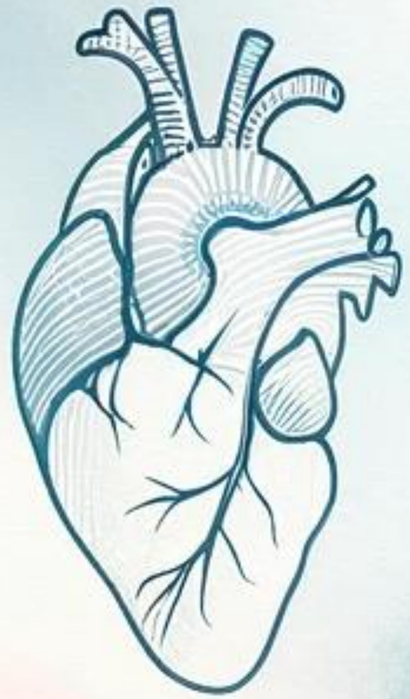
HEART DISEASE RISK PREDICTION



By Zaid Chaudhary, Vishal Vuppula,
Shandilya Motupally, Amruta Godase

INTRODUCTION

- Heart disease: A Top global health threat and leading cause of death.
- Many at-risk individuals remain undiagnosed until advanced stages.
- Early detection and intervention are critical to reducing mortality.

RESEARCH QUESTION

How can we construct a predictive model that reliably identifies individuals at risk of heart disease to facilitate early and effective intervention?

PROBLEM STATEMENT

- **Project Objective: Develop a predictive model for heart disease risk which focuses on different lifestyle factors and behavioral patterns.**
- **Benefit: Early identification and prevention for at-risk individuals.**

IMPORTANCE OF THE PROBLEM

- Global Issue: Cardiovascular diseases are leading causes of death.
- Concerning Statistics: High fatality rates among 35-40 year old's.
- Goal: Achieve a model with high recall to minimize missed diagnoses.
- Impact: Reduce prevalence and fatalities from cardiovascular diseases.

EVALUATION METRICS

- Primary Metric: Accuracy — aiming for 90% or higher detection rate.
- Tool Objective: A prediction tool based on lifestyle and behavior.
- End Goal: Develop a highly accurate system to lower heart disease rates.

DATA SOURCE

- Uses 2015 CDC BRFSS survey with 229,787 participants; behavior and lifestyle data inform prediction models for non-heart disease individuals.

EXPLORATORY DATA ANALYSIS

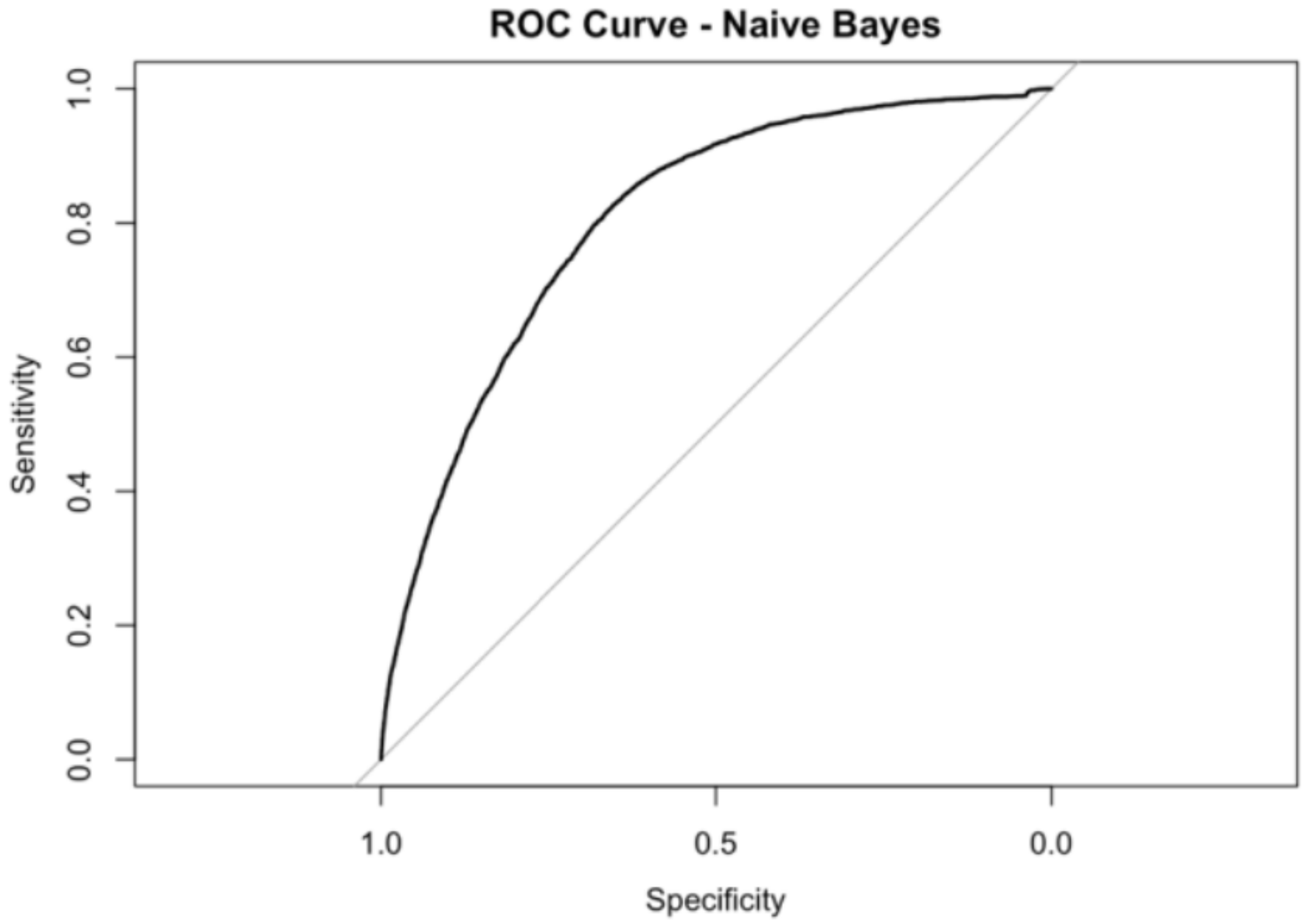
- Data Analysis: Examined correlations using exploratory analysis and visualizations.
- Dataset Integrity: Checked for missing values in a dataset with 253,680 observations and 22 variables.
- Variables: 1 binary target variable “HeartDiseaseorAttack” and 21 binary/ordinal predictors.
- Visualization Tools: Used ggplot2 and corrplot for bar charts and correlation matrices.

NAIVE BAYES - MODEL INTERPRETATION

- **Confusion Matrix and Accuracy Results**

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 39042  2191
         1  6955  2548

               Accuracy : 0.8197
                 95% CI : (0.8164, 0.8231)
    No Information Rate : 0.9066
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2664

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8488
            Specificity : 0.5377
         Pos Pred Value : 0.9469
         Neg Pred Value : 0.2681
             Prevalence : 0.9066
         Detection Rate : 0.7695
   Detection Prevalence : 0.8127
      Balanced Accuracy : 0.6932
```

- **ROC Curve for Naive Bayes**

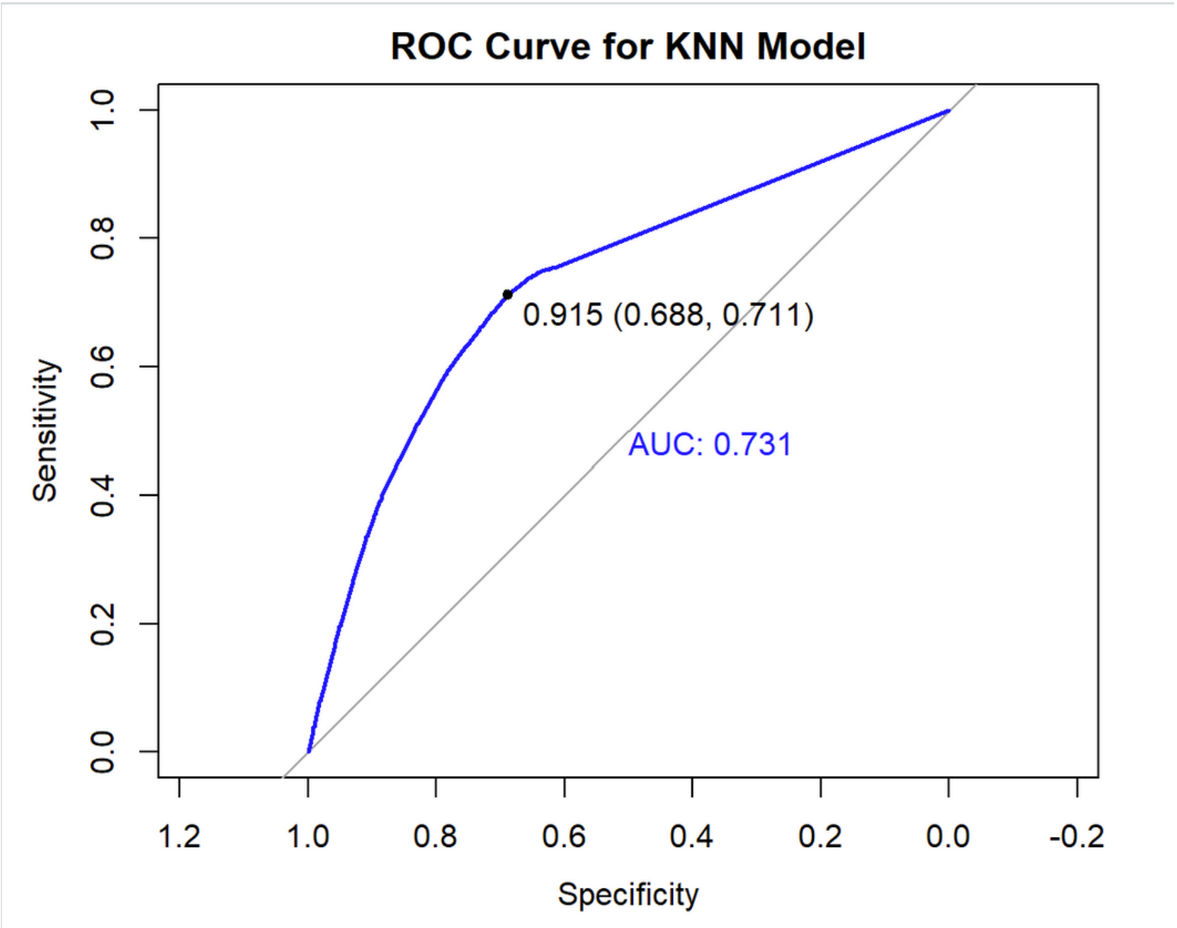


K NEAREST NEIGHBOUR- MODEL INTERPRETATION

- **Confusion Matrix and Accuracy Results**

```
[1] "Accuracy of KNN model: 89.99 %"

Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 67869  6463
         1  1154   618

               Accuracy : 0.8999
                 95% CI : (0.8978, 0.902)
    No Information Rate : 0.907
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1063

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.98328
            Specificity : 0.08728
         Pos Pred Value : 0.91305
         Neg Pred Value : 0.34876
             Prevalence : 0.90696
         Detection Rate : 0.89179
   Detection Prevalence : 0.97672
      Balanced Accuracy : 0.53528

       'Positive' Class : 0
```

- **ROC Curve for KNN Model**

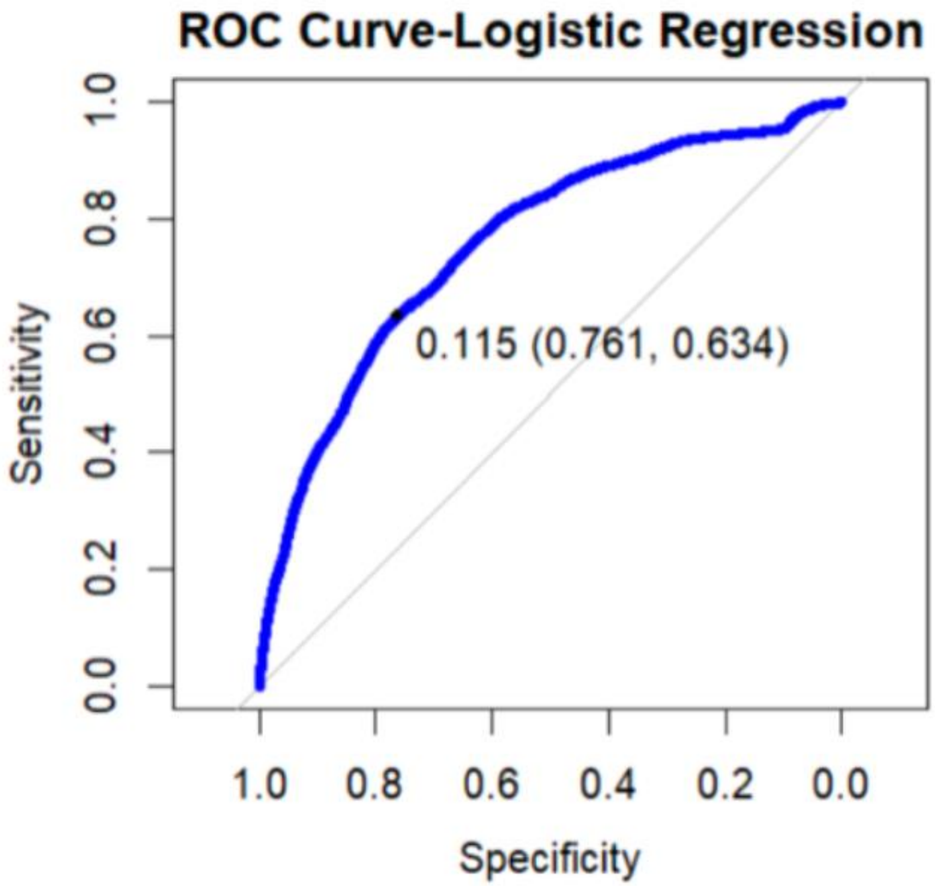


LOGISTIC REGRESSION - MODEL INTERPRETATION

- **Confusion Matrix and Accuracy Results**

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 45500  4163
         1   497   576

               Accuracy : 0.9082
                 95% CI : (0.9056, 0.9107)
    No Information Rate : 0.9066
    P-Value [Acc > NIR] : 0.1154

                  Kappa : 0.1696

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.9892
            Specificity : 0.1215
         Pos Pred Value : 0.9162
         Neg Pred Value : 0.5368
             Prevalence : 0.9066
         Detection Rate : 0.8968
   Detection Prevalence : 0.9789
      Balanced Accuracy : 0.5554

       'Positive' Class : 0
```

- **ROC Curve for KNN Model**



CHALLENGES FACED

Feature Selection

Employed rigorous analysis for logistic regression to avoid overfitting.

Data Volume

Overcame obstacles in managing a dataset with 250,000+ observations.

Computational Demands

Balanced computational efficiency with optimal 'k' in KNN model.

Algorithmic Challenges

Addressed limitations like Naive Bayes' independence assumption affecting specificity.

Model Optimization

Fine-tuned parameters to navigate the trade-off between accuracy and computational load.

RESULTS AND CONCLUSIONS

- **Logistic Regression Success:** Achieved standout performance with strong predictors, such as smoking and stroke history.
- **Impressive Accuracy:** Model verified with 90.53% accuracy.
- **ROC Curve Validation:** AUC of 0.759, indicating high model reliability.
- **KNN Algorithm Proficiency:** Exhibited high accuracy at 89.99%, effectively using demographic and health data.
- **Naive Bayes Insights:** Registered 81.97% accuracy, with high sensitivity but lower specificity, indicating an area for improvement.
- **Model Contributions:** Each offers unique insights, paving the way for advanced predictive health models.

THANK YOU