Heart disease classification

PRESENTED BY:

GROUP NUMBER -12

SAI NARAYANA DAS MAKARAPU

NARENDRA GUDE

INTRODUCTION

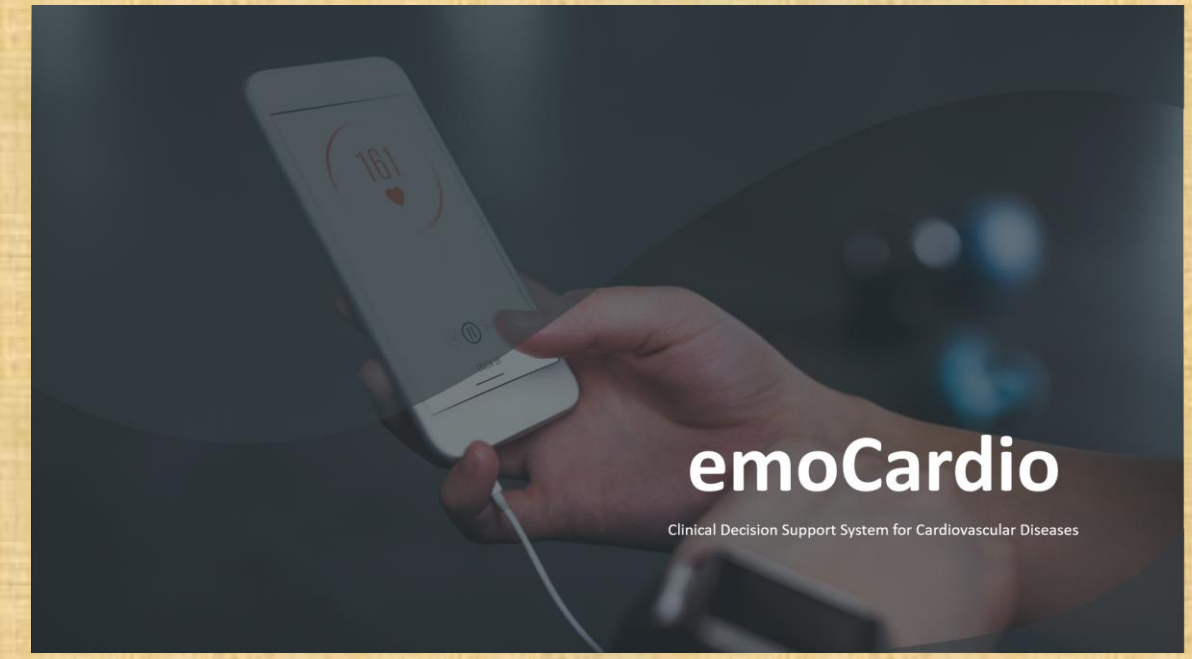
- Cardiovascular diseases are a group of disorders of the heart and blood vessels. It is a chronic condition that can cause very poor outcomes from COVID-19 and is on the rise globally.
- It is the leading cause of death globally, taking an estimated 17.9 million lives each year. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease.
- The **population selection** for cardiovascular disease is often based on the presence of risk factors such as high blood pressure, high cholesterol, smoking, obesity, diabetes, and family history of heart disease.
- The **measurement** of cardiovascular disease can vary depending on the specific condition being studied. Common measurements include blood pressure, cholesterol levels, body mass index (BMI), and blood glucose levels

Problem?

- Smoking? Obese? No Physical activity?
- Early prediction of CVD using risk factors is the main problem that needs to be solved.
- By predicting CVD using CDSS, it can help in the proactive management of cardiovascular health and the reduction of CVD-related risks.

Clinical Decision Support Systems (CDSS) can play a crucial role:

- Risk Stratification
- Complex data integration
- Early detection of risk factors
- Evidence-Based Decision support
- Personalized treatment plans
- Alerts for critical information
- Optimizing medication management
- Clinical Workflow integration



Cost:

1. A burden that contributes to most of the more than $320 billion in annual healthcare costs and lost productivity caused by cardiovascular disease.
2. In recent years, public funding for cardiovascular research has topped $2 billion annually.

Recent advancements:

1. Kwon et al. developed a DL-based algorithm combining a multilayer perceptron (MLP) and CNN, which aims to detect moderate or severe aortic stenosis (AS) using ECGs.
2. Shelly et al. developed a CNN model for screening moderate or severe AS using ECG and echocardiogram from 129,788 adult patients. AI-ECG performed well in the testing group, including 102,926 participants with an AUC of 0.85 and an accuracy of 74%, and the negative predictive value was 98.9%.
3. Attia et al. have implemented a CNN to identify patients with AF during normal sinus rhythm using standard 10 s, 12-lead ECG. The model was trained using nearly 500,000 ECGs. Applying the model in a testing set yielded an AUC of 0.87 for detecting AF from sinus-rhythm ECGs, with an overall accuracy of 79.4%. B

PEER-REVIEWED ARTICLES

- S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.
- K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.
- Krittanawong, C., Virk, H.U.H., Bangalore, S. *et al.* Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* **10**, 16057 (2020). https://doi.org/10.1038/s41598-020-72685-1
- Baghdadi, N.A., Farghaly Abdelaliem, S.M., Malki, A. *et al.* Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *J Big Data* **10**, 144 (2023). https://doi.org/10.1186/s40537-023-00817-1
- Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang,.A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data, Medical Engineering & Physics,Volume 105,2022,103825,ISSN 1350-4533, https://doi.org/10.1016/j.medengphy.2022.103825.
- Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. BMJ. 2020 Nov 4;371:m3919. doi: 10.1136/bmj.m3919. PMID: 33148619; PMCID: PMC7610202.
- RT Journal Article, A1 Allan, Simon,A1 Olaiya, Raphael, A1 Burhan, Rasan, T1 Reviewing the use and quality of machine learning in developing clinical prediction models for cardiovascular disease, JF Postgraduate Medical Journal,JO Postgrad. Med. J., YR 2021, DO 10.1136/postgradmedj-2020-139352, RD 11/27/2023. UL https://doi.org/10.1136/postgradmedj-2020-139352

Dataset

- Size: Dataset contains 319796 records, 18 variables (9 booleans, 5 strings and 4 decimals).
- Time Frame: The most recent dataset as of February 15, 2022 (includes data from 2020).
- Source: The dataset come from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents.

Data Preparation and Preprocessing:

- Import the dataset
- Exploratory Data Analysis
- There is no missing data
- Removed duplicate records of 18078 identical cases to avoid any overfitting in the model.
- Handling outliers

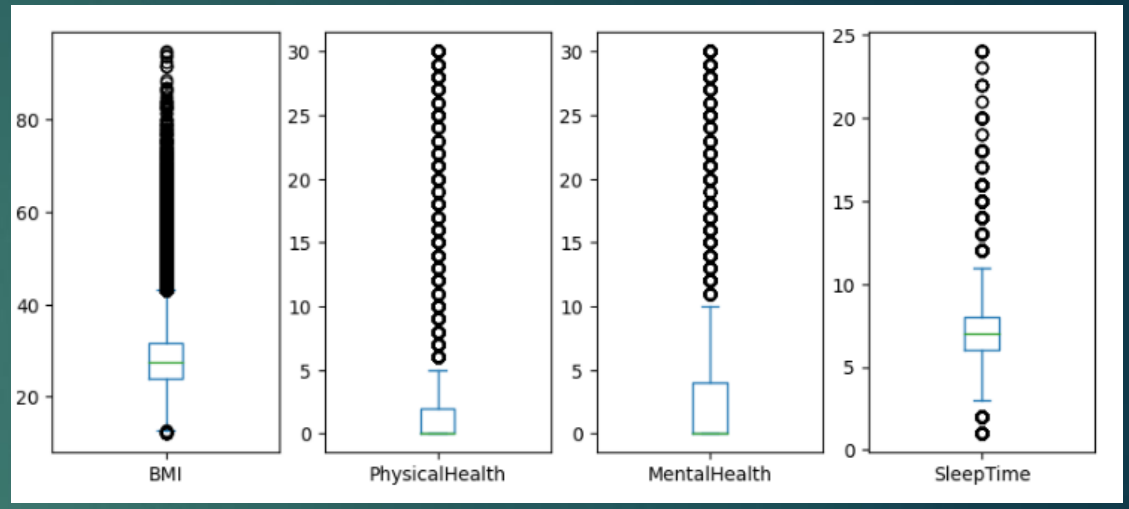
Handling outliers

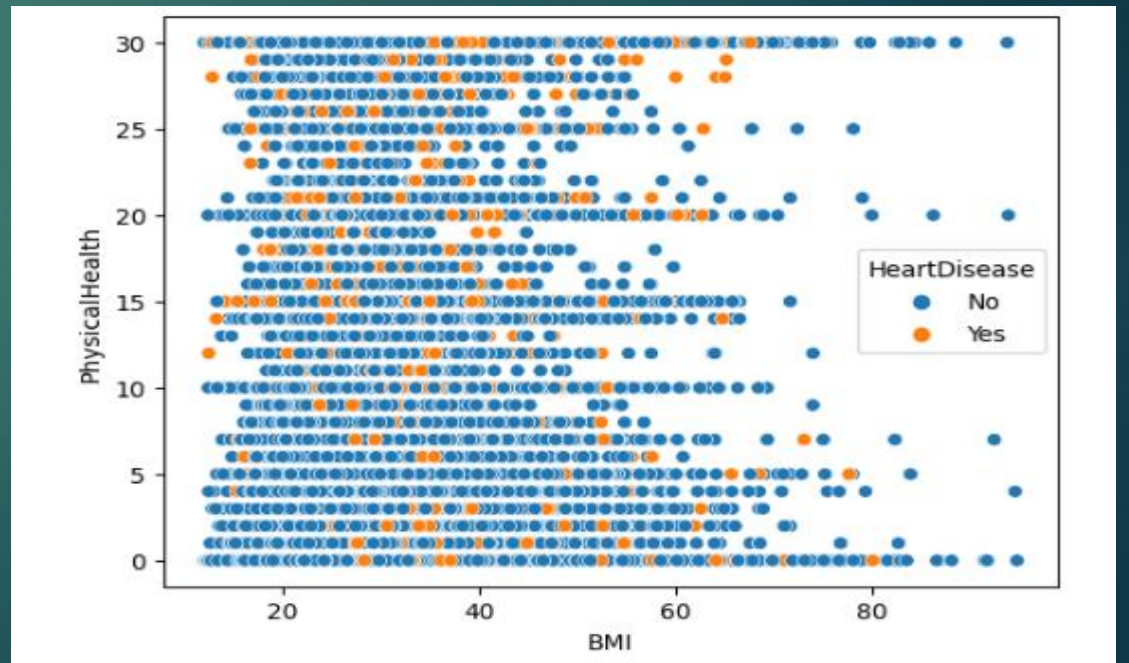
- Visual inspection:
  Box plot of BMI, Physical health, Mental health, SleepTime.
  Scatter plot of X-label = ‘BMI’, ylabel =‘PhysicalHealth’ with heartdisease.

- Need to get only the absolute values for a single comparison operation

- Used Z-score as a statistical tool to provide a standardized measure in identifying outlier.

- A 13% magnitude of outliers is substantial, making it impractical to remove them. Even if the percentage were lower and outlier removal were typically considered in normal cases, it is advisable to retain outliers, especially in the context of imbalanced data.





- **Feature Encoding**

  Initially encoded the categorical data using LabelEncoder

- **Feature scaling**

  As we have outliers in our data, we used the Robust scaler. Later, applied the MinMaxScaler specifically to the features while excluding the target variable.

  Reason: Using a robust scaler during the initial preprocessing stage helps maintain the integrity of the feature scaling. Dataset has been effectively treated for outliers, transitioning to a min-max scaler could be beneficial for normalization to a specific range.

Data Splitting

- Splitting the dataset into train and test sets:

  train_size = 0.7, test_size: 0.3, random_state=42
- Used stratified splitting technique to divide a dataset into training and testing subsets while preserving the class distribution of the target variable.
- This technique is useful when dealing with imbalanced datasets, where the number of instances belonging to different classes is significantly unequal.
- Goal of stratified splitting is to ensure that the training and testing sets have similar class proportions as the original dataset.
- Applying oversampling or undersampling techniques before the split can introduce data leakage, where the same observations end up in both sets.

ML models used

- Creates an instance of the **Logistic Regression model** and assigns it to the variable LR.
- Creates an instance of **the Random Forest Classifier model** with specific parameters (4 trees in the forest and a random state of 42) and assigns it to the variable RF.
- Creates an instance of the **Decision Tree Classifier model** with a specific random state (0) and assigns it to the variable DT.
- Creates an instance of the **K-Nearest Neighbors Classifier model** with a specified number of neighbors (100) and assigns it to the variable KNN.
- Creates an instance of **the XGBoost Classifier model** and assigns it to the variable XGB.

Results before applying any oversampling or undersampling techniques:

	LogR	RandF	DT	KNN	XGB
Precision	0.716365	0.605636	0.575108	0.750186	0.717305
Recall	0.543354	0.540273	0.583992	0.517122	0.546982
F1	0.556943	0.550171	0.579060	0.511173	0.562569
Accuracy	0.910182	0.896361	0.853374	0.910646	0.910270
Train_Score	0.910337	0.968864	0.997386	0.910881	0.917860
Test_Score	0.910182	0.896361	0.853374	0.910646	0.910270

1st Technique: "Over-Sampling" (Randomly):

Oversampling aims to balance class distribution by randomly increasing minority class examples by replicating them.

Apply the random oversampling technique, which is importing and using the ROSE (Random Oversampling Examples) method from the "imblearn" package

	LOGR	RANDF	DTS	KNN	XGB
Precision	0.593474	0.595988	0.571763	0.588717	0.596287
Recall	0.745168	0.564393	0.573449	0.742219	0.748947
F1	0.585943	0.574964	0.572586	0.571766	0.592010
Accuracy	0.732644	0.882838	0.857837	0.710747	0.740521
Train_Score	0.748170	0.996950	0.998563	0.771705	0.795300
Test_Score	0.732644	0.882838	0.857837	0.710747	0.740521

2nd Technique: "Under-Sampling" (Randomly):

Under-Sampling aims to balance class distribution by randomly delete majority class examples.

Apply the random undersampling technique, using the undersampling method from the "imblearn" package.

	LogR	RandF	DTs	KNN	XGB
Precision	0.593044	0.580346	0.558382	0.589058	0.593204
Recall	0.745338	0.689057	0.664127	0.742685	0.751725
F1	0.584457	0.581649	0.524229	0.572649	0.580088
Accuracy	0.730147	0.757093	0.666755	0.711996	0.719928
Train_Score	0.752817	0.940366	0.998638	0.754991	0.807708
Test_Score	0.730147	0.757093	0.666755	0.711996	0.719928

3rd Technique: "Over-Sampling" using SMOTE (Intelligently)

	LogR	RandF	DTs	KNN	XGB
Precision	0.592619	0.589083	0.563915	0.585385	0.633524
Recall	0.744725	0.573885	0.584339	0.740714	0.590272
F1	0.583525	0.580233	0.571047	0.559024	0.605565
Accuracy	0.728932	0.873658	0.835974	0.689977	0.890947
Train_Score	0.752392	0.981295	0.998558	0.782103	0.926670
Test_Score	0.728932	0.873658	0.835974	0.689977	0.890947

- One of the most used oversampling methods to solve the imbalance problem.
- SMOTE synthesises new minority instances between existing minority instances.
- It generates the virtual training records by linear interpolation for the minority class.
- These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class.
- After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

4th Technique: "Under-Sampling" using Tomek Links Technique (Intelligently):

- It is one of a modification from Condensed Nearest Neighbors (CNN).
- It can be used to find desired samples of data from the majority class that is having the lowest Euclidean distance with the minority class data and then remove it.

	LogR	RandF	DTs	KNN	XGB
Precision	0.704391	0.613852	0.579601	0.721279	0.700708
Recall	0.559891	0.558233	0.597564	0.531306	0.566581
F1	0.580856	0.572102	0.586801	0.537172	0.589539
Accuracy	0.909110	0.892770	0.848612	0.910303	0.908558
Train_Score	0.909189	0.971066	0.997291	0.909537	0.917944
Test_Score	0.909110	0.892770	0.848612	0.910303	0.908558

5th Technique: Combining both Over-Sampling "SMOTE" and Under-Sampling "Tomek Links" (Intelligently): Using "SMOTETomek"

- A combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance than only under-sampling the majority class.

The process of SMOTE-Tomek Links is as follows.

1.Start of SMOTE: choose random data from the minority class. Calculate the distance between the random data and its k nearest neighbors.

2. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.

3. Repeat step number 2–3 until the desired proportion of minority class is met (End of SMOTE).

4. Start of Tomek Links: choose random data from the majority class.

5. If the random data's nearest neighbor is the data from the minority class (i.e. create the Tomek Link), then remove the Tomek Link.

	LogR	RandF	DTs	KNN	XGB
Precision	0.592444	0.592686	0.567409	0.585185	0.634251
Recall	0.744858	0.578093	0.588965	0.740319	0.597743
F1	0.582854	0.584302	0.575015	0.558552	0.611620
Accuracy	0.727772	0.874000	0.837476	0.689359	0.889412
Train_Score	0.760334	0.985134	0.998518	0.789122	0.933338
Test_Score	0.727772	0.874000	0.837476	0.689359	0.889412

Metrics for evaluation

- Heart disease is a critical disease which may cost our patients their lives thus we want to take appropriate measures to detect it in order to be able to treat it properly.
- In cases of high risk prediction, we always count on the **"Recall"** which is the ability of a model to find all the relevant cases within a data set. The number of **true positives** divided by the number of **true positives plus the number of (false negatives)**.
- We want to avoid false negatives as much as possible.
- In most high-risk detection cases (like our example here: heat disease), recall is a more important evaluation metric than precision.

Conclusion

According to our results, we conclude that:

- The SMOTE technique, and the combination technique (SMOTETOMEK) were the best approaches.
- We found that both XGB and Logistic Regression had the best results, regarding both the accuracy and the recall (with about 90% accuracy and 75% recall).
- Our original data without any resampling, had the highest and most misleading accuracy as expected.

To achieve better results

- There is much more to do when it comes to Cross Validation and hyperparameter tuning.
- Also, should've encoded our feature with two different methods according to whether they are ranked or nominal, where we should've encoded only ranked data with the label encoder, while used one hot encoding with the other (nominal) data or features.

DATA

Strengths

- Large Sample size
- Diversity of data
- High Granularity
- Annual frequency
- Official source

Weaknesses

- Selection Bias
- Duplicate Data
- Data Imbalance

References

- Baghdadi, N.A., Farghaly Abdelaliem, S.M., Malki, A. *et al.* Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *J Big Data* **10**, 144 (2023). https://doi.org/10.1186/s40537-023-00817-1
- Aamir Javaid, Fawzi Zghyer, Chang Kim, Erin M. Spaulding, Nino Isakadze, Jie Ding, Daniel Kargillis, Yumin Gao, Faisal Rahman, Donald E. Brown, Suchi Saria, Seth S. Martin, Christopher M. Kramer, Roger S. Blumenthal, Francoise A. Marvel,medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology,American Journal of Preventive Cardiology, Volume 12,2022,100379,ISSN 266677, https://doi.org/10.1016/j.ajpc.2022.100379.
- https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
- Zhao, Y., Wood, E. P., Mirin, N., Cook, S. H., & Chunara, R. (2021). Social determinants in machine learning cardiovascular disease prediction models: a systematic review. American journal of preventive medicine, 61(4), 596-605.
- Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
- Elreedy, Dina, and Amir F. Atiya. "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance." Information Sciences 505 (2019): 32-64.


THANK YOU