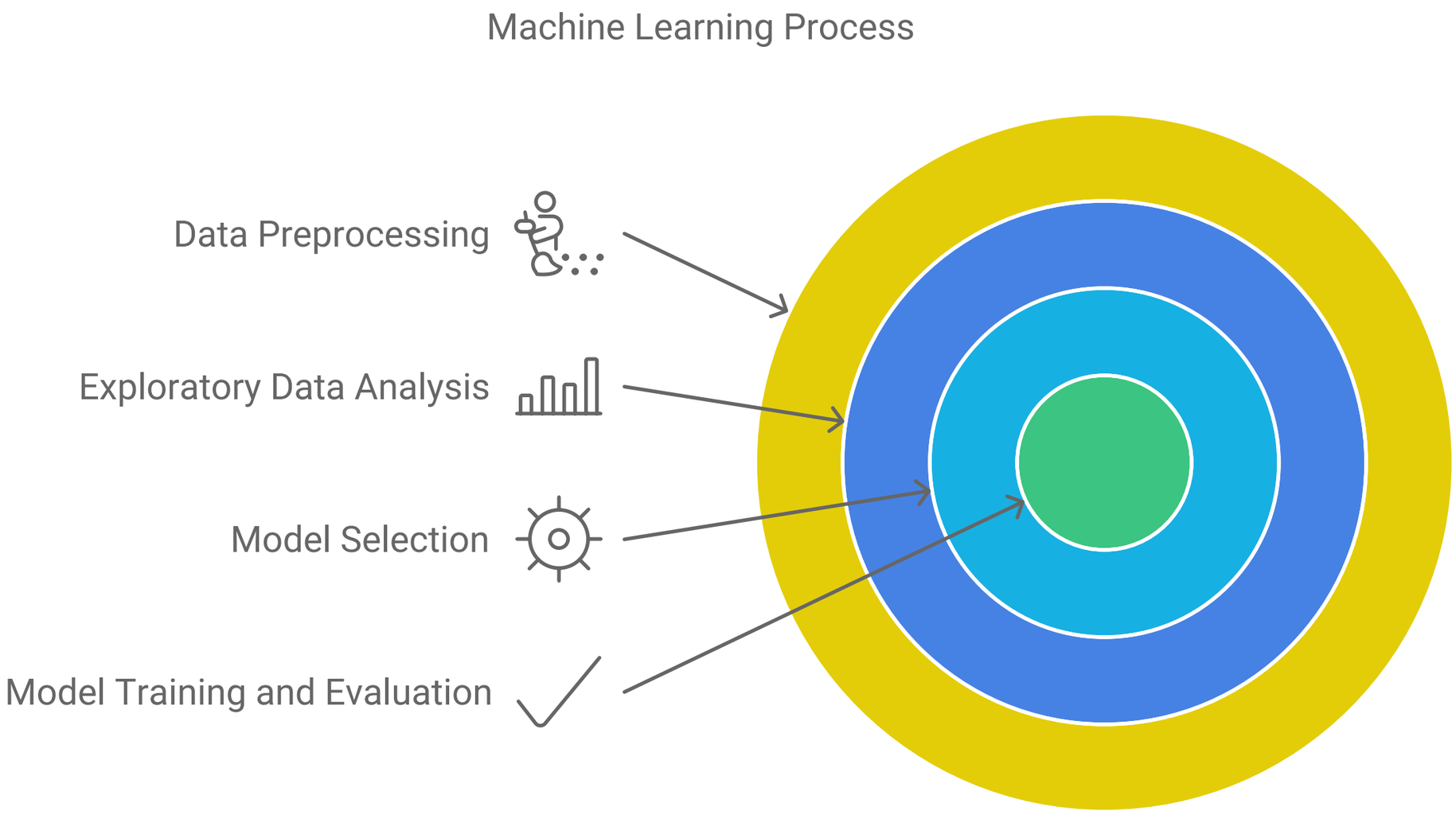
- To predict lifestyle choices based on various features.
- To analyze the relationships between different variables and lifestyle choices.

Features

- **Input Features**: Gender, Age, Annual Vacation Days, Average Monthly Spend on Entertainment, Number of Online Purchases, etc.
- **Output Feature**: Lifestyle Choice (categorical variable).

Methodology

1. **Data Preprocessing**: Clean and preprocess the dataset to handle missing values and convert categorical variables into numerical formats.
2. **Exploratory Data Analysis (EDA)**: Visualize the data to understand distributions and relationships among variables.
3. **Model Selection**: Choose appropriate machine learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forests) for classification.
4. **Model Training and Evaluation**: Split the dataset into training and testing sets, train the model, and evaluate its performance using metrics like accuracy, precision, and recall.



Model

Selected Algorithms

- **Logistic Regression**: For binary classification of lifestyle choices.
- **Random Forest**: To capture complex relationships and interactions between features.
- **Support Vector Machine (SVM)**: For high-dimensional data classification.

Model Training

- **Training Data**: 80% of the dataset.
- **Testing Data**: 20% of the dataset.

Visualizations

EDA Visualizations

- **Distribution of Age**: Histogram showing age distribution among individuals.
- **Spending Habits**: Box plots comparing average monthly spending across different lifestyle choices.
- **Correlation Matrix**: Heatmap illustrating correlations between features.

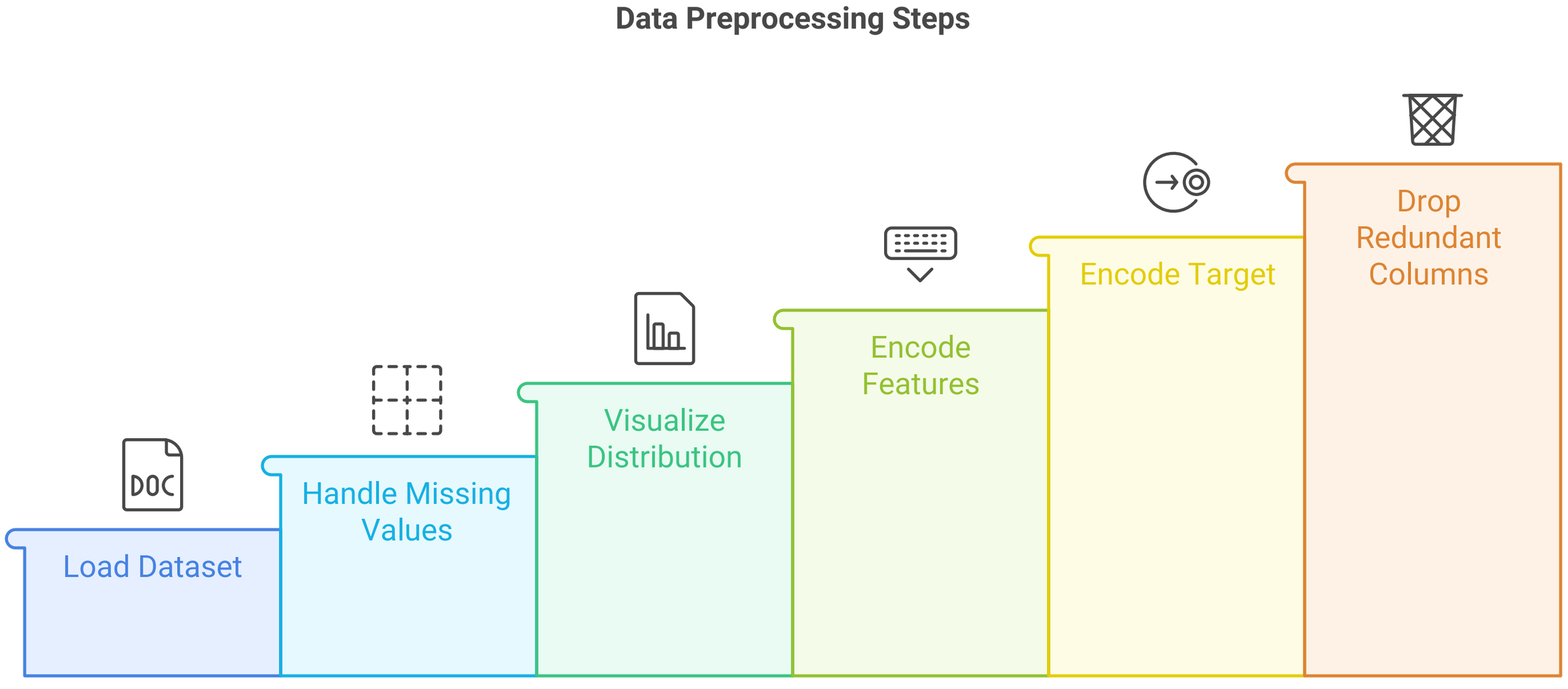
Model Performance

- Accuracy score and classification model

Tutorials

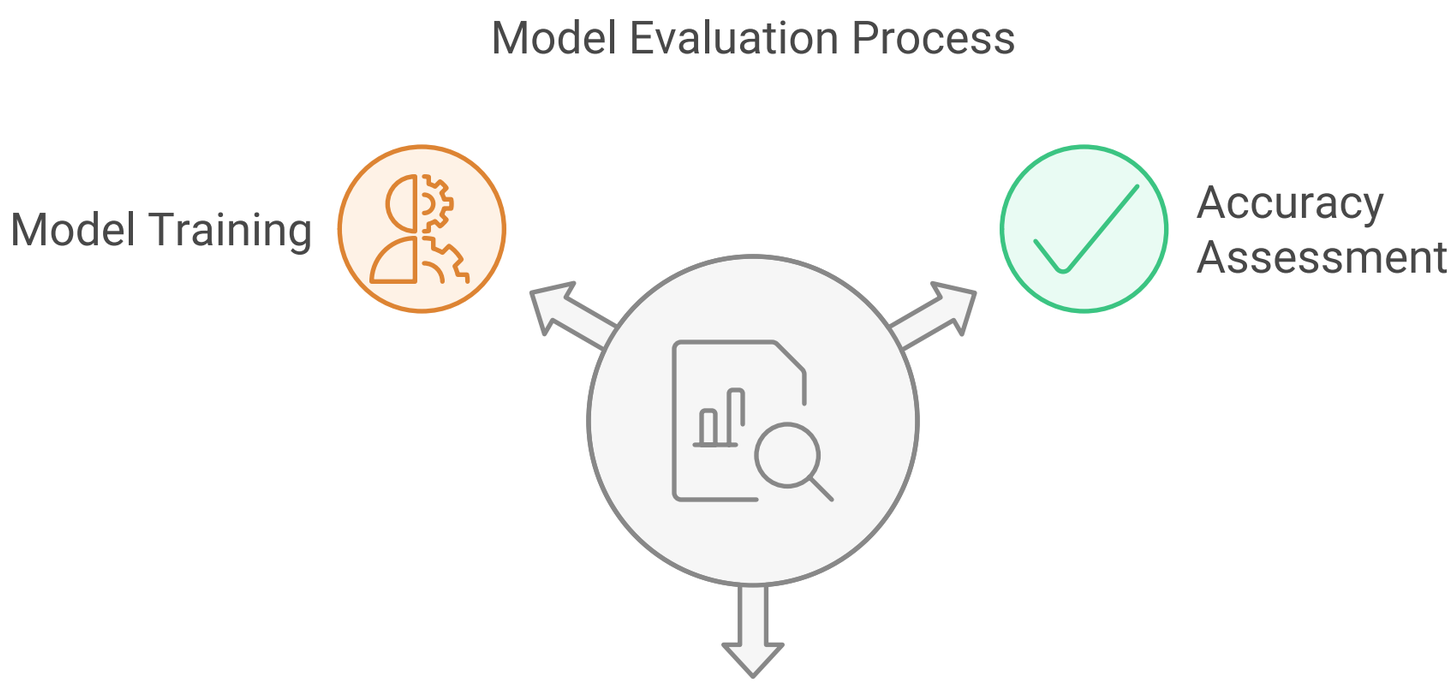
Data Preprocessing Tutorial

1. Load the dataset using Pandas.
2. Handle missing values.
3. visualize target variable distribution, to check for variable imbalance.
4. Encode categorical features
5. Encode target
6. Drop high collinearity, high dimensionality, and redundant columns



Model Training Tutorial

1. Split the dataset into training and testing sets.
2. Train the model using selected algorithms.
3. Evaluate the model using the accuracy and confusion matrix.



Report

Findings

- The model achieved an accuracy of 71% on the testing dataset.
- Key features influencing lifestyle choices include 'Tech-Savviness Score', 'Investment Portfolio Value', 'Risk Tolerance in Investments, 'Average Weekly Exercise Hours', 'Environmental Awareness Rating', 'Financial Wellness Index', 'Social Media Influence Score', 'Health Consciousness Rating', 'Average Monthly Spend on Entertainment', 'Stress Management Score', 'Investment Risk Appetite', 'Number of Online Purchases in Last Month', 'Average Daily Screen Time', 'Education Level', 'Lifestyle Balance Score', 'Time Management Skill', 'Work-Life Balance Indicator', 'Eco-Consciousness Metric'

- The model can effectively predict lifestyle choices based on the input features.

Future Work

- Explore additional features for improved predictions.
- perform hyperparameter tuning to improve the accuracy score
- Consider model deployment
- Implement user feedback mechanisms to enhance the model.

Model Enhancement and Deployment Process