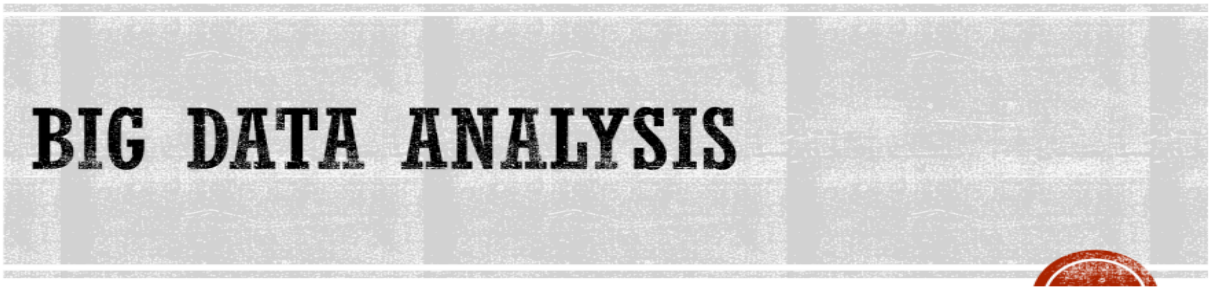


FINAL PROJECT

Student: Mariam Gevorgyan
Professors: Claudiu Brandas,Gabriela Mariutac

Timișoara 2025

Objective

The aim of this project is to apply data analysis and machine learning techniques to predict heart disease risk based on a public health dataset. The dataset, obtained from Kaggle, contains demographic, clinical, and lifestyle information on 70,000 individuals. The project involved data cleaning, feature engineering, exploratory analysis, clustering, and the implementation of classification models.

Database

The dataset used for this project was obtained from Kaggle:
🔗 Heart Disease Risk Prediction Dataset

This dataset contains synthetic health data for 70,000 individuals, with the goal of predicting the risk of heart disease. It consists of binary features (0 or 1), a continuous feature for age, and a binary target variable (`risk_label`).

The dataset was originally generated using Python libraries like NumPy and pandas. It maintains realistic clinical patterns and ensures a balanced distribution between low-risk and high-risk cases

Features Used

The dataset includes the following variables.

Variable	Type	Description
`chest_pain`	Binary	Presence of chest pain, a typical symptom of heart disease.
`shortness_of_breath`	Binary	Difficulty in breathing, often linked to cardiac conditions.
`fatigue`	Binary	Unusual tiredness without obvious cause.
`palpitations`	Binary	Rapid or irregular heartbeat.
`dizziness`	Binary	Episodes of fainting or lightheadedness.
`swelling`	Binary	Swelling in legs/ankles, often indicating heart failure.

radiating_pain	Binary	Pain radiating to the arm, jaw, neck, or back.
cold_sweats	Binary	Cold sweats or nausea, common in cardiac events.
hypertension	Binary	History of high blood pressure.
cholesterol_high	Binary	Presence of high cholesterol levels.
diabetes	Binary	Whether the individual is diabetic.
smoker	Binary	Smoking history.
obesity	Binary	Whether the individual is obese.
family_history	Binary	Family history of cardiovascular diseases.
age	Numeric	Age of the individual (in years).
risk_label	Binary	Target variable: 0 = low risk, 1 = high risk of heart disease.

Methodology / Data Processing

The dataset was loaded using `pandas` and examined for structure, missing values, and data types:

```
df = pd.read_csv('heart_disease_risk_dataset_earlymed.csv')

expected_columns = ['Chest_Pain', 'Shortness_of_Breath', 'Fatigue', 'Palpitations', 'Dizziness', 'Swelling',
 'Pain_Arms_Jaw_Back', 'Cold_Sweats_Nausea_Nausea', 'High_BP', 'High_Cholesterol', 'Diabetes',
 'Smoking', 'Obesity', 'Sedentary_Lifestyle', 'Family_History', 'Chronic_Stress',
 'Gender', 'Age', 'Heart_Risk']


missing_columns = [col for col in expected_columns if col not in df.columns]
if missing_columns:
    print(f"Warning: Missing columns in dataset: {missing_columns}")
    print(f"Available columns: {list(df.columns)}")
    print("Please check your column names and update the expected_columns list if needed.")

print("Dataset loaded successfully!")
print(f"Dataset shape: {df.shape}")
print(f"Available columns: {list(df.columns)}")

print("Dataset Overview:")
print(f"Dataset Shape: {df.shape}")
print(f"Features: {df.shape[1] - 1}")
print(f"Samples: {df.shape[0]}")
print("\nFirst 5 rows:")
print(df.head())

print("\nDataset Info:")
print(df.info())

print("\nBasic Statistics:")
print(df.describe())

print("\nTarget Variable Distribution:")
print(df['Heart_Risk'].value_counts())
print(f"Risk Distribution: {df['Heart_Risk'].value_counts(normalize=True)}")
```

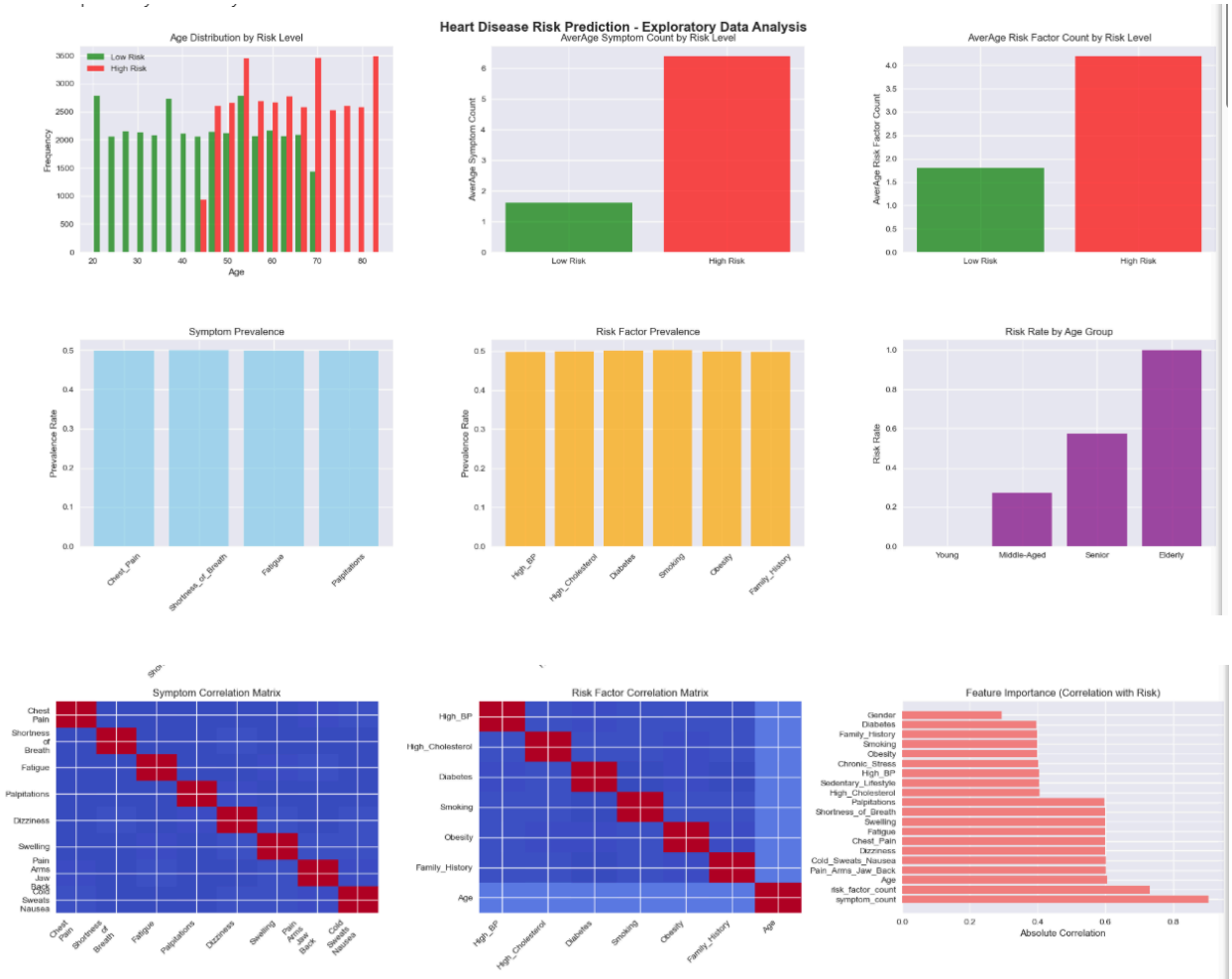
The dataset included 70,000 samples and 19 columns, with no missing values. Most features were binary (0 or 1), and `age` was continuous

```
df['symptom_count'] = (df['Chest_Pain'] + df['Shortness_of_Breath'] + df['Fatigue'] +
                       df['Palpitations'] + df['Dizziness'] + df['Swelling'] +
                       df['Pain_Arms_Jaw_Back'] + df['Cold_Sweats_Nausea'])

df['risk_factor_count'] = (df['High_BP'] + df['High_Cholesterol'] + df['Diabetes'] +
                           df['Smoking'] + df['Obesity'] + df['Family_History'])

df['Age_group'] = pd.cut(df['Age'], bins=[0, 30, 50, 70, 100],
                         labels=['Young', 'Middle-Aged', 'Senior', 'Elderly'])

print("New features created:")
print("- symptom_count: Total number of symptoms")
print("- risk_factor_count: Total number of risk factors")
print("- Age_group: Categorical Age groups")

print("\n3.3 Data Preprocessing for Machine Learning")

feature_cols = [col for col in df.columns if col not in ['Heart_Risk', 'Age_group']]
X = df[feature_cols]
y = df['Heart_Risk']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Before applying machine learning models, the dataset was split into training and test sets (80/20 split). Features were also standardized:

Data visualization and discussion

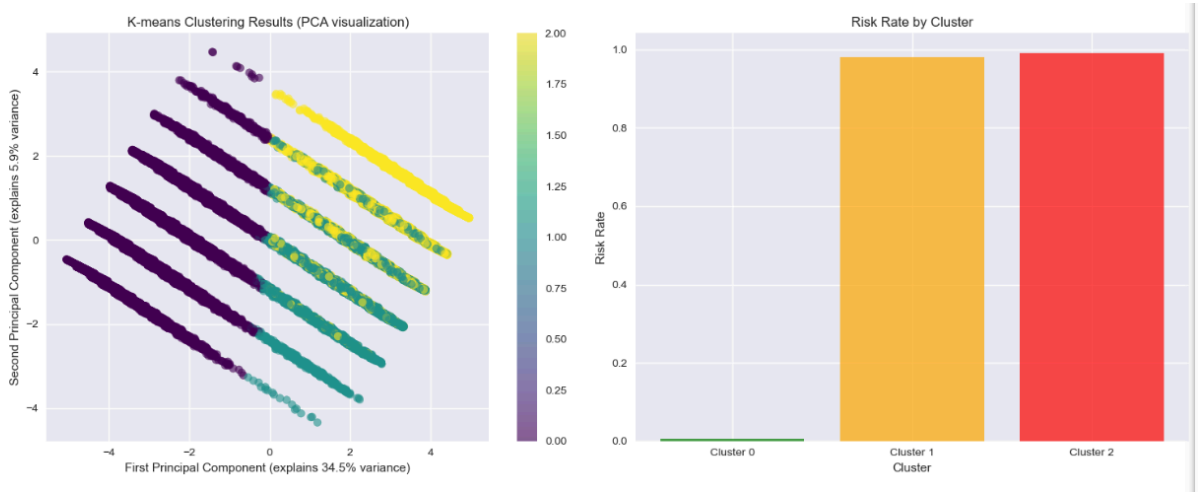
A Pearson **correlation analysis** was performed to understand the relationship between features and the target variable (`risk_label`). The newly engineered features — `symptom_count` (r = 0.905) and `risk_factor_count` (r = 0.731) — showed the highest correlation with heart disease risk, followed by `age` (r = 0.605). These findings suggest that the number of reported symptoms and risk factors are strong indicators of cardiovascular risk.



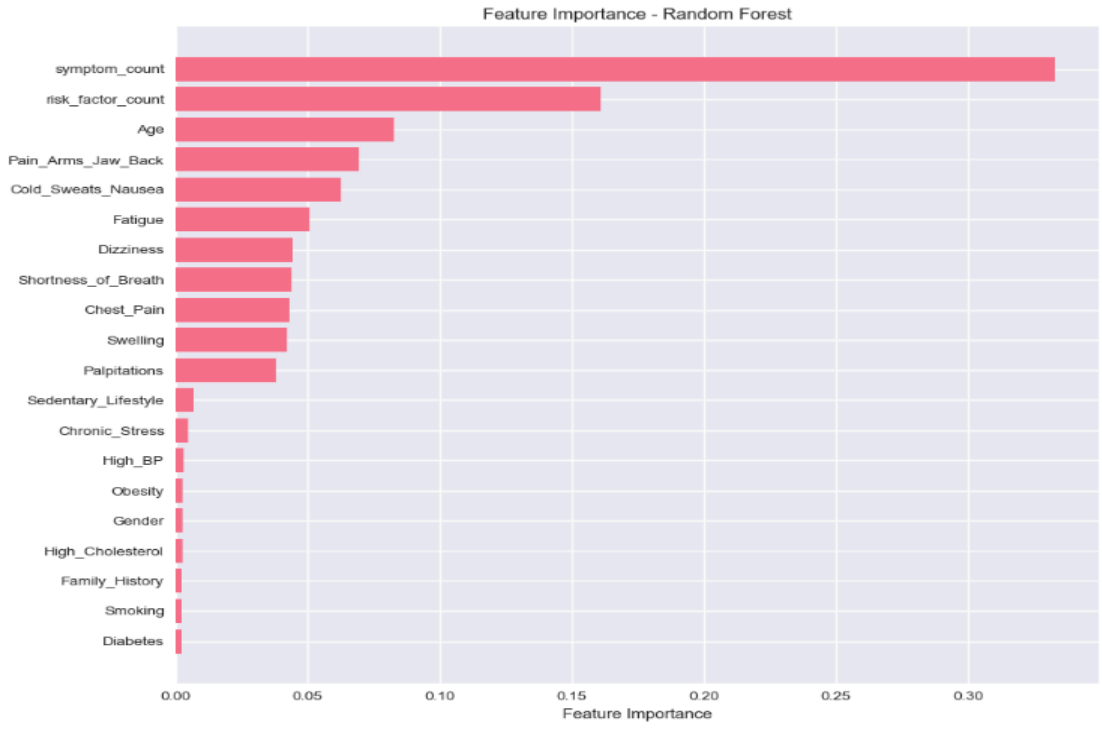
K-Means clustering (k = 3) was applied to identify patterns among patients. The clusters revealed three distinct profiles:

- **Cluster 0**: Younger individuals (~45 years), low symptom and risk factor counts.

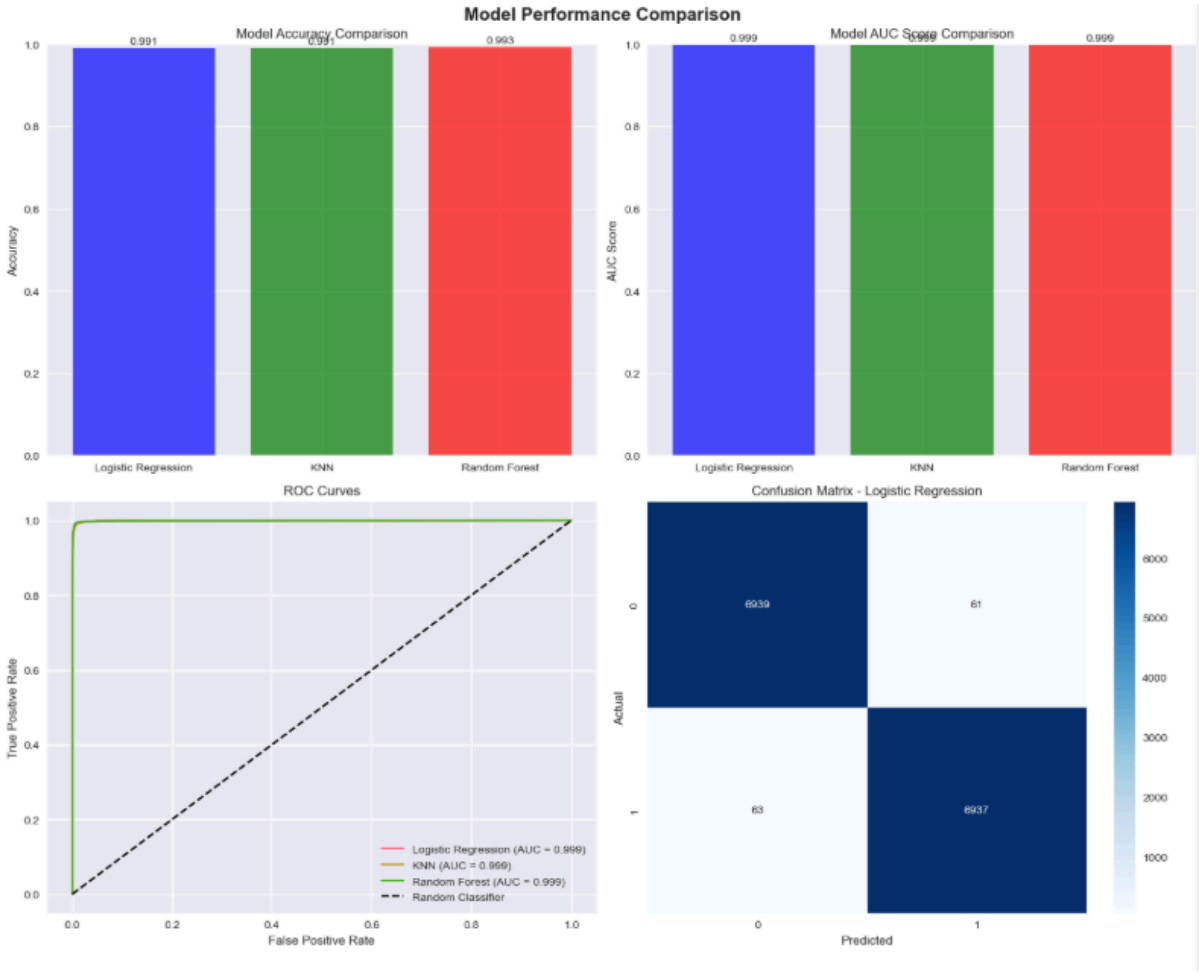
- **Cluster 1 & 2**: Older individuals (~64 years) with high symptom burden; Cluster 2 had even more risk factors on average.



A Random Forest classifier was used to determine which features contributed most to predicting heart disease risk. The top features included:



Three models were evaluated on the test set of 14,000 samples:



Model	Accuracy	AUC-ROC
Logistic Regression	0.991	0.999
K-Nearest Neighbors	0.991	0.999
Random Forest	**0.993**	**0.999**

The combination of clinical symptoms, lifestyle risk factors, and demographic variables provided a strong foundation for accurate heart disease risk prediction. The engineered features (`symptom_count` and `risk_factor_count`) played a crucial role in boosting model performance and simplifying interpretation. The clustering analysis also helped identify patient groups with similar profiles, offering insights for stratified risk management.

Conclusions

In this project, we explored a heart disease risk prediction dataset with the goal of identifying patterns and building accurate machine learning models. After cleaning and examining the data, we engineered new features—such as the number of reported symptoms and risk factors—which proved to be highly correlated with heart disease risk. Through clustering analysis, we discovered distinct groups of patients based on age and health characteristics, providing insights into how risk varies across different population segments.

We then evaluated several classification models, including Logistic Regression, K-Nearest Neighbors, and Random Forest. All models performed exceptionally well, achieving over 99% accuracy and near-perfect AUC scores. Among them, the Random Forest model stood out slightly for its accuracy and ability to highlight important features contributing to risk.

Overall, the project showed how combining data preprocessing, thoughtful feature engineering, and machine learning can lead to highly effective health risk prediction systems. In the future, this approach could be extended to real-world clinical data to support early detection and preventive care strategies for cardiovascular diseases.

```
5. CONCLUSIONS
--------------------------------------------------
1. Dataset Analysis: Successfully analyzed 70000 patient records with 21 features
2. Feature Correlations: symptom_count shows the highest correlation (0.905) with heart disease risk
3. Clustering: Identified 3 distinct patient groups with varying risk profiles
4. Model Performance: Logistic Regression achieved the best performance with 0.999 AUC score
5. Key Risk Factors: Age, symptom count, and specific risk factors are strong predictors
6. Clinical Relevance: Models can assist in early identification of high-risk patients

Key Insights:
- Older patients show higher risk rates across all Age groups
- Combination of symptoms is more predictive than individual symptoms
- Traditional risk factors (hypertension, diabetes, smoking) remain important
- Machine learning models can effectively predict heart disease risk

Recommendations:
- Implement regular screening for high-risk groups identified by clustering
- Focus on patients with multiple symptoms and risk factors
- Consider ensemble methods for improved prediction accuracy
- Validate models on external datasets before clinical deployment
```