DATA REPORT ON CARDIOVASCULAR RISK ASSESSMENT

Resources

GitHub repository: [https://github.com/]

Members

1. Moses Kigo moses.wanja@student.moringaschool.com
2. Erik Lekishon eric.lekishon@student.moringaschool.com
3. Josephine Gathenya josephine.wanjiru@student.moringaschool.com
4. Chepkemoi Chepkemoi chepkemoi.chepkemoi@student.moringaschool.com
5. Eunita Nyengo eunita.nyengo@student.moringaschool.com

Business Understanding

Business Overview

According to the World Health Organisation (WHO), cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age.

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use, and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These "intermediate risk factors" can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure, and other complications.

Cessation of tobacco use, reduction of salt in the diet, eating more fruit and vegetables, regular physical activity, and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease. Health policies that create conducive environments for making healthy choices affordable and available are essential for motivating people to adopt and sustain healthy behaviors.

Identifying those at the highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. Access to non-communicable disease medicines and basic health technologies in all primary healthcare facilities is essential to ensure that those in need receive treatment and counseling.[1]

Therefore, there was a need to create a system that acts as a tool to assess the cardiovascular risk individuals have within various populations around the world. This system aims to assess and detect individuals who require prompt diagnosis and follow-up at the nearest health facility by giving an output ranging from mild, to moderate to severe risk with specific advice given in each case.

[1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Business Objectives

The **main objective** of this study is to build a machine-learning model that can predict the risk of cardiovascular disease based on the severity of the risk factors present in an individual.

Other objectives are to:

1. Categorize individuals' cardiovascular risk as mild, moderate or severe.
2. Use the outputs predicted together with the accompanying advice given by the model to help the users make informed decisions about their cardiovascular health and promptly seek medical attention.
3. Create a basis for future algorithms that are tailor-made for various populations in the world using data that is collected from those regions.

Success Criteria

This project will be successful when the following are achieved:

1. Completing the project within the stipulated timeframe and with the required resources
2. Meeting technical requirements by implementing various data science skills such as:
    a. Performing Data Cleaning and Exploratory Data Analysis
    b. Feature engineering & Modeling
    c. Model deployment
3. Meeting non-technical requirements such as being able to deploy the model as an interactive application to provide real-time assessments for end users.
4. Meeting all the objectives.

Methods

1. Personnel:

- ***a.*** **Project Manager:** Coordinates the project, ensuring timely progress and effective communication.
- ***b.*** **Data Analysts (2):** Conduct data cleaning and exploratory analysis, providing initial insights.
- ***c.*** **Data Scientists (2):** Develop the predictive model and perform clustering analysis, focusing on rapid iteration and validation.

2. ***Datasets***
   Cardio_train.csv from kaggle.com

3. ***Assumptions***
   We assumed that the dataset was not geographically or racially biased and that it was collected from real individuals.

4. ***Constraints***
   a. Short timeframe of 3.5 weeks
   b. Data is not representative of all populations around the world which is expected to be different for every region based on factors such as race and genetics

Data Understanding

Overview

The dataset used in this project was Cardio_train.csv from Kaggle.com.

Data Description

The dataset was derived from the Cardio_train.csv file which contains 70,000 respondents with features such as age, gender, blood pressure, cholesterol level, and other medical history indicators. This comprehensive dataset provides a basis for developing robust predictive models.

There were 3 types of input features:

1. **Objective**: factual information;
2. **Examination**: results of medical examination;
3. **Subjective**: information given by the patient.

Statistical Summary:

1. Age: Age is in days, translated age from days to years to make it more interpretable. The age ranges from about 29.6 to 64.9 years.
2. Height: There were potentially erroneous entries, as the minimum height is 55 cm and the maximum is 250 cm.

3. Weight: The weight ranged from 10 kg to 200 kg also suggests some potentially erroneous data.
4. Blood Pressure (ap_hi, ap_lo): Some values were negative or excessively high, which are errors (e.g., -150 mmHg and 16020 mmHg for systolic, -70 mmHg and 11000 mmHg for diastolic).
5. Cholesterol and Glucose: These were categorical variables encoded as 1 (normal), 2 (above normal), and 3 (well above normal).

Verifying data quality

Data Preparation

1. **Loading Data**

2. **Data cleaning**

Preprocessing

There were no missing values or duplicated rows in the dataset. 5 columns have 2 unique values and 2 columns have three unique values each.

Feature Engineering

We performed feature engineering to get new features that would give us more objective variables that will be used in creating our model. The following steps were followed:

1. Created a new column with age in years instead of days, for easier interpretation and categorized age into 7 category age groups.
2. We then created a new BMI column which is a better risk factor to use to determine the risk of cardiovascular disease and categorized the BMI into underweight, normal, overweight, and obese using codes 1,2,3,4 respectively.
3. Other new columns created were for the blood pressure category and the calculated pulse pressure category
4. We checked the correlation between all the features with cardiovascular risk and noted features with positive correlation and those with negative correlation.
5. We then assessed for any outliers and removed them by using a limit range for our analysis i.e.

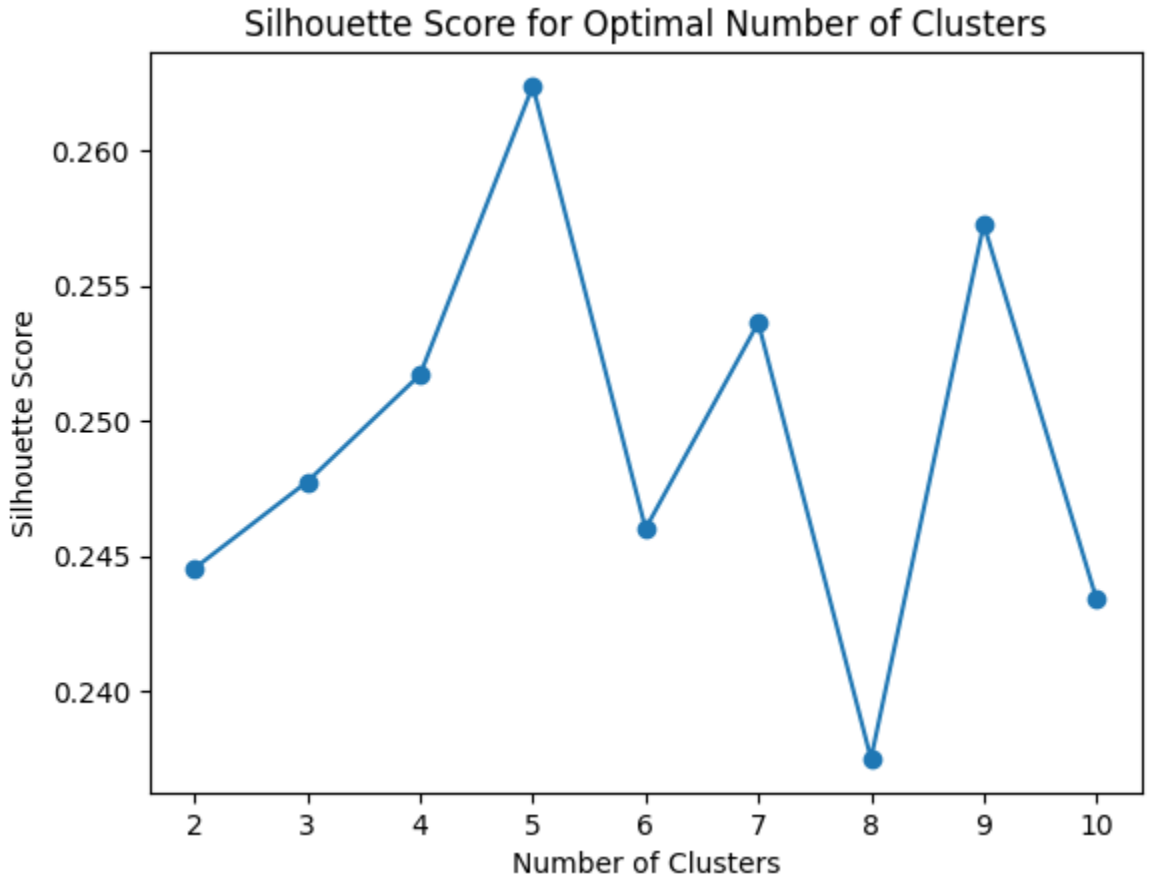
a. **Blood Pressure:** We limited systolic (ap_hi) and diastolic (ap_lo) blood pressure to reasonable clinical adult ranges, such as systolic from 90 to 250 mmHg, and diastolic from 60 to 150 mmHg.
b. **Height and Weight**: Remove heights below 50 cm and above 250 cm. In the same way, filtered out weights below 30 kg or above 200 kg to improve accuracy.

6. We re-assessed the correlation between the feature variables, including the new features, and cardiovascular risk.

Clustering and Segmentation

We performed clustering & segmentation on our new dataset using the features with the highest positive correlation to cardiovascular risk i.e. age group category, blood pressure category, pulse pressure category, category, cholesterol, and glucose.

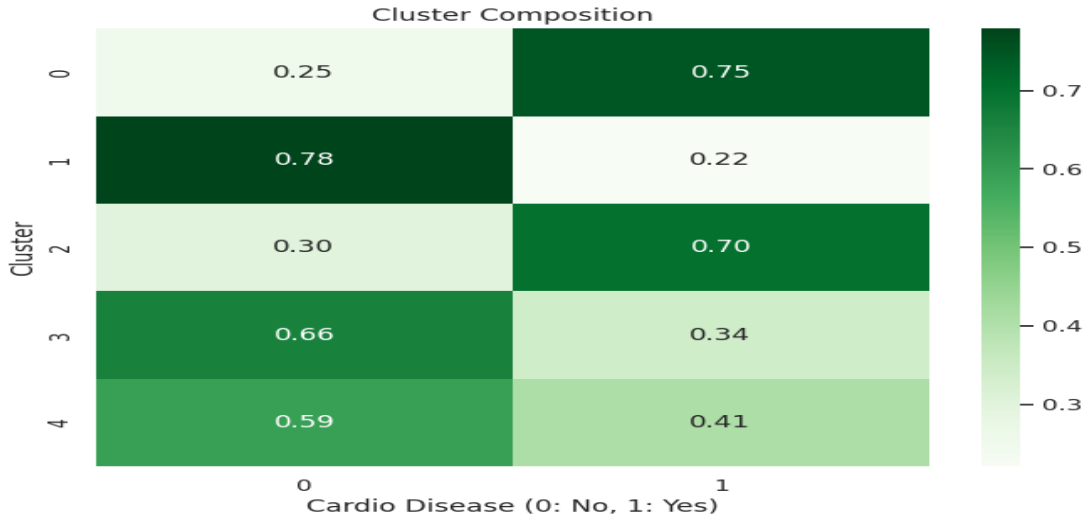
We identified the optimal number of clusters as 5 from the Silhouette score, using K-means clustering.



We then explored the feature distribution and the mean value for every feature in each of the 5 clusters.

We deployed the clustering model and segmented it into user-relatable outputs using animals related to each of the clusters as follows;

a. Cluster 0 - The High-Risk Elderly: The Wise Old Elephant
b. Cluster 1 - The Younger, Healthier Subset: The Agile Gazelle
c. Cluster 2 - The Metabolic Challenge Group: The Burdened Bear
d. Cluster 3 - Young, Yet at Risk: The Cheetah Cub
e. Cluster 4 - Well-Managed Mature Group: The Seasoned Horse



Predictive Modeling

Using the same features that had the highest positive correlation with cardiovascular risk, we trained, validated, and tested out the dataset using 6 models as follows:

Linear Regression

We decided to use linear regression as our baseline model because of its simplicity and easy interpretability.

```
# Create and train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Calculate metrics
accuracy = accuracy_score(y_test, np.round(y_pred))
roc_auc = roc_auc_score(y_test, y_pred)
recall = recall_score(y_test, np.round(y_pred))
precision = precision_score(y_test, np.round(y_pred))
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
```

The model performed as follows;

1. Linear Regression Model Performance:
2. Accuracy: 0.7133406034169393
3. ROC AUC Score: 0.7763864270613108
4. Recall: 0.6443636363636364
5. Precision: 0.7473009446693657
6. Root Mean Squared Error (RMSE): 0.44255496716948906

Modeling using other modeling methods

We created other models using the following methods:

1. Logistic regression
2. KNN
3. Random Forest

4. XG Boost
5. Decision trees

We used one code to measure different metrics in all the above models, before and after tuning;

```
# Helper function to calculate metrics
def evaluate_model(model, X_train, y_train, X_val, y_val):
    model.fit(X_train, y_train)
    y_pred_train = model.predict(X_train)
    y_pred_val = model.predict(X_val)
    y_pred_prob_val = model.predict_proba(X_val)[:, 1] if hasattr(model, "predict_proba") else y_pred_val

    metrics = {
        'Train Accuracy': round(accuracy_score(y_train, y_pred_train) * 100, 2),
        'Validation Accuracy': round(accuracy_score(y_val, y_pred_val) * 100, 2),
        'Validation ROC AUC': round(roc_auc_score(y_val, y_pred_prob_val) * 100, 2) if hasattr(model, "predict_proba") else N
        'Validation Recall': round(recall_score(y_val, y_pred_val) * 100, 2),
        'Validation Precision': round(precision_score(y_val, y_pred_val) * 100, 2)
    }
    if hasattr(model, "predict_proba"):
        metrics['Validation RMSE'] = round(math.sqrt(mean_squared_error(y_val, y_pred_prob_val)), 2)
    return metrics

# Define models and parameters
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000, random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'KNN': KNeighborsClassifier(),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'XGBoost': xgb.XGBClassifier(random_state=42, use_label_encoder=False, eval_metric='logloss')
}
```

Evaluation

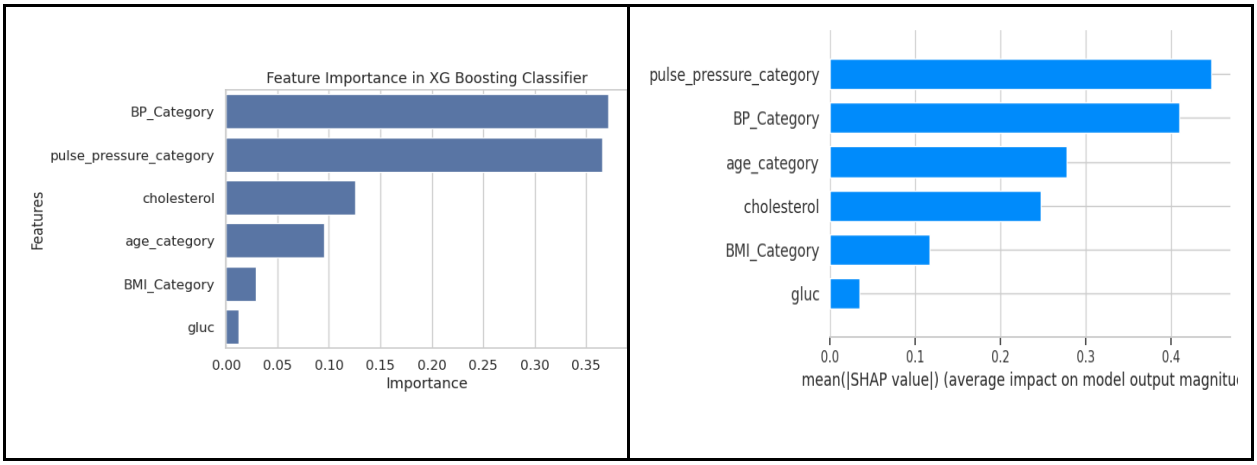
The following is a summary of our findings:

	Before Tuning						After Tuning					
Metric	Train Accuracy	Validation Accuracy	Validation Precision	Validation RMSE	Validation ROC AUC	Validation Recall	Train Accuracy	Validation Accuracy	Validation Precision	Validation RMSE	Validation ROC AUC	Validation Recall
Model												
Decision Tree	72.34	70.34	72.83	0.44	76.64	64.16	72.14	70.51	72.87	0.44	76.97	64.62
KNN	69.54	67.54	70.70	0.47	72.07	59.05	70.59	69.07	72.10	0.46	74.47	61.45
Logistic Regression	70.81	70.38	73.11	0.44	76.87	63.75	70.81	70.38	73.11	0.44	76.87	63.75
Random Forest	72.34	70.43	72.71	0.44	76.88	64.66	72.05	70.73	73.22	0.44	77.57	64.66
XGBoost	72.10	70.64	73.16	0.44	77.49	64.50	71.73	71.09	73.22	0.44	77.83	65.79

The areas highlighted in red indicate model metrics that decreased after tuning, while the ones in green show metric improvement. Those that were not highlighted remained unchanged. Overall, the Random Forest and XG Boost models performed best.

Feature importance and impact analysis

We performed a feature importance and a feature impact analysis using XG Boost which was the best-performing model and noted that the pulse pressure and blood pressure category features had the highest impact on CVD risk.



Model Deployment

We deployed the model using the joblib library.

```
# Export Models for Deployment

# Save the clustering and predictive models
joblib.dump(stacking_clf, 'stacking_classifier_model.pkl')
joblib.dump(kmeans, 'patient_clustering_model.pkl')
joblib.dump(scaler, 'scaler.pkl')

# Export requirements.txt
!pip freeze > requirements.txt
```

The expected output would be one of the five clusters represented by an animal icon with the specific advice given, as shown below: