Advanced Data Mining for Data-Driven Insights and Predictive Modeling

- Course: MSCS-634 Advanced Big Data and Data Mining
- Project: Comprehensive Data Mining Pipeline

Team Members

- Sushil Khanal
- Sri Hari Gunji
- Sauhard Shakya
- Sahaj Shrestha

Project Overview

Dataset: UCI Heart Disease Dataset (1,035 records, 14 attributes)

Objective: Predict heart disease presence using multiple ML techniques

Deliverables Completed:

- Data Cleaning & EDA
- Regression Modeling
- Classification & Clustering
- Association Rule Mining

Dataset & Data Preprocessing

Dataset Overview:

- 1,035 patient records, 14 medical attributes
- Target: Heart disease presence (0=No, 1=Yes)
- Age range: 29-77 years (mean ~54)

Data Cleaning Steps:

- Missing values: Median imputation
- Duplicates: Removed duplicate records
- Outliers: IQR method applied
- Categorical encoding: One-hot encoding

Exploratory Data Analysis Results

Key Feature Correlations:

- Chest pain type → Strong positive correlation
- Max heart rate → Strong negative correlation
- ST depression → Strong disease indicator

Dataset Characteristics:

- Balanced classes: 50.5% disease, 49.5% normal
- Gender split: 68% male, 32% female
- Age distribution: Normal curve, peak 54-58 years

Regression Modeling Results

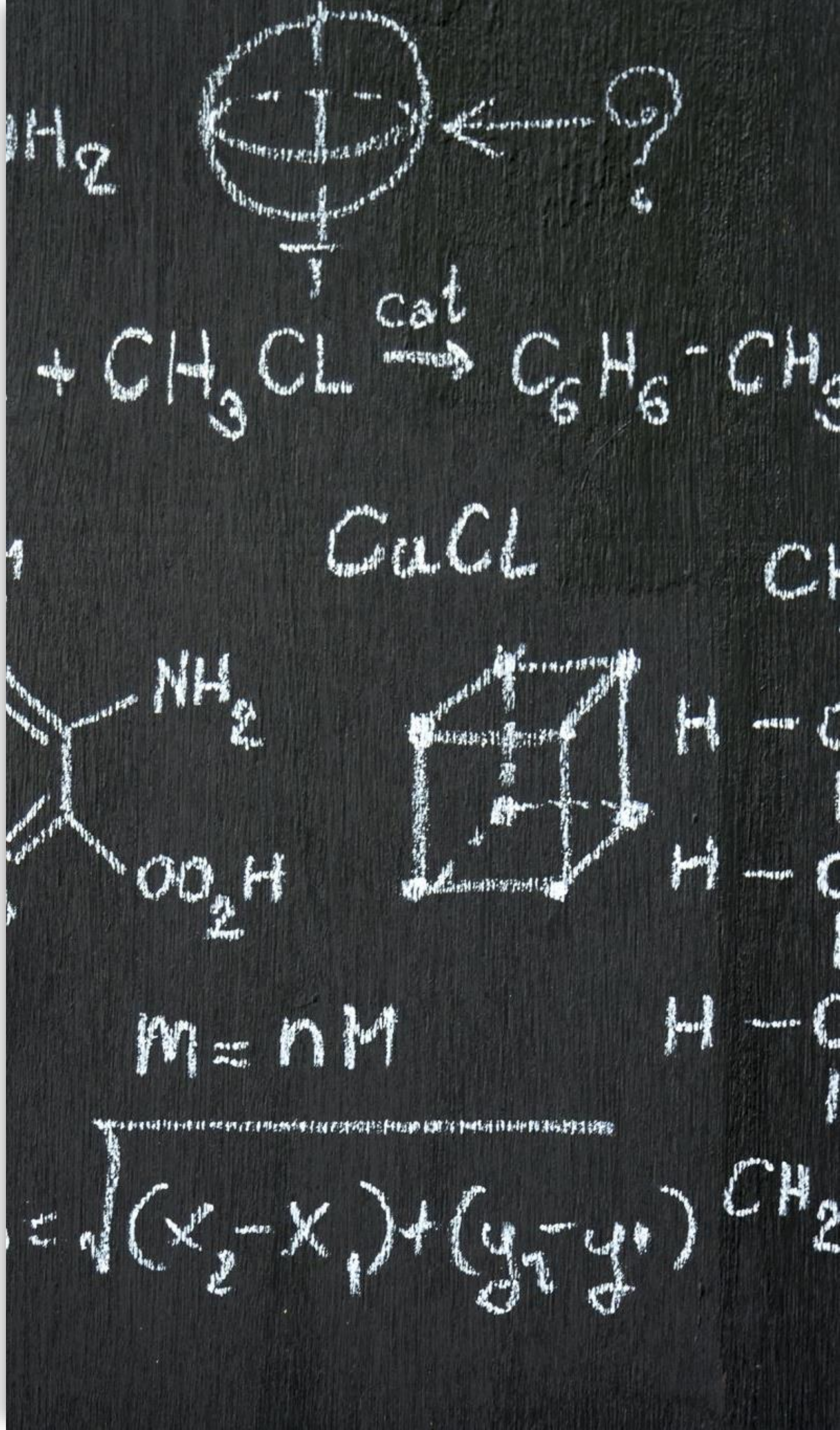
Cholesterol Prediction Models:

- Linear Regression (baseline)
- Ridge Regression (L2 regularization)

Performance Results:

- Linear: $R^2 = 0.12$, RMSE = 65.2
- Ridge: $R^2 = 0.18$, RMSE = 61.8 ✓

Key Finding: Ridge performed better with regularization



Classification Model Performance

Heart Disease Classification

Models Developed:

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

Results:

- Decision Tree: 83.2% accuracy
- KNN (k=5): 86.1% accuracy ✓

Optimal Parameters:

- Decision Tree: max_depth = 4
- KNN: k = 5

Clustering & Pattern Mining

K-Means Clustering (k=2):

• Clusters aligned with disease status (78% accuracy)

• PCA showed clear data separability

• Confirms inherent disease patterns

Association Rule Mining:

• Strongest Rule: chest pain type → heart disease

• Confidence: 78.5%, Lift: 1.53

• Provides clinical decision support

Model Comparison & Best Performers

Summary of Best Models

Winners by Category:

- Classification: KNN (k=5) - 86.1% Accuracy
- Regression: Ridge Regression - Better generalization
- Clustering: K-Means confirmed natural separation
- Pattern Mining: Chest pain type = strongest predictor

Key Success Factors:

- Feature standardization crucial for KNN
- Regularization prevents overfitting

Real-World Applications

Healthcare Applications:

- Risk assessment tool for physicians
- Early screening protocol development
- Patient triage automation

Business Value:

- Reduce diagnostic costs
- Improve patient outcomes
- Support preventive healthcare
- Rule-based diagnostic assistance

Technical Challenges & Solutions

Project Challenges Overcome

Data Quality Issues:

- Challenge: Missing values, outliers, duplicates
- Solution: Median imputation, IQR capping, deduplication

Model Performance:

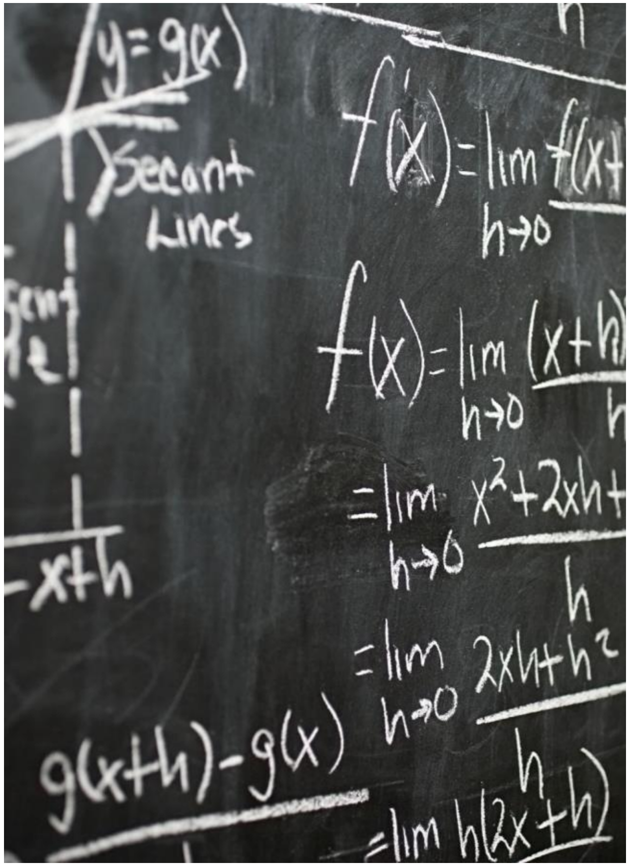
- Challenge: Low R² scores in regression
- Solution: Focused on more successful classification task

Feature Engineering:

- Challenge: Mixed data types (categorical/numerical)
- Solution: One-hot encoding, standardization

Validation:

- Challenge: Overfitting concerns
- Solution: Cross-validation, hyperparameter tuning



Key Insights & Findings

Major Discoveries

Data Science Insights:

- Heart disease prediction achievable with 86.1% accuracy
- Chest pain type is strongest single predictor
- Multiple weak predictors combine effectively

Medical Insights:

- Age, gender, chest pain type form strong prediction trio
- Maximum heart rate decline strongly indicates disease

Technical Insights:

- Regularization prevents overfitting in medical data
- Distance-based algorithms work well with standardized features

Future Recommendations

Next Steps & Improvements

Model Enhancement:

- Ensemble methods (Random Forest, Gradient Boosting)
- Deep learning for complex pattern recognition
- Feature selection optimization

Data Expansion:

- Larger, more diverse patient populations
- Additional biomarkers and lab results
- Longitudinal patient tracking

Deployment Considerations:

- Real-time prediction API development
- Integration with electronic health records
- Regulatory compliance (HIPAA, FDA)

Conclusion

Project Success Summary

Accomplished Objectives:

- Complete data mining pipeline implemented
- Multiple ML techniques successfully applied
- High-accuracy predictive models developed
- Actionable medical insights discovered

Key Metrics Achieved:

- 86.1% classification accuracy
- 78.5% confidence association rules
- Robust cross-validation performance

Business Impact:

- Practical healthcare decision support tool
- Cost-effective screening methodology
- Evidence-based medical rule discovery

Questions & Discussion

Thank You!