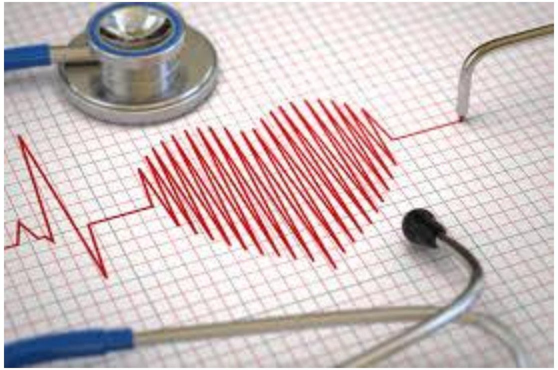


Heart Attack Prediction Report

By Samith Lakka

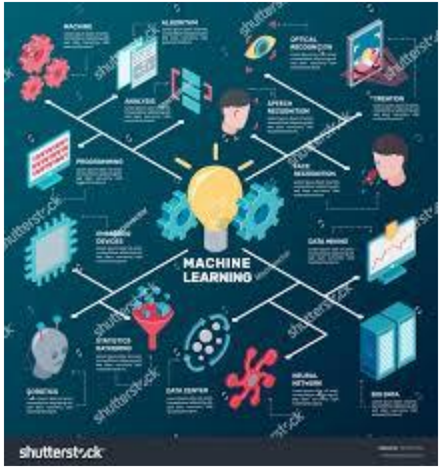
Springboard Capstone 2 Project

Problem

- Can machine learning predict the likelihood of heart attacks?
- Identify key risk factors.
- Assist in early diagnosis and prevention.

Project Goals

- Develop a predictive model for heart attacks.
- Use machine learning to analyze risk factors.
- Evaluate model performance using accuracy, precision, recall, and F1-score.

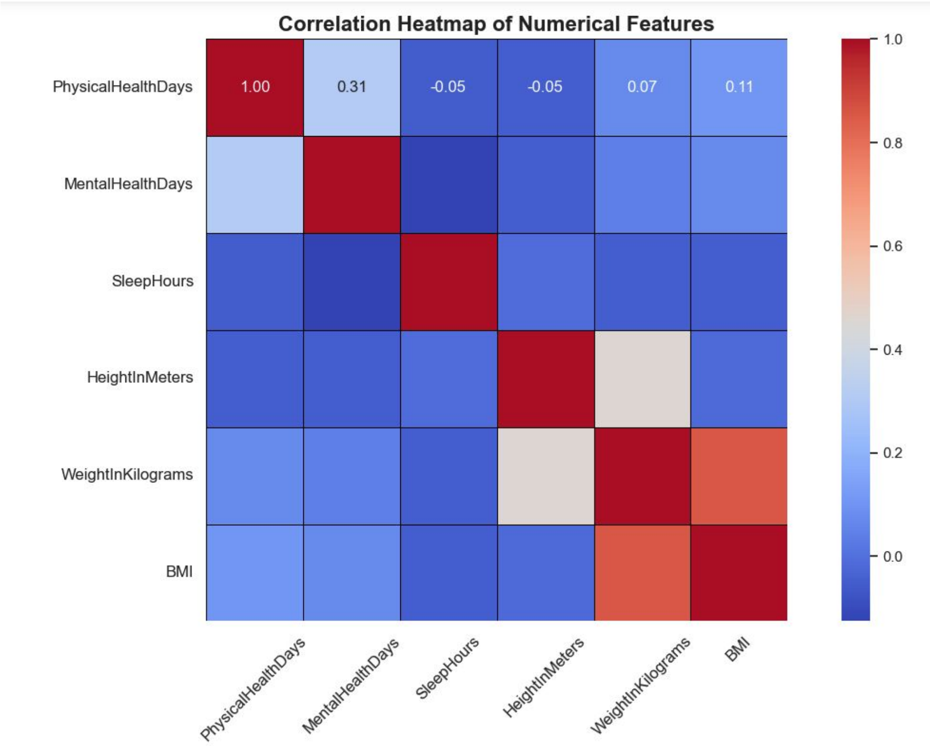


Data Collection and Preprocessing

- **Dataset:** Kaggle's Heart.csv.
- Features include Age, Sex, Blood Pressure, Cholesterol, etc.
- Steps:
  - Handled missing values.
  - Encoded categorical variables.
  - Normalized numerical features.
  - Split into training & testing sets.

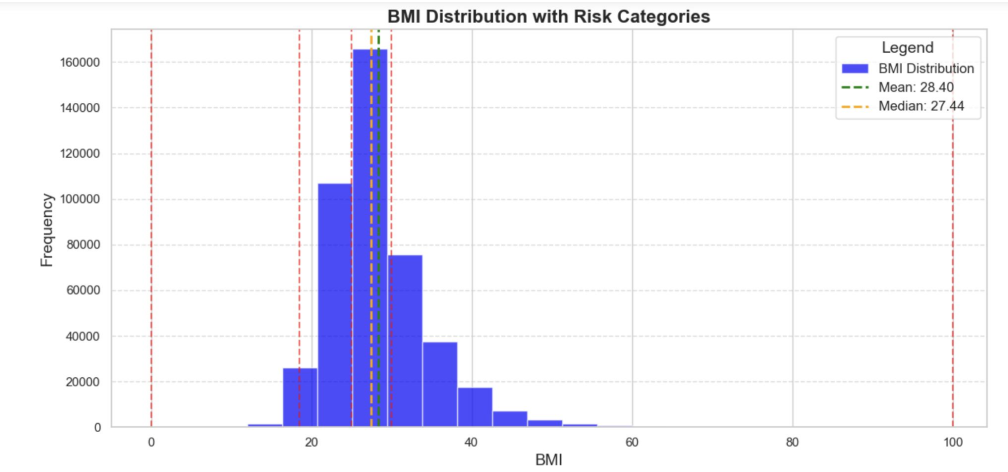
Exploratory Data Analysis

- Initial insights from the dataset.
- Identification of key predictors.
- Understanding data distribution.



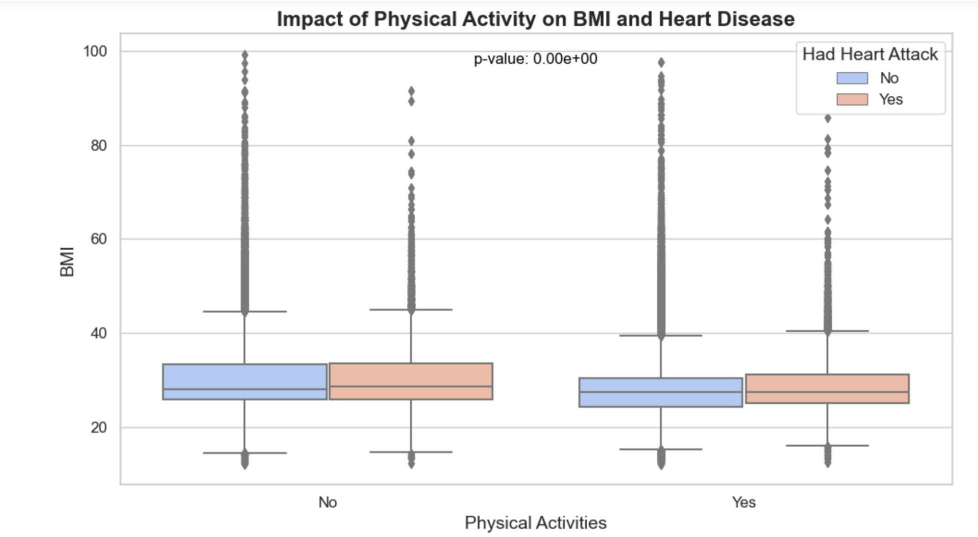
BMI Distribution and Heart Attack Risk

- Majority of individuals fall in Overweight/Obese categories.
- High BMI correlates with heart attack risk.



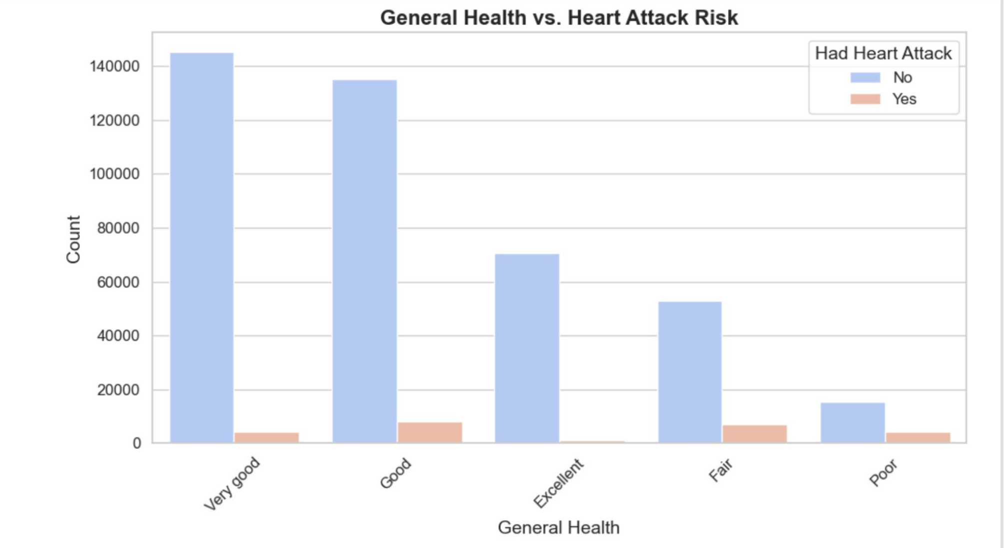
Impact of Physical Activity on Heart Disease

- Individuals with no physical activity tend to have higher BMI and greater heart attack risk.
- Statistically significant differences in BMI between active and inactive groups.



General Health and Heart Attack Risk

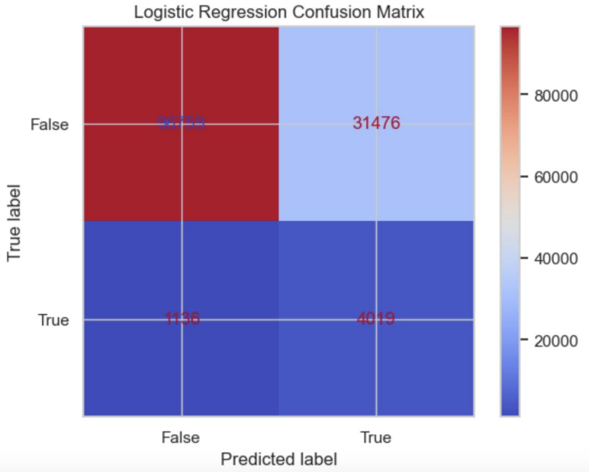
- Poorer self-reported health correlates with higher heart attack occurrences.
- Subjective health ratings provide key insights for prediction.

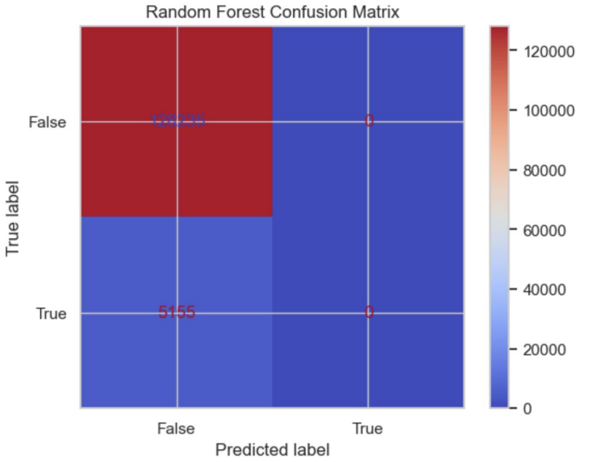


Baseline Modeling

- **Logistic Regression:**
  - Accuracy: 75.55%
  - Struggles with high-risk cases.
- **Random Forest (Default):**
  - Accuracy: 96.13%
  - Completely fails to classify high-risk cases.

	precision	recall	f1-score	support
False	0.988396	0.754544	0.855782	128235.000000
True	0.113227	0.779631	0.197737	5155.000000
accuracy	0.755514	0.755514	0.755514	0.755514
macro avg	0.550811	0.767088	0.526759	133390.000000
weighted avg	0.954574	0.755514	0.830351	133390.000000



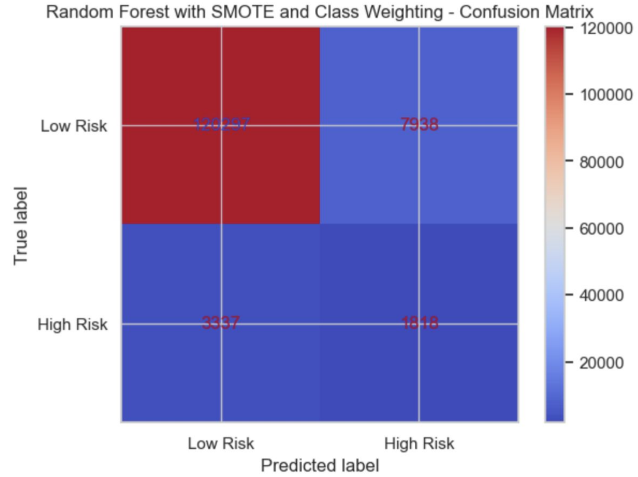


Random Forest Model Evaluation:

	precision	recall	f1-score	support
False	0.961354	1.000000	0.980296	128235.000000
True	0.000000	0.000000	0.000000	5155.000000
accuracy	0.961354	0.961354	0.961354	0.961354
macro avg	0.480677	0.500000	0.490148	133390.000000
weighted avg	0.924201	0.961354	0.942412	133390.000000

Improving Model Performance on RFM

- Applied **SMOTE** to oversample high-risk cases.
- Adjusted class weights to prioritize minority class.
- Optimized model parameters with GridSearchCV.
- Best Parameters:
    - max_depth: 20
    - max_features: ‘sqrt’
    - min_samples_split: 2
    - n_estimators: 100
    - Accuracy improved to **92%**.
    - High-risk cases still challenging.




```
Best Parameters for Random Forest with SMOTE: {'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min
_samples_split': 2, 'n_estimators': 100}

Classification Report:
              precision    recall  f1-score   support

    Low Risk       0.97      0.94      0.96    128235
   High Risk       0.19      0.35      0.24      5155

    accuracy                           0.92    133390
   macro avg       0.58      0.65      0.60    133390
weighted avg       0.94      0.92      0.93    133390
```

Key Findings

- The optimized model improved performance but struggled with high-risk classifications.
- Accuracy and F1-scores were good for low-risk cases but weak for high-risk cases.

Conclusion and Next Steps

- Incorporate more clinically relevant features.
- Test advanced techniques (deep learning, ensemble models).
- Improve data balancing techniques.
- Deploy interpretability tools (SHAP, LIME).

Recommendations for StakeHolders

- **Healthcare Professionals:** Use the model as a supplementary tool.
- **Researchers:** Focus on class imbalance solutions.
- **Policymakers:** Promote awareness campaigns.
- **Patients:** Use predictions as part of preventive care.