PGPDSE FT Capstone Project- Final Report

No. of Batchmates	8
Name	Akshay Yadav, Shubham, Harsh Sharma, Nabendu Mondal, Aviral Meharishi, Harshit Kumar, Avinash Mishra, Kandara Kanishk
Project domain	Healthcare
Project Title	Prediction of Cardiovascular Disease Risk Using Lifestyle Factors
Mentor Name	Jatin Bedi
Group Number	3

1. Summary of Problem Statement

Cardiovascular diseases (CVDs) are a major cause of mortality worldwide. Traditional clinical risk assessments focus on physiological factors but often overlook lifestyle influences. This project aims to develop a predictive model for CVD risk using behavioral and lifestyle factors from the Behavioral Risk Factor Surveillance System (BRFSS) dataset.

The dataset consists of 308,854 entries with 25 variables, including general health, medical checkups, exercise habits, diet, BMI, smoking history, and more. Exploratory Data Analysis (EDA) has revealed key correlations between lifestyle factors and heart disease, highlighting the importance of early lifestyle interventions.

Additionally, this project seeks to assess the probability of a person getting cardiovascular disease based on various environmental and lifestyle factors using different machine learning models.

Data description:

Column Name	Data Type	Description
ID	Int	Identifier
General Health	Object	General health rating of individuals (eg. excellent, good, poor)
Checkup	Object	Frequency of medical checkups

Exercise	Object	Indicator of whether the individual exercises regularly (eg. Yes, No)
Heart Disease	Object	Target variable indicating the presence of heart disease (eg. Yes, No)
Skin Cancer	Object	Indicator of whether the individual has skin care (eg. Yes, No)
Other Cancer	Object	Indicator of whether the individual has other types of cancer (eg. Yes, No)
Depression	Object	Indicator of whether the individual has been diagnosed with depression
Diabetes	Object	Indicator of whether the individual has diabetes (eg Yes, No)
Arthritis	Object	Indicator of whether one has arthritis (eg. Yes, No)
Sex	Object	Sex of the individual (eg. Male, Female)
Age Category	Object	Categorical representation of age groups (eg. 18-24,25-34)
Height (cm)	float64	Height of the individual in centimeters
Weight (kg)	float64	Weight of the individual in kilogram
BMI	float64	Body mass index of the individual
Smoking History	Object	History of smoking (eg. current, former, never)
Alcohol consumption	float64	Amount of alcohol consumed
Fruit consumption	float64	Frequency of fruit consumption
Green vegetable consumption	float64	Frequency of green vegetable consumption

Fried potato consumption	float64	Frequency of fried potato consumption
BP	Object	Blood pressure of the individual (eg. 0- Normal BP, 1- Abnormal BP)
Blood gp	Object	The blood group of the individual (eg. A, B, O)
Vaccinated	Object	Whether the individual is vaccinated or not (eg 0- No, 1- Yes)
Marital Status	Object	Whether the individual is married or not (eg 0- Divorced/Single, 1- Married)
Diet Preference	Object	Diet preference of the individual (eg. Veg, Non-Veg)

Out of the 25 features, there are 3 numerical features and 22 categorical features

- **Numeric Variables (3):**
  - Height_(cm)
  - Weight_(kg)
  - BMI

- **Categorical Variables (22):**
  - General_Health
  - Checkup
  - Exercise
  - Heart_Disease
  - Skin_Cancer
  - Other_Cancer
  - Depression
  - Diabetes
  - Arthritis
  - Sex
  - Age_Category
  - Smoking_History
  - Alcohol_Consumption
  - Fruit_Consumption
  - Green_Vegetables_Consumption
  - FriedPotato_Consumption
  - BP
  - Blood_gp
  - Vaccinated
  - Marital_status

- ○ Diet_pref

Preliminary Findings:

There were a few observations that we made before starting any sort of analysis of the data:

1) Since the ID column is not a unique identifier hence we drop it.
2) Columns mentioned in the list are categories but assigned data type float which is to be changed to object.
3) The target columns (Heart Disease) have the following properties-
    - The variable is binary
    - The majority of the data indicates no heart disease (91.91%)
    - A small percentage indicates heart disease (8.09%)
    - There's a significant class imbalance. The "No" class heavily outweighs the "Yes" class. This is a common issue in medical datasets and can impact the performance of machine learning models.

2. Overview of the Final Process

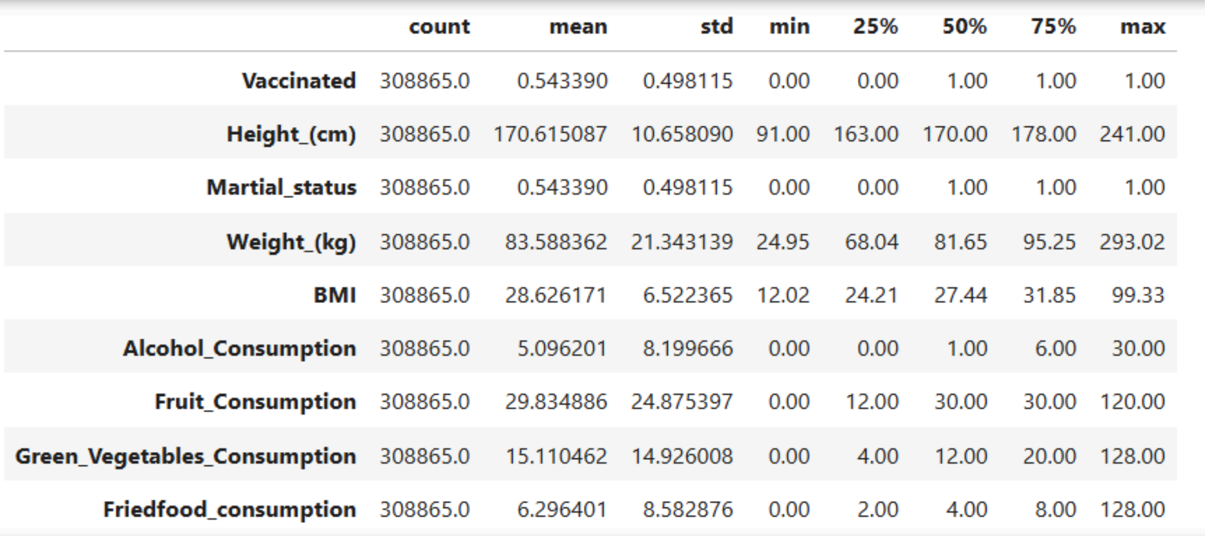
We followed a structured data science approach:

- **Descriptive Statistics:** Understanding our data before we move on to more complex analyses or modeling
- **Data Preprocessing**: Handled missing values, corrected data types, and removed redundant columns.
- **Exploratory Data Analysis**: Identified key lifestyle factors influencing CVD risk.
- **Statistical analysis:** To identify significant features in the data.
- **Feature Engineering**: Encoded categorical variables, checked for multicollinearity, and assessed feature importance.
- **Model Selection and Training**: Evaluated various machine learning algorithms to predict heart disease risk.
- **Evaluation and Interpretation**: Assessed model performance and compared it with benchmark standards.
- **Deployment:** Explored potential real-world applications and limitations and finally deployed the model

3. Step-by-step walkthrough of the Solution

Descriptive statistics:

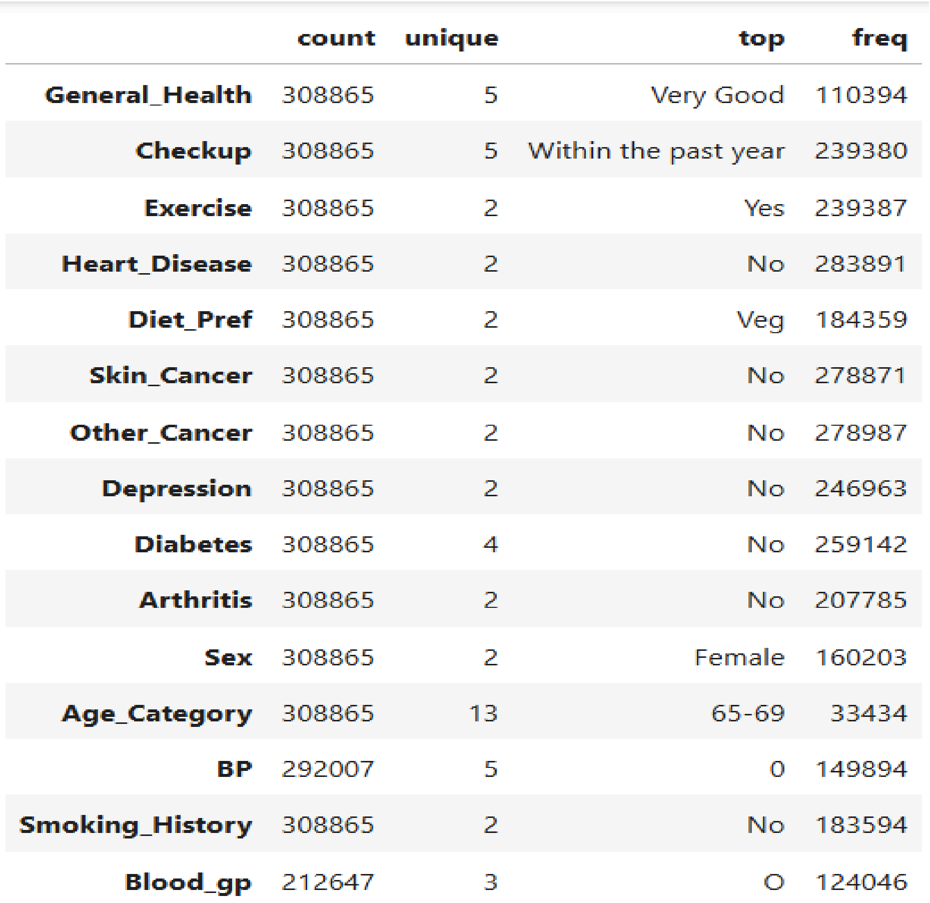
When descriptive statistics was applied to the numerical columns the following inferences were made:



	count	mean	std	min	25%	50%	75%	max
Vaccinated	308865.0	0.543390	0.498115	0.00	0.00	1.00	1.00	1.00
Height_(cm)	308865.0	170.615087	10.658090	91.00	163.00	170.00	178.00	241.00
Martial_status	308865.0	0.543390	0.498115	0.00	0.00	1.00	1.00	1.00
Weight_(kg)	308865.0	83.588362	21.343139	24.95	68.04	81.65	95.25	293.02
BMI	308865.0	28.626171	6.522365	12.02	24.21	27.44	31.85	99.33
Alcohol_Consumption	308865.0	5.096201	8.199666	0.00	0.00	1.00	6.00	30.00
Fruit_Consumption	308865.0	29.834886	24.875397	0.00	12.00	30.00	30.00	120.00
Green_Vegetables_Consumption	308865.0	15.110462	14.926008	0.00	4.00	12.00	20.00	128.00
Friedfood_consumption	308865.0	6.296401	8.582876	0.00	2.00	4.00	8.00	128.00

- Some individuals have a max height of 241 cm and a minimum height of 91 cm.
- There are individuals who are highly low weight and also heavyweight.
- The average BMI is 28.62 which falls under the overweight category.

When the same was done for the categorical columns:



	count	unique	top	freq
General_Health	308865	5	Very Good	110394
Checkup	308865	5	Within the past year	239380
Exercise	308865	2	Yes	239387
Heart_Disease	308865	2	No	283891
Diet_Pref	308865	2	Veg	184359
Skin_Cancer	308865	2	No	278871
Other_Cancer	308865	2	No	278987
Depression	308865	2	No	246963
Diabetes	308865	4	No	259142
Arthritis	308865	2	No	207785
Sex	308865	2	Female	160203
Age_Category	308865	13	65-69	33434
BP	292007	5	0	149894
Smoking_History	308865	2	No	183594
Blood_gp	212647	3	O	124046

- Around 33% of individuals present in the data set have very good general health.
- More than 50% got their checkup done within a year and exercised regularly.
- Data has an equal distribution of males and females.

Data Preprocessing:

- **Checking for Duplicates & Treating:**
  33 duplicate columns were dropped immediately

- **Checking for Null/Missing Values & Treating**
  2 columns had missing values:
    - BP- 16860
    - Blood_group- 96220

  All the missing values in these columns were dropped

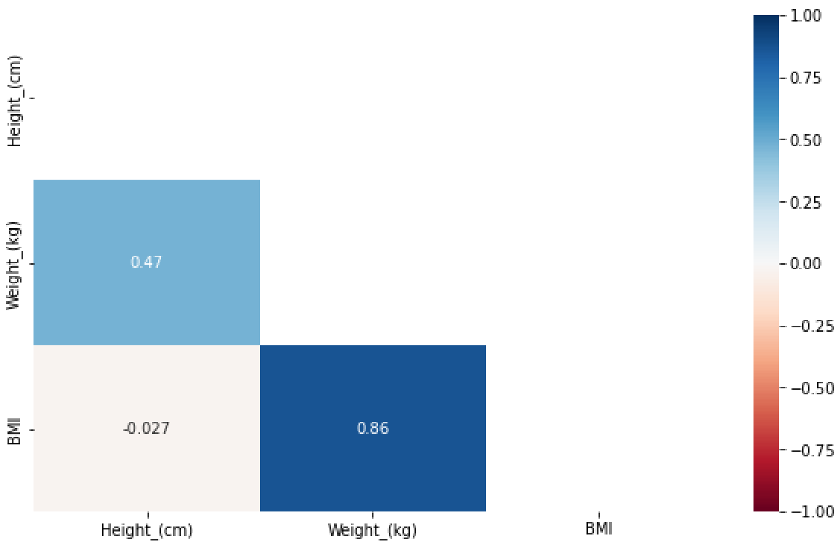
- **Redundant Columns:**
  - There is one redundant column - ID which is dropped

- **Datatype Correction:**
  - Various columns are categorical in nature but are typecast as float, we corrected their data types.

Exploratory Data Analysis:

Relationship between variables

Correlation plot (Numerical values):



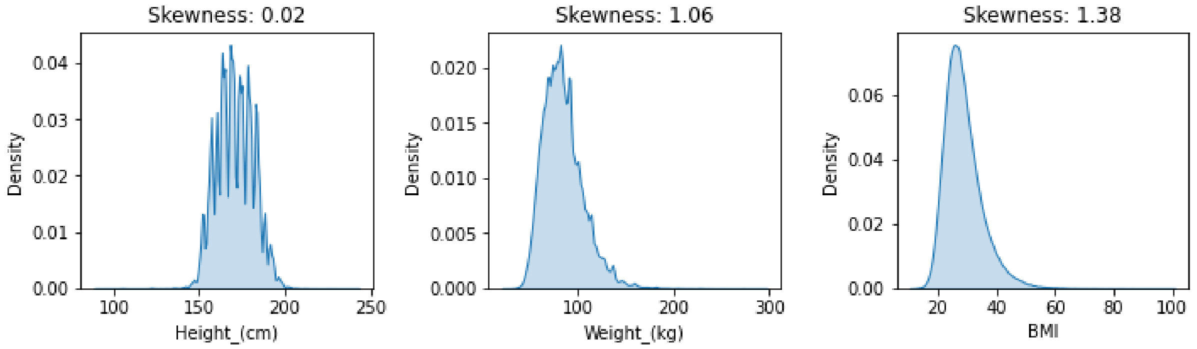
Inferences

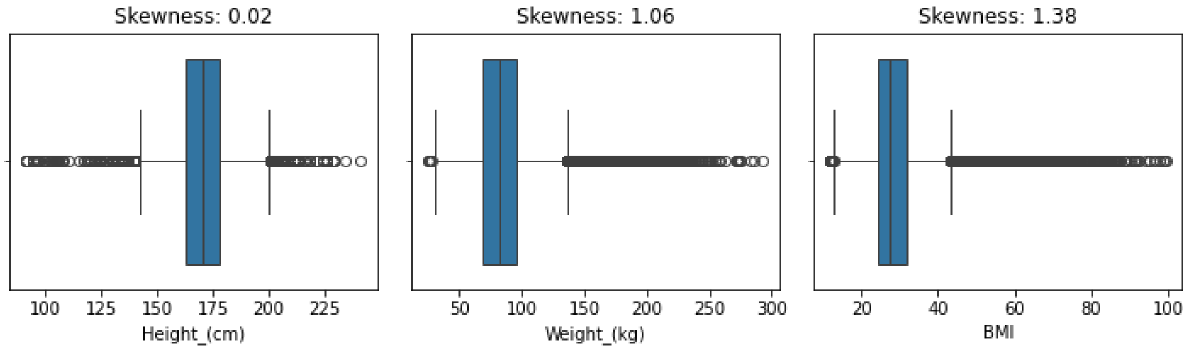
- Weight and BMI have a strong positive correlation (0.86), which makes sense since BMI is calculated using weight. This indicates that as weight increases, BMI tends to increase substantially as well.
- Height and Weight have a moderate positive correlation (0.47). This suggests that taller people tend to weigh more, but the relationship isn't as strong as you might expect, likely because weight can vary significantly among people of the same height.

Interestingly, Height and BMI have a very slight negative correlation (-0.027), which is almost zero. This tells us that height has virtually no direct relationship with BMI e.i being taller doesn't necessarily mean you'll have a higher or lower BMI

Univariate Analysis:

Inferences for numerical columns:

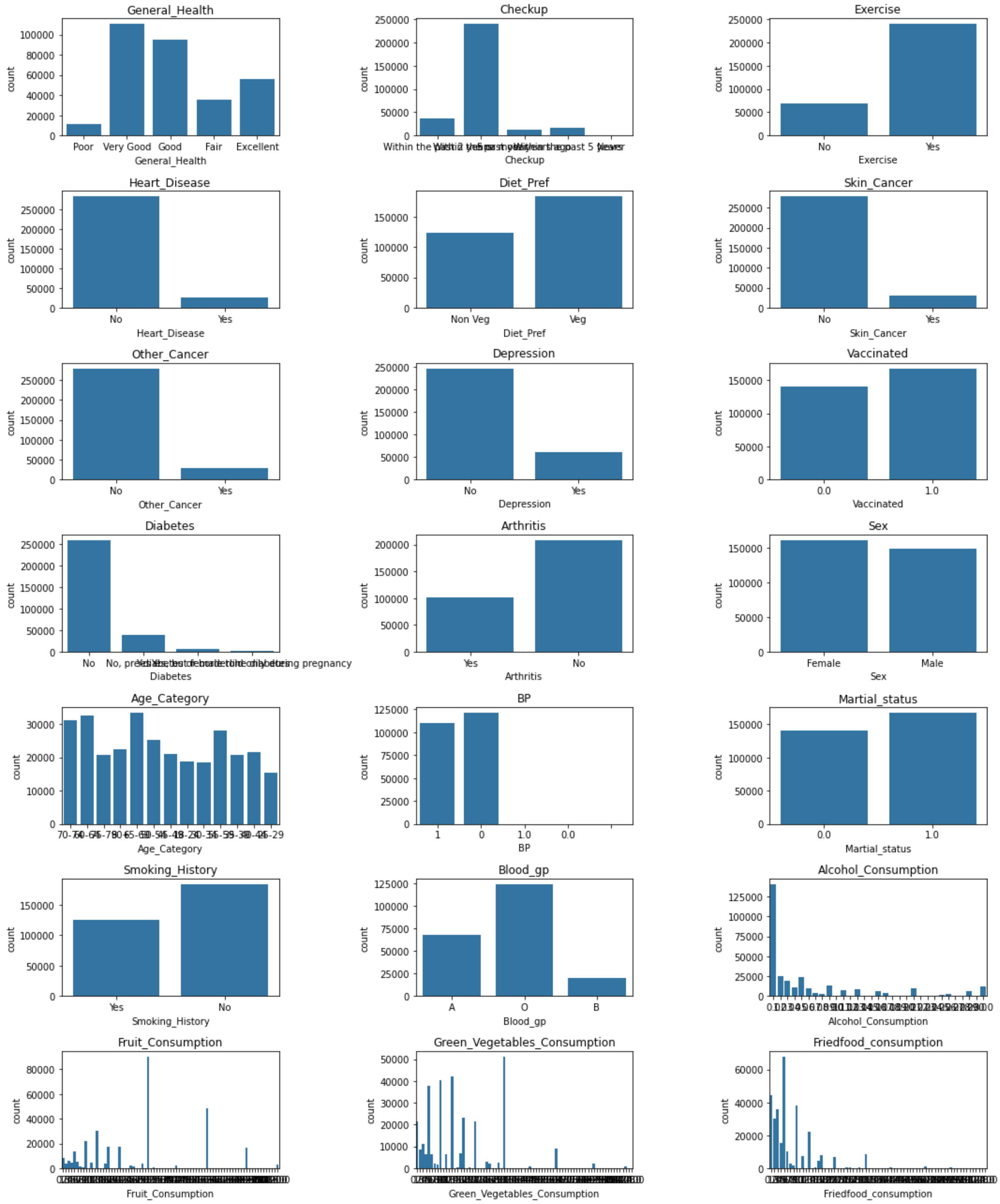




1. The data follows a normal distribution but there is moderate positive skewness in columns like weight and BMI.

2. The Box plot shows there might be outliers but the density flow says that they might not be outliers.

Inferences for categorical columns:



1. Most individuals report fair or good general health and very few has poor health.

2. There are more individuals who have had a checkup within a year compared to those who have not.

3. Most individuals do not exercise regularly.

4. Most individuals prefer vegetarian food.

5. Most individuals do not have skin cancer or other types of cancer.

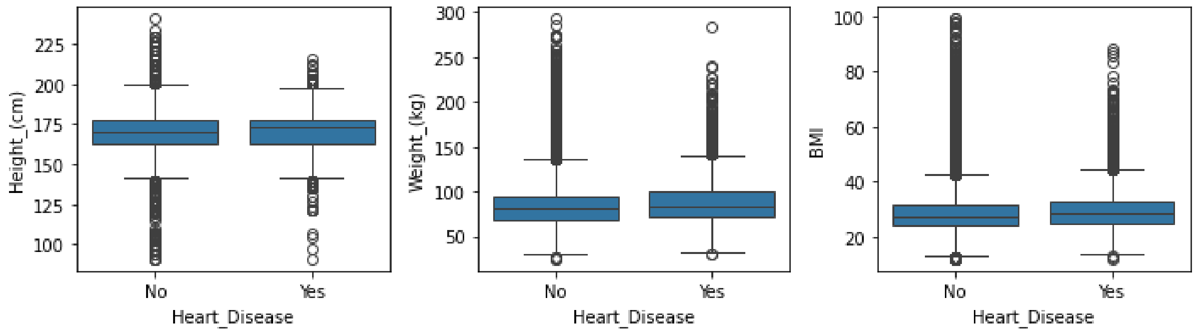
6. A large majority of individuals do not have depression and also do not have diabetes.

7. Most individuals do not smoke.

8. The distribution of blood groups appears to be relatively even with blood group O being majority.

Bivariate Analysis:

Inferences for numerical columns:



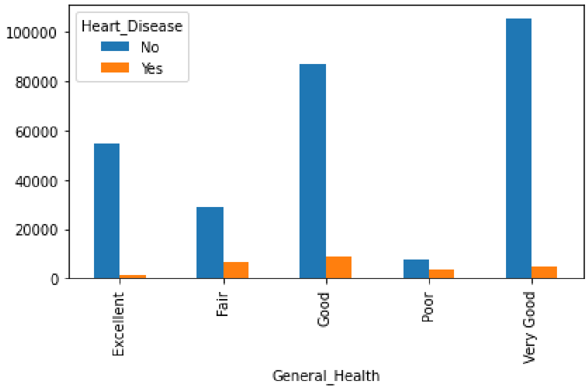
1. The distribution of heights appears to be similar between individuals with and without heart disease.

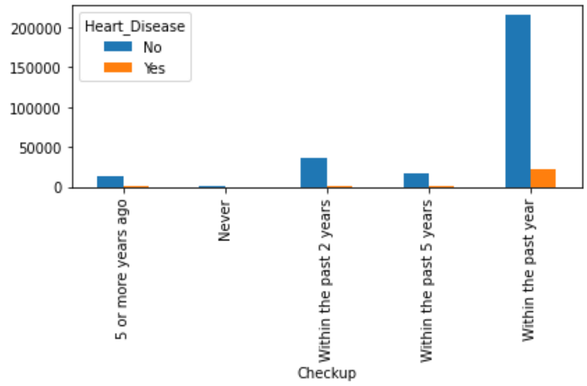
2. There is a slight suggestion that individuals with heart disease might have a slightly lower median height, but the difference is not very pronounced.

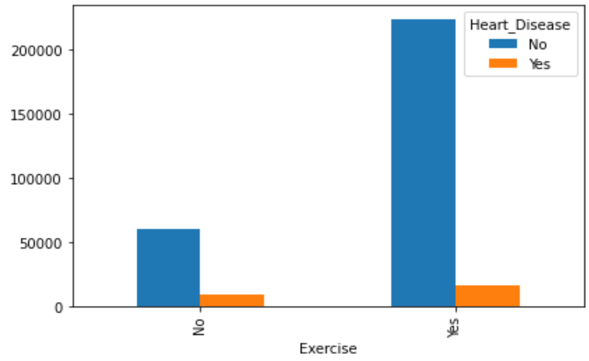
3. Individuals with heart disease seem to have a higher median weight compared to those without heart disease.

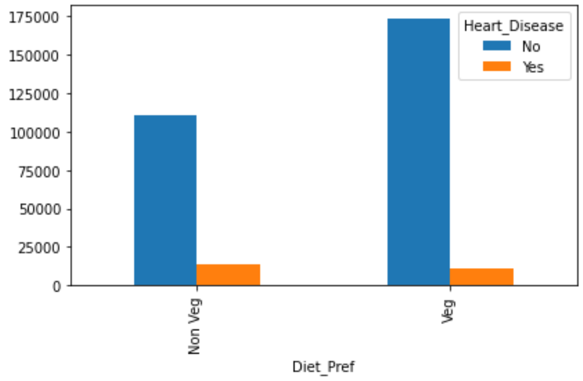
4. Like weight, individuals with heart disease tend to have a higher median BMI compared to those without heart disease.

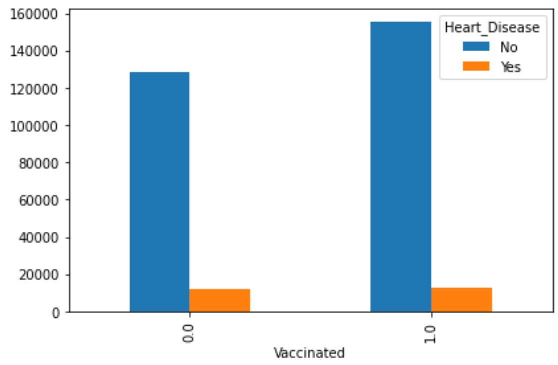
Inferences for categorical columns:

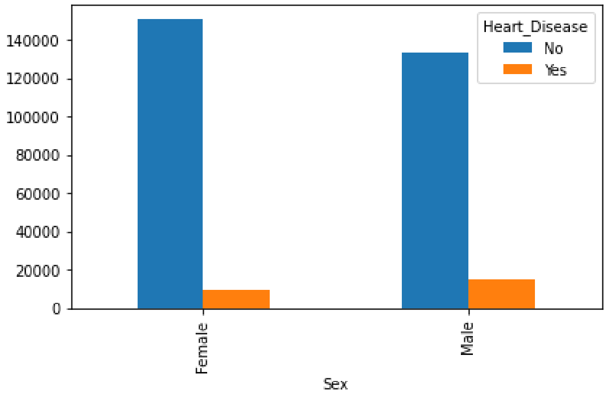












1. As general health declines (from Excellent to Poor), the proportion of individuals with heart disease increases.

2. The chart shows a clear trend that individuals who have not had a checkup in a long time (5 or more years ago or never) have a higher proportion of heart disease compared to those who have had a checkup more recently.

3. The chart clearly shows that individuals who do not exercise regularly have a significantly higher proportion of heart disease compared to those who do exercise.

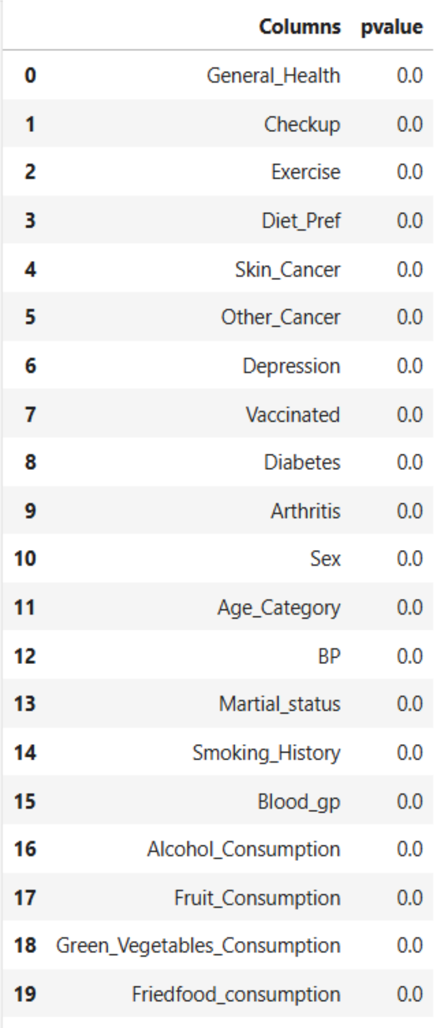
4. Individuals with a "Non-Veg" diet preference have a significantly higher proportion of heart disease compared to those with a "Veg" diet preference.

5. Individuals being vaccinated or not does not increases their chances of having a heart disease.

6. The chart shows that male tends to suffer from heart disease more than females.

Check for statistical significance of variables

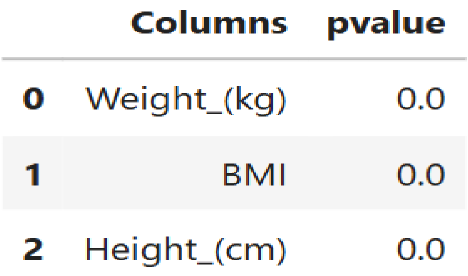
We used the chi-square test to first check the significance of the categorical columns in the data against the target variable



	Columns	pvalue
0	General_Health	0.0
1	Checkup	0.0
2	Exercise	0.0
3	Diet_Pref	0.0
4	Skin_Cancer	0.0
5	Other_Cancer	0.0
6	Depression	0.0
7	Vaccinated	0.0
8	Diabetes	0.0
9	Arthritis	0.0
10	Sex	0.0
11	Age_Category	0.0
12	BP	0.0
13	Martial_status	0.0
14	Smoking_History	0.0
15	Blood_gp	0.0
16	Alcohol_Consumption	0.0
17	Fruit_Consumption	0.0
18	Green_Vegetables_Consumption	0.0
19	Friedfood_consumption	0.0

All the columns listed have a p-value of 0.0. This indicates a significant association between these variables and the target column (Heart_Disease_Encoded). Since we used the chi-square test, it suggests that these variables are likely important predictors of heart disease.

Next, we used the ANOVA test to check the significance of numerical columns in the data against the target variable



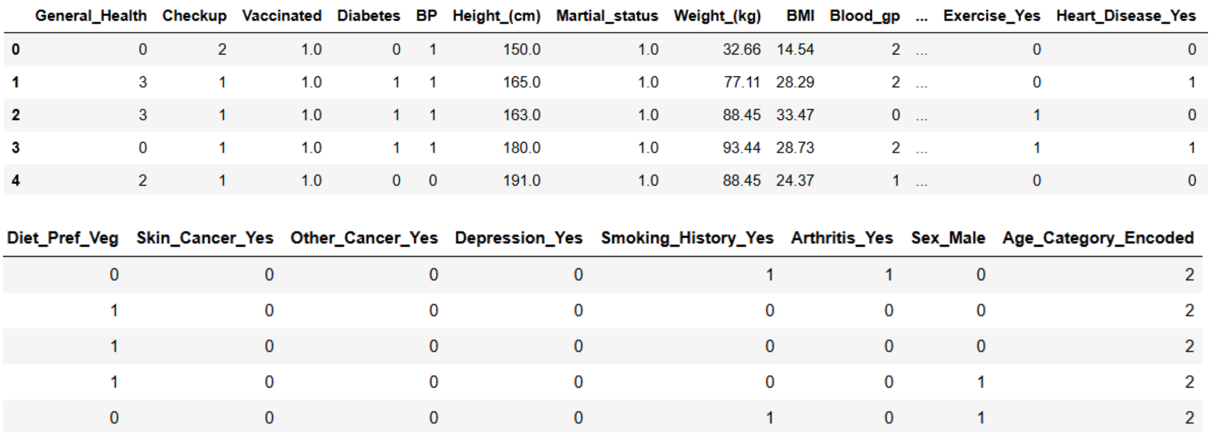
	Columns	pvalue
0	Weight_(kg)	0.0
1	BMI	0.0
2	Height_(cm)	0.0

1. Highly Significant Relationships: The p-values of 0.0 for Weight, BMI, and Height indicate these variables have extremely statistically significant relationships with heart disease.
2. This means:
   The null hypothesis of no relationship between these variables and heart disease can be rejected
   These physical measurements are strongly associated with heart disease occurrence
   The relationships are not due to random chance

Feature Engineering:

Feature encoding:

- Feature encoding is performed on various columns like Checkup, Diabetes, Blood_gp, etc.
- Many of these columns had 2 categories which were encoded using N-1 Dummy encoding
- The remaining columns were encoded using mapping as they had more than 2 categories.

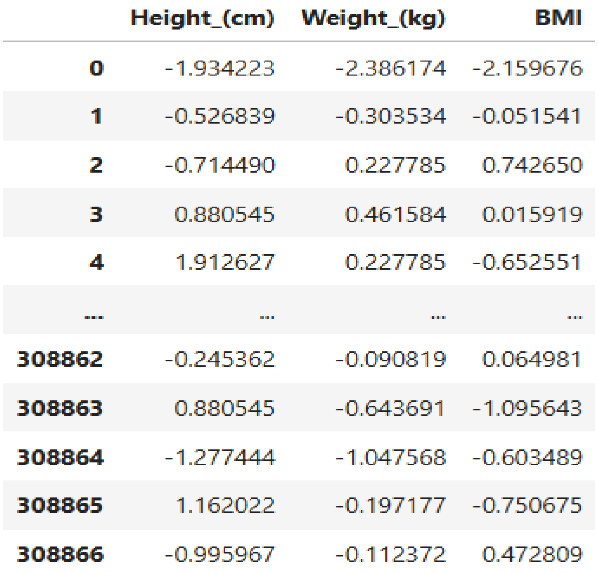


	General_Health	Checkup	Vaccinated	Diabetes	BP	Height_(cm)	Martial_status	Weight_(kg)	BMI	Blood_gp	...	Exercise_Yes	Heart_Disease_Yes
0	0	2	1.0	0	1	150.0	1.0	32.66	14.54	2	...	0	0
1	3	1	1.0	1	1	165.0	1.0	77.11	28.29	2	...	0	1
2	3	1	1.0	1	1	163.0	1.0	88.45	33.47	0	...	1	0
3	0	1	1.0	1	1	180.0	1.0	93.44	28.73	2	...	1	1
4	2	1	1.0	0	0	191.0	1.0	88.45	24.37	1	...	0	0

Diet_Pref_Veg	Skin_Cancer_Yes	Other_Cancer_Yes	Depression_Yes	Smoking_History_Yes	Arthritis_Yes	Sex_Male	Age_Category_Encoded
0	0	0	0	1	1	0	2
1	0	0	0	0	0	0	2
1	0	0	0	0	0	0	2
1	0	0	0	0	0	1	2
0	0	0	0	1	0	1	2

Feature scaling:

The numerical columns were scaled to get a better score for the KNN classifier. The columns' Height_(cm)', 'Weight_(kg)', and 'BMI' were standardized using StandardScaler.

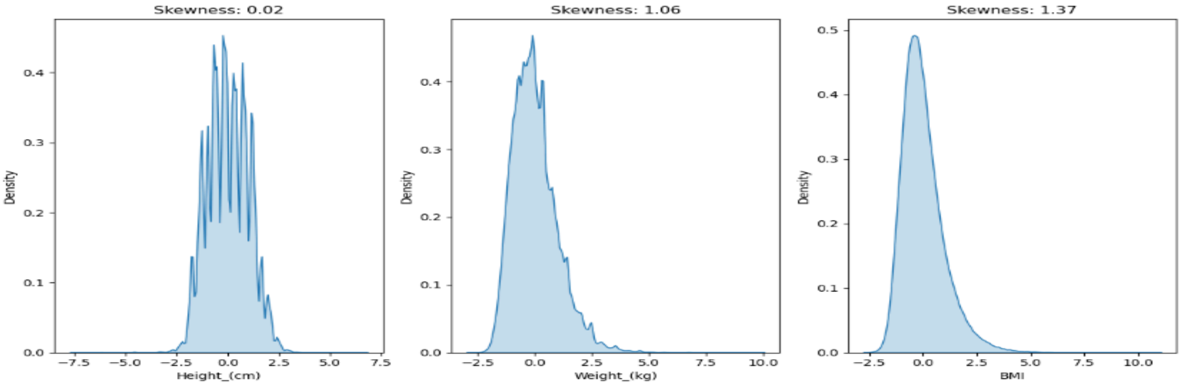


	Height_(cm)	Weight_(kg)	BMI
0	-1.934223	-2.386174	-2.159676
1	-0.526839	-0.303534	-0.051541
2	-0.714490	0.227785	0.742650
3	0.880545	0.461584	0.015919
4	1.912627	0.227785	-0.652551
...	...	...	...
308862	-0.245362	-0.090819	0.064981
308863	0.880545	-0.643691	-1.095643
308864	-1.277444	-1.047568	-0.603489
308865	1.162022	-0.197177	-0.750675
308866	-0.995967	-0.112372	0.472809

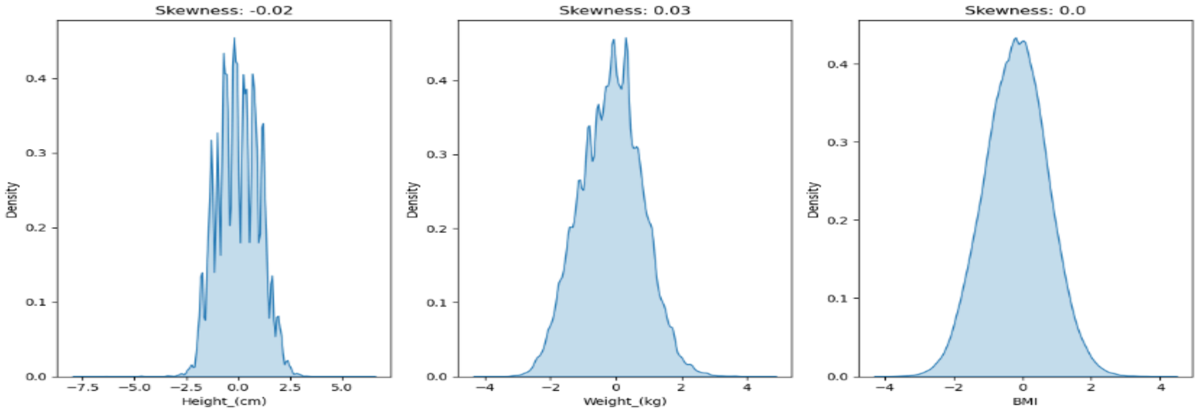
Feature Transformation:

The numerical columns were further transformed (PowerTransformer) since they had high skewness.

Before transformation:



After transformation:



4. Model selection and training:

Splitting the data into train and test:

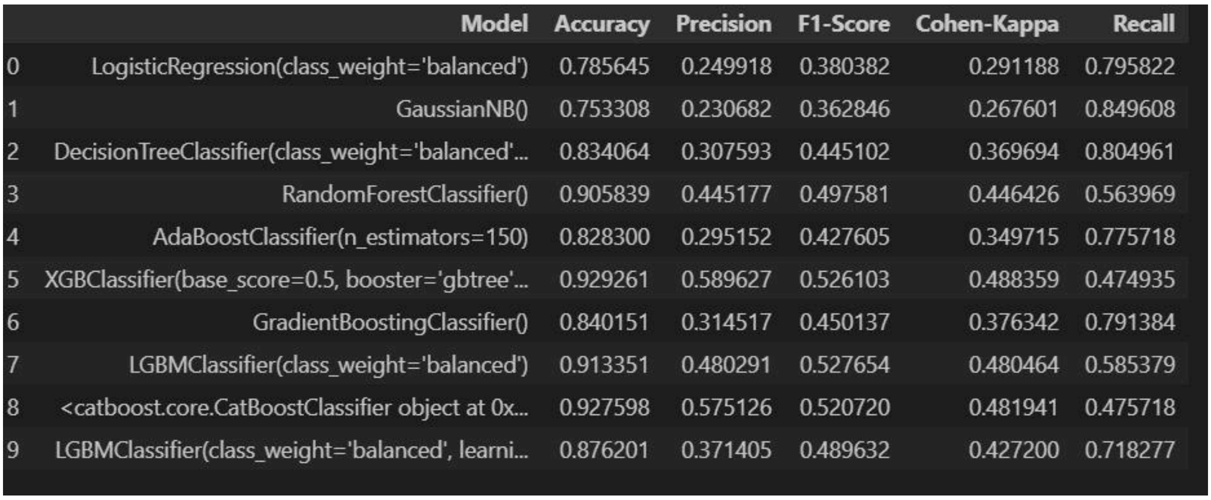
The dataset was split into train and test data (85/15 split) based on features (x) and the binary target variable (y).

The datasets had the following shapes:

	Rows	Columns
X-train	262508	22
X-test	46325	22
Y-train	262508	-
Y-test	46325	-

To evaluate the appropriate model we fit different machine learning models to check which model gave the highest score

A scorecard was made to analyze the different scores and parameters of each model to get a better idea of which model performs the best.



```
                                              Model  Accuracy  Precision  F1-Score  Cohen-Kappa    Recall
0       LogisticRegression(class_weight='balanced')  0.785645   0.249918  0.380382     0.291188  0.795822
1                                      GaussianNB()  0.753308   0.230682  0.362846     0.267601  0.849608
2  DecisionTreeClassifier(class_weight='balanced'...  0.834064   0.307593  0.445102     0.369694  0.804961
3                          RandomForestClassifier()  0.905839   0.445177  0.497581     0.446426  0.563969
4             AdaBoostClassifier(n_estimators=150)  0.828300   0.295152  0.427605     0.349715  0.775718
5  XGBClassifier(base_score=0.5, booster='gbtree'...  0.929261   0.589627  0.526103     0.488359  0.474935
6                      GradientBoostingClassifier()  0.840151   0.314517  0.450137     0.376342  0.791384
7           LGBMClassifier(class_weight='balanced')  0.913351   0.480291  0.527654     0.480464  0.585379
8   <catboost.core.CatBoostClassifier object at 0x...  0.927598   0.575126  0.520720     0.481941  0.475718
9   LGBMClassifier(class_weight='balanced', learni...  0.876201   0.371405  0.489632     0.427200  0.718277
```

Final Model: LGBM Classifier:

The LightGBM (LGBM) classifier was chosen as the final model for predicting the probability of cardiovascular disease in individuals. This decision was based on a thorough evaluation of various machine learning models using multiple performance metrics.

How does LGBM work:

LightGBM (Light Gradient Boosting Machine) is a powerful gradient-boosting framework designed for speed and efficiency, making it highly suitable for large datasets. It is based on Gradient Boosting Decision Trees (GBDT), where trees are built sequentially, with each new tree learning from the errors of the previous ones by minimizing the loss function using gradient descent. Unlike traditional boosting methods, LightGBM introduces several optimizations to enhance training speed and accuracy. One of its key features is histogram-based learning, which bins continuous values into discrete bins instead of evaluating every data point, significantly reducing computation time. Additionally, LightGBM grows trees leaf-wise (best-first) rather than level-wise (breadth-first) like traditional methods, resulting in deeper and more accurate trees.

Objective of the Project:

The goal was to predict the likelihood of cardiovascular disease in individuals based on various health, lifestyle, and demographic factors. Given the dataset's mix of categorical and numerical features, models were trained to classify individuals as at risk or not at risk for cardiovascular disease.

Prominent Model Parameters

For the final LGBMClassifier, the following parameters played a significant role:

- class_weight='balanced': This helped address any class imbalance in the dataset.
- learning_rate: A controlled learning rate was used to optimize model convergence.
- Boosting Type (Gradient Boosting Trees): LightGBM's tree-based boosting mechanism helped efficiently capture complex patterns in the data.

Performance Evaluation & Robustness

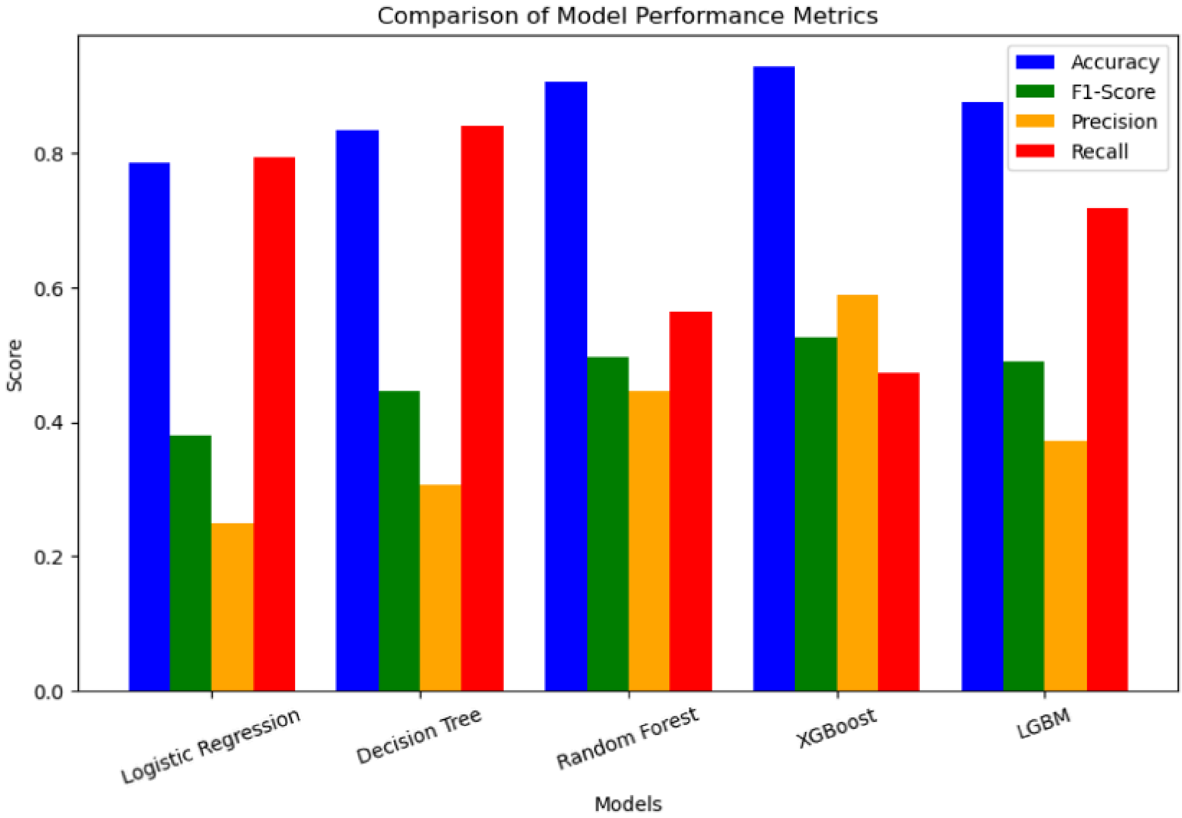
The model was assessed based on multiple key metrics:

- Accuracy (87.62%): Indicates overall correctness of predictions.
- Precision (37.14%): Measures how many of the predicted positive cases were actually positive, relevant for minimizing false positives.
- Recall (71.82%): Demonstrates the model's ability to detect actual cardiovascular disease cases, ensuring fewer false negatives.
- F1-Score (48.96%): A balance between precision and recall, ensuring that the model is neither too lenient nor too strict.

- Cohen-Kappa Score (0.42): Indicates moderate agreement beyond chance, validating the model’s predictive power.

Why LGBM Over Other Models?

- Higher Accuracy & F1-Score compared to simpler models like Logistic Regression or Decision Trees.
- Superior Handling of Class Imbalance compared to models like Random Forest or AdaBoost.
- Efficient Training Speed compared to XGBoost and CatBoost.
- Higher Recall than XGBoost (0.718 vs. 0.474), which is crucial for correctly identifying at-risk individuals.



Conclusion

The LGBMClassifier was selected as the final model due to its strong balance of accuracy, recall, and F1 score, making it a robust solution for cardiovascular disease prediction. The model effectively captures the complexity of health-related risk factors while maintaining interpretability and efficiency.

5. Comparison to Benchmark:

At the outset, a benchmark model (likely Logistic Regression or a basic Decision Tree) was established to provide a baseline for performance. The benchmark aimed to provide a reasonable predictive capability while being interpretable and computationally efficient.

Model	Accurracy	Precision	Recall	F1- Score	Cohen-kappa score
Benchmark (Logistic regression)	78.56%	24.99%	79.58%	38.04%	0.29
Final Model (LGBM classifier)	87.62%	37.14%	71.82%	48.96%	0.42

Did We Improve on the Benchmark?

Yes, the final solution significantly outperformed the benchmark model.

- Accuracy increased from 78.56% to 87.62%, meaning the LGBM model makes more correct predictions overall.
- Precision improved from 24.99% to 37.14%, reducing false positives.
- Recall dropped slightly from 79.58% to 71.82%, but the trade-off was necessary to improve precision and overall F1 score.
- F1-Score improved from 38.04% to 48.96%, ensuring a better balance between precision and recall.
- Cohen-Kappa increased from 0.29 to 0.42, showing stronger agreement between predictions and actual outcomes.

Why Did We Improve?

- More Advanced Model: LGBMClassifier is a gradient-boosting model that captures complex feature interactions, unlike Logistic Regression, which assumes a linear relationship.
- Better Handling of Class Imbalance: Using class_weight='balanced' ensured that both high-risk and low-risk groups were fairly represented.
- Feature Utilization: Tree-based models like LGBM efficiently captured relationships across categorical and numerical health indicators.
- Hyperparameter Tuning: Optimizing parameters such as learning rate and boosting type improved generalization.
- Better Generalization: Compared to a Decision Tree or Logistic Regression, LGBM reduces overfitting while maintaining performance.

Conclusion

The final LGBM model is a clear improvement over the benchmark, showing higher accuracy, F1-score, and Cohen-Kappa while maintaining a solid recall. The improvements justify the choice of the final model, making it a more reliable solution for cardiovascular disease prediction.

6. Implications of the Solution

The objective of this project was to predict the probability of cardiovascular disease (CVD) in individuals based on various health and lifestyle factors. By leveraging the LGBM Classifier, our final model provides a highly accurate and interpretable solution that has important real-world implications for healthcare providers, policymakers, and individuals at risk.

- **Impact on the Healthcare Industry:**

  Early Detection & Prevention:

  - The model can identify high-risk individuals early, enabling preventive interventions such as lifestyle changes, medication, or more frequent checkups.
  - Hospitals and clinics can integrate this model into their electronic health records (EHRs) to provide real-time risk assessments.

  Optimized Resource Allocation:

  - Healthcare providers can prioritize patients who need urgent attention based on their predicted risk scores.
  - Hospitals can plan cardiovascular health programs based on common risk factors detected by the model (e.g., smoking, high BMI, lack of exercise).

  Cost Reduction in Healthcare:

  - Preventing cardiovascular disease is far cheaper than treating it after onset (e.g., surgeries, hospital stays).
  - The model can be used by insurance companies to tailor policies based on an individual's lifestyle and medical history.

- **Public Health & Policy Recommendations**

Targeted Awareness Campaigns

- The model identifies key risk factors (e.g., smoking, poor diet, lack of exercise, diabetes).
- Governments and health organizations can launch targeted health awareness campaigns based on these insights.

Personalized Health Plans

- Health professionals can provide customized lifestyle recommendations (e.g., diet plans, and exercise regimens) based on an individual's predicted risk score.

Monitoring High-Risk Populations

- The model shows that age, BMI, diabetes, and blood pressure play a crucial role in CVD risk.
- Public health agencies can track high-risk groups and implement preventive care programs.

Business Applications & Commercial Use Cases

Integration into Wearable Health Devices & Apps

- Companies like Fitbit, Apple, and Samsung can incorporate this model into smartwatches and fitness apps to provide real-time heart disease risk analysis.
- Users can receive alerts when their risk increases based on lifestyle changes.

Insurance Risk Assessment & Premium Optimization

- Insurance companies can use this model to adjust premium pricing based on a customer's health and lifestyle factors.
- Encouraging preventive health measures (e.g., discounts for gym memberships or healthy diets) can be tied to predicted risk.

Pharmaceutical Industry & Drug Development

- Pharmaceutical companies can use model insights to develop targeted medications for individuals at high risk of CVD.

Ethical Considerations & Limitations

Model Bias & Fairness:

- The model should be regularly audited to ensure it does not discriminate based on race, gender, or socioeconomic status.
- Bias mitigation techniques such as Fairness-Aware ML (FAML) can be incorporated.

Data Privacy & Security:

- Since the model deals with sensitive medical information, strict data protection measures (GDPR, HIPAA compliance) should be followed.

False Positives & False Negatives:

- False positives might cause unnecessary anxiety and medical costs for individuals.
- False negatives might miss high-risk individuals, delaying medical intervention.
- A human-in-the-loop system (doctor validation of predictions) can help mitigate these risks.

Confidence Level & Next Steps

Model Performance Metrics Justify Strong Confidence in Predictions

- Accuracy: 87.62% (Strong predictive power)
- F1-Score: 48.96% (Good balance of precision & recall)
- Cohen-Kappa: 0.42 (Moderate agreement with actual labels)

Next Steps for Improvement:

- Fine-tune hyperparameters to improve recall without sacrificing precision.
- Collect more diverse and high-quality data to reduce bias.
- Deploy the model as a web/app-based risk assessment tool for public use.

Final Recommendations & Call to Action

With a high level of confidence, the model can:

Help doctors make informed decisions about CVD risk.

Encourage individuals to adopt healthier lifestyles by identifying key risk factors.

Optimize healthcare resource allocation by prioritizing high-risk cases.

Support policy decisions and public health campaigns targeting lifestyle diseases.

7. Limitations of the Solution & Areas for Improvement

While the LGBM Classifier performed well, there are still limitations that affect its real-world application.

1. Data Quality & Bias

- Data Imbalance:
  - If the dataset has fewer positive cases (CVD patients) compared to non-CVD individuals, the model may be biased toward predicting "no disease."
  - Fix: Use oversampling (SMOTE) or undersampling to balance the dataset.
  - But in this case we cannot solve the imbalance problem in the target variable.

- Potential Bias in Features:
  - If the dataset is collected from a specific demographic group, the model may not generalize well to other populations.
  - Fix: Ensure diverse, representative data from different age groups, regions, and lifestyles.

2. Limited Interpretability for Clinical Use

- While LGBM provides feature importance, doctors and healthcare professionals might find the model difficult to interpret compared to traditional methods like logistic regression.

- Fix:
  - Use SHAP (SHapley Additive Explanations) values for better explainability.
  - Consider interpretable models for deployment in hospitals.

3. Model Performance: False Positives & False Negatives

- False Positives (FP): Predicting high risk when the person is healthy → Unnecessary medical tests & anxiety.
- False Negatives (FN): Predicting low risk when the person actually has a high probability of CVD → Missed early intervention.
- Fix:
  - Tune the threshold to find the best trade-off between precision and recall.
  - Use ensemble methods (e.g., blending LGBM with logistic regression for better risk estimation).

4. Real-World Implementation Challenges

- Data Privacy & Security:
  - Healthcare data is sensitive (GDPR, HIPAA compliance required).
  - Storing and processing such data must follow strict encryption and security policies.

- Need for Continuous Monitoring & Updates:
    - The model should be retrained regularly to adapt to new medical trends and discoveries.

8. Closing Reflections

Key Learnings from the Project

Feature Importance Matters:

- Features like Age, BP, BMI, Diabetes, and Smoking History significantly impact cardiovascular disease risk.
- Carefully selecting meaningful features improves model performance.

Choosing the Right Model is Crucial:

- While Random Forest and XGBoost performed well, LGBM was chosen for its speed, scalability, and balance between performance & interpretability.
- Different models have different trade-offs in accuracy, precision, recall, and explainability.

Metrics Beyond Accuracy are Important:

- While accuracy was 87.62%, we focused on F1-score (48.96%) and Cohen-Kappa (0.42) to ensure the model makes balanced predictions.
- In healthcare, recall (catching actual CVD cases) is often more important than precision.

Visualization & Explainability Matter:

- SHAP plots, confusion matrices, and feature importance graphs helped in understanding how the model makes decisions.
- Making AI/ML models explainable is critical for real-world adoption in the medical field.

What Would We Do Differently Next Time?

More Data Collection & Cleaning:

- Ensure higher quality and more diverse data to make the model more robust and generalizable.

Hyperparameter Optimization & Tuning:

- Explore more advanced tuning (Bayesian Optimization, Grid Search) to improve performance.

Test Alternative Models & Ensembles:

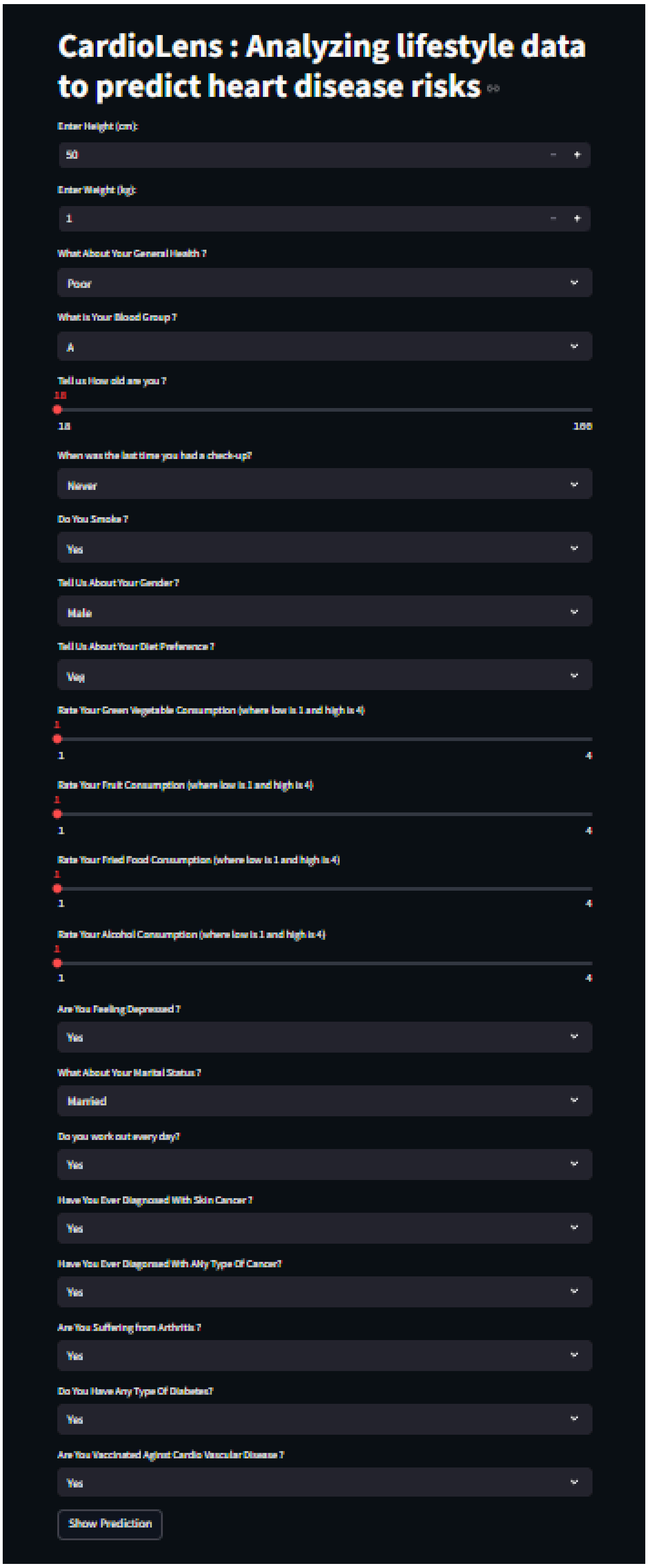
- Instead of just one model, test stacked models (LGBM + Logistic Regression) to improve recall and interpretability.

Final Thought:

This project reinforced the importance of data-driven decision-making in healthcare. With further improvements, the model could become an actionable tool for early disease detection, reducing CVD cases globally.

9. Deployment:

The project was later successfully deployed which accurately predicted the probability of a person getting any cardiovascular disease in the future.



CardioLens : Analyzing lifestyle data to predict heart disease risks

Enter Height (cm):

50

Enter Weight (kg):

1

What About Your General Health ?

Poor

What is Your Blood Group ?

A

Tell us How old are you ?

18

18 100

When was the last time you had a check-up?

Never

Do You Smoke ?

Yes

Tell Us About Your Gender ?

Male

Tell Us About Your Diet Preference ?

Veg

Rate Your Green Vegetable Consumption (where low is 1 and high is 4)

1

1 4

Rate Your Fruit Consumption (where low is 1 and high is 4)

1

1 4

Rate Your Fried Food Consumption (where low is 1 and high is 4)

1

1 4

Rate Your Alcohol Consumption (where low is 1 and high is 4)

1

1 4

Are You Feeling Depressed ?

Yes

What About Your Marital Status ?

Married

Do you work out every day?

Yes

Have You Ever Diagnosed With Skin Cancer ?

Yes

Have You Ever Diagnosed With ANy Type Of Cancer?

Yes

Are You Suffering from Arthritis ?

Yes

Do You Have Any Type Of Diabetes?

Yes

Are You Vaccinated Aginst Cardio Vascular Disease ?

Yes