Capstone Project - 3

Cardiovascular Risk Prediction

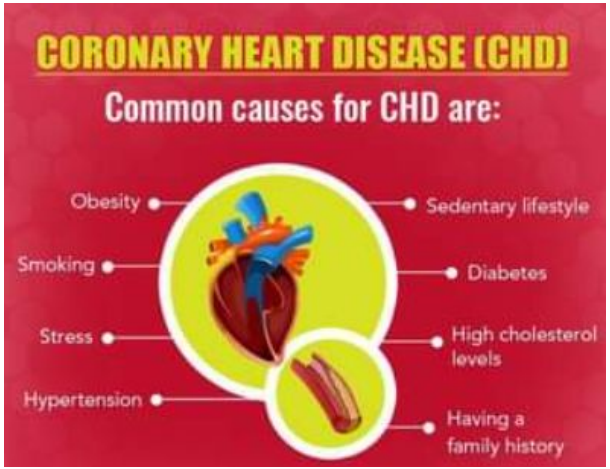
Team Members

Akshada

Nikita

Introduction

We'll be working on a cardiovascular risk prediction, where we'll go through the entire EDA process with data visualization, and model implementation to predict whether the patient has a 10-year risk of future coronary heart disease (CHD), based on factors like heart rate, cholesterol & blood pressure etc. This is a classification problem. Cardiovascular disease (CVD) is the biggest concern in the medical sector at present. It is one of the most lethal and chronic diseases, leading to the highest number of deaths worldwide.



Objective

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are demographic, behavioral, and medical risk factors.

In order to achieve this, the supervised machine learning model was developed using classification algorithms. The main objective is to understand which algorithm provides a better result in predicting whether the patient has a 10 year risk of future coronary heart disease.

Coronary Heart Disease(CHD)

Cardiovascular disease is the group of heart and blood vessels disorder that includes coronary heart disease. CHD is a kind of heart disease where the arteries of the heart cannot supply enough oxygen-rich blood to the heart. Chest pain, breathlessness, pain throughout the body, and feeling faint and sick are the main symptoms of CHD. But not everyone has similar symptoms and some people may not have any before coronary heart disease is diagnosed.

The deposit of fatty substance in the coronary arteries causes CHD, that block or interrupt the heart's blood supply. Over time, the walls of the artery can become furred up with fatty deposits. It can be caused due to lifestyle factors, such as smoking and regularly drinking excessive amounts of alcohol. A person is also more at risk if he/she has conditions like high cholesterol, high blood pressure (hypertension), and diabetes.

Variable Breakdown

- id: Patient identification number

Demographic:

- age: Age of the patient
- education: no further information provided
- sex: male or female

Behavioral:

- is_smoking: whether or not the patient is a current smoker
- cigsPerDay: the number of cigarettes that the person smoked on average in one day

Variable Breakdown (Continued)

Information on medical history:

- **BPMeds: whether or not the patient was on blood pressure medication**
- **prevalentStroke: whether or not the patient previously had a stroke**
- **prevalentHyp: whether or not the patient was hypertensive**
- **diabetes: whether or not the patient had diabetes**

Information on current medical condition:

- **totChol: total cholesterol level**
- **sysBP: systolic blood pressure**

Variable Breakdown (Continued)

- **diaBP: diastolic blood pressure**
- **BMI: Body Mass Index**
- **heartRate : heart rate**
- **glucose: glucose level**

Target variable to predict:

- **TenYearCHD: 10 year risk of coronary heart disease (CHD), binary variable where “1” means Yes and “0” means No.**

Steps involved in ML

Step 1: Exploratory Data Analysis	Step 2: Feature Selection	Step 3: Data Preprocessing
We performed exploratory data analysis on the cardiovascular risk dataset, where we looked at descriptive statistics, treated for null or missing values, handled imbalanced data and performed label encoding for categorical variables in the given dataset.	It is the process of reducing the number of input features when developing a predictive model. We decided to remove the four least important features in our dataset.	Here we split the given data into training and validation sets so that we could learn the model's parameters and evaluate the model's performance.

Steps involved in ML(Continued)

Step 4: Model Implementation	Step 5: Hyperparameter finetuning using GridSearchCV on selected models	Step 6: Comparing models performance
In the model implementation, out of eight different models, three best performing models were selected to be trained.	The selection of hyperparameters involves testing the performance of the model against different combinations of hyperparameters, and the best combinations are then selected on the basis of chosen metric and validation method.	The performances of selected models and stacking on selected models were compared to find the best model for the given dataset.

Descriptive statistics

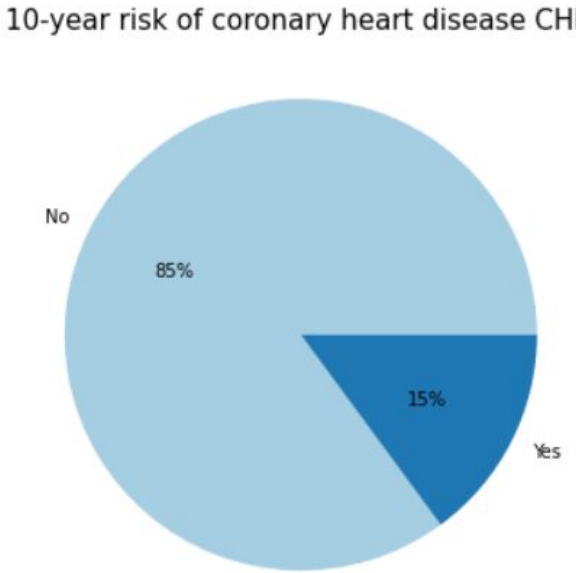
- As the categorical data of the dataset were already converted into discrete numerical values therefore they are also included in descriptive statistics.
- The age of the patients varies from 32 to 70.
- Only 25% of patients in the given dataset are consuming cigarettes on an average of at least 20 cigarettes per day.

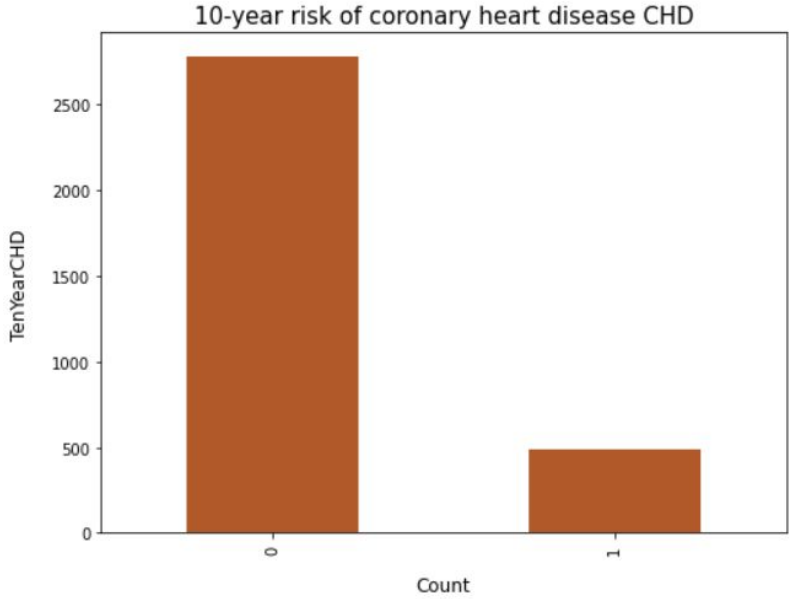
	count	mean	std	min	25%	50%	75%	max
age	3272.0	49.442237	8.559199	32.00	42.00	49.00	56.000	70.0
sex	3272.0	0.438570	0.496288	0.00	0.00	0.00	1.000	1.0
is_smoking	3272.0	0.493276	0.500031	0.00	0.00	0.00	1.000	1.0
cigsPerDay	3272.0	9.054095	11.866441	0.00	0.00	0.00	20.000	70.0
BPMeds	3272.0	0.029645	0.169633	0.00	0.00	0.00	0.000	1.0
prevalentStroke	3272.0	0.005807	0.075993	0.00	0.00	0.00	0.000	1.0
prevalentHyp	3272.0	0.313264	0.463892	0.00	0.00	0.00	1.000	1.0
diabetes	3272.0	0.025672	0.158180	0.00	0.00	0.00	0.000	1.0
totChol	3272.0	237.036675	45.114324	113.00	206.00	234.00	264.000	696.0
sysBP	3272.0	132.409077	22.123491	83.50	117.00	128.00	143.625	295.0
diaBP	3272.0	82.864456	11.952790	48.00	74.50	82.00	90.000	142.5
BMI	3272.0	25.800339	4.123262	15.96	23.01	25.39	28.040	56.8
heartRate	3272.0	75.955990	12.024657	45.00	68.00	75.00	83.000	143.0
glucose	3272.0	81.420232	23.195784	40.00	72.00	77.00	85.000	394.0
TenYearCHD	3272.0	0.149144	0.356285	0.00	0.00	0.00	0.000	1.0

Descriptive statistics (Continued)

- The average total cholesterol level of the patients in this dataset is 237.03.
- The minimum and maximum systolic blood pressure of the patients are 83.5 and 295, respectively.
- The minimum and maximum diastolic blood pressure of the patients are 48 and 142.5, respectively.
- The minimum and maximum body mass index of the patients are 15.96 and 56.8, respectively.
- The average heart rate of the patients in this dataset is 75.95.
- The average glucose level of the patients in this dataset is 81.42.

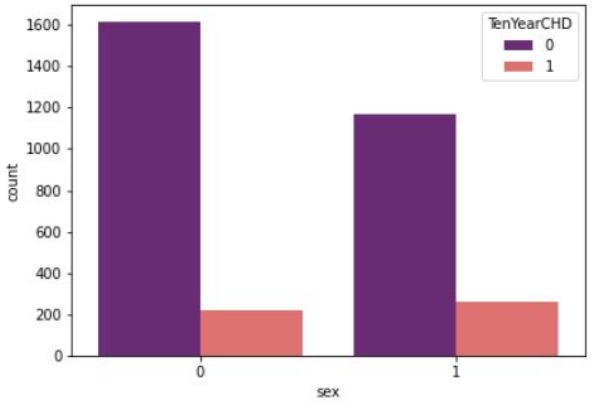
Checking the 10-year risk of CHD

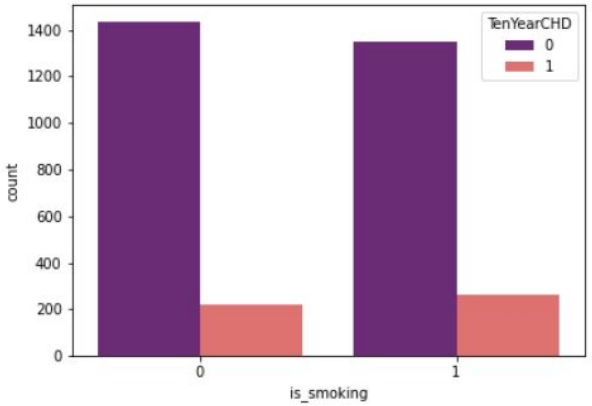


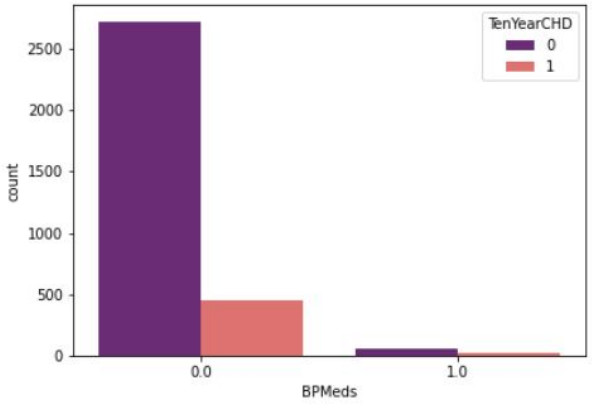


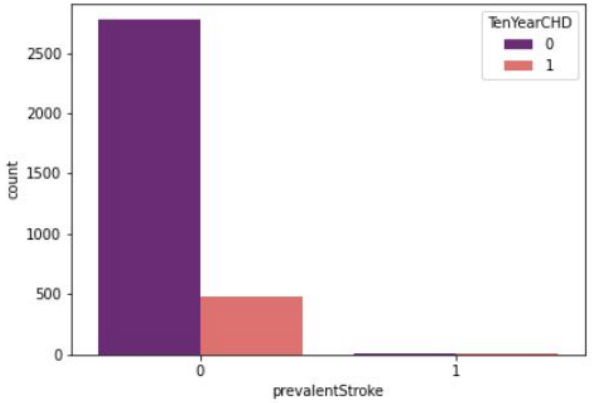
- From above observation, we can conclude that our dataset is not balanced i.e. Yes is 488 (~15%) and No is 2784 (~86%). Analysis shows that ten year risk of future coronary heart disease (CHD) for patients in the given dataset is 15%.

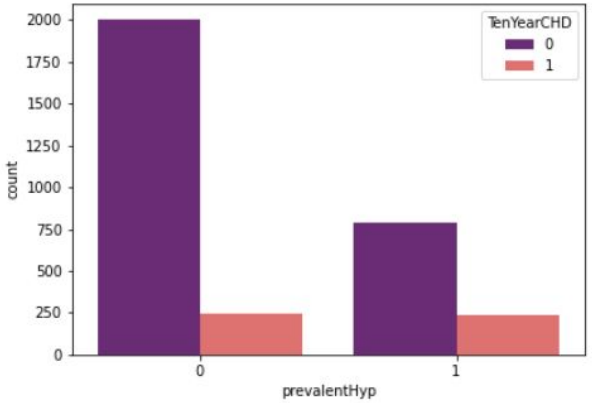
10-year risk of CHD according to categorical variables

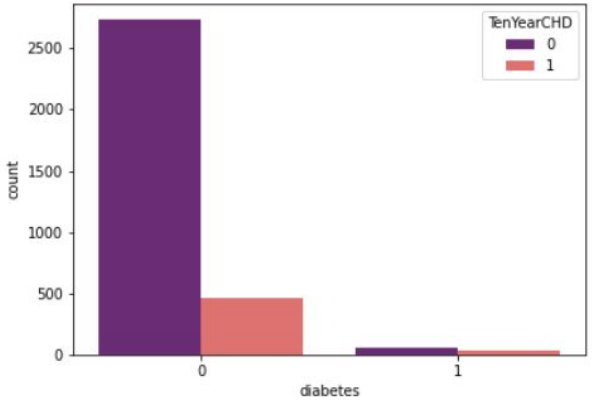








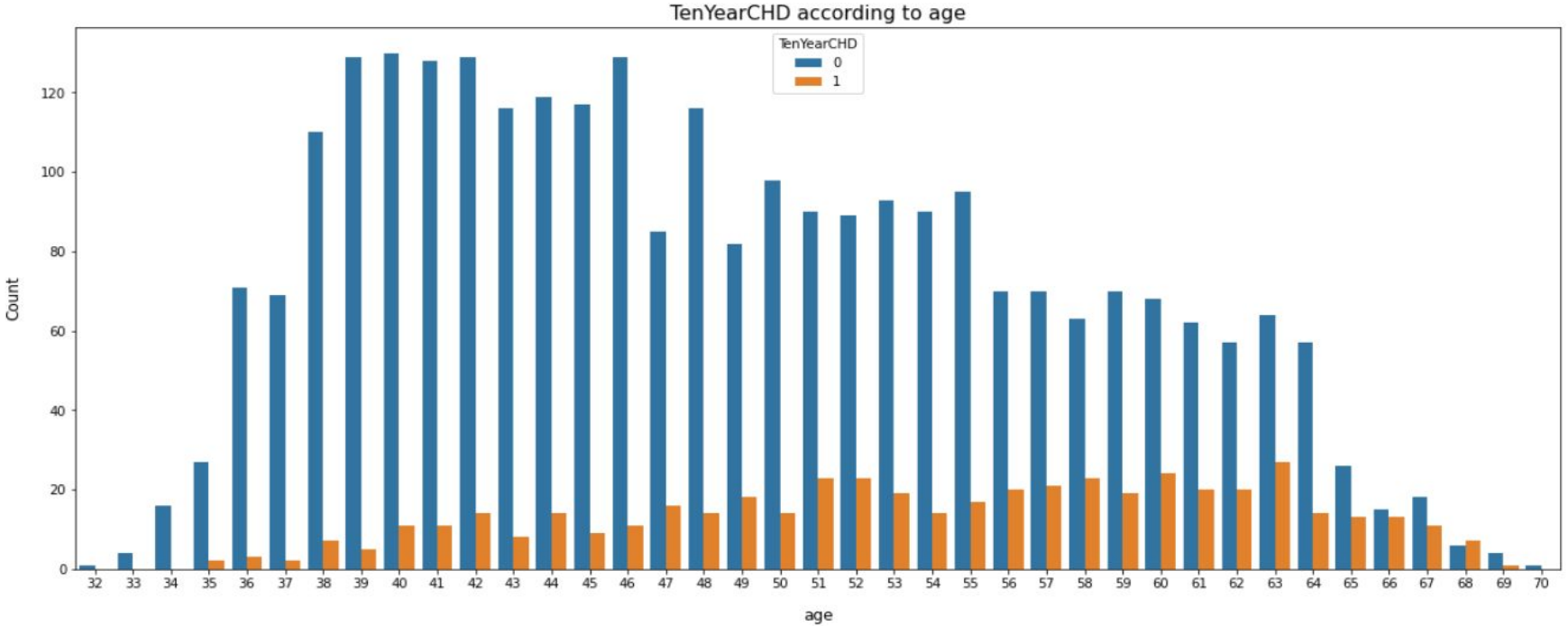




10-year risk of CHD according to categorical variables(Continued)

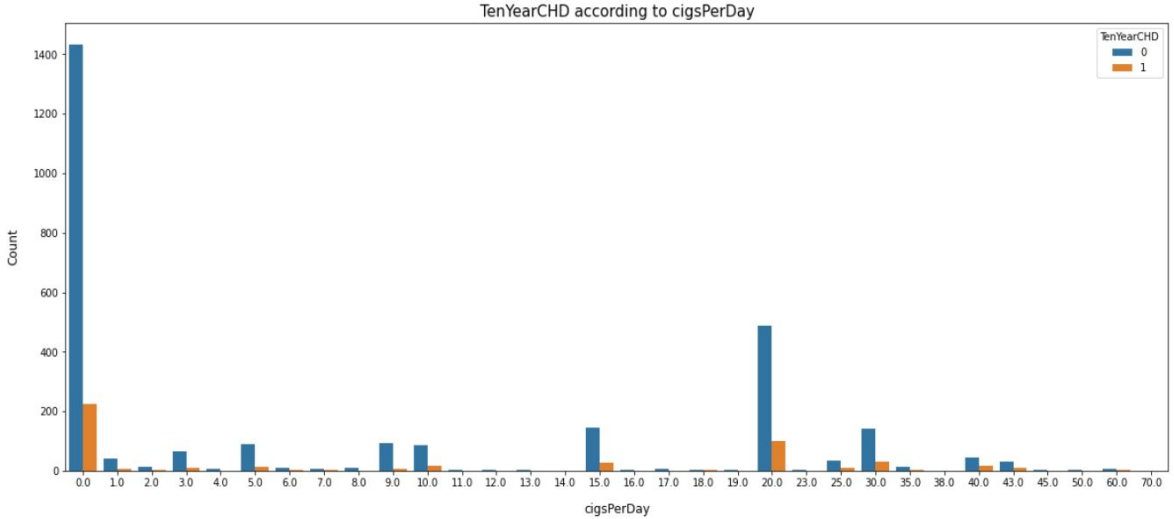
- From count plots, we can observe that BPmeds, prevalentStroke, and diabetes are highly imbalanced.
- The number of females not affected by CHD is more than the number of males.
- The number of current smokers and non-smokers is almost the same.
- Non-hypertensive patients not affected by CHD are more than hypertensive patients.
- Female patients are less prone to the 10-year risk of future coronary heart disease (CHD).
- Patients taking blood pressure medication are having low risk of CHD as compared to those who are not taking medication.

10-year risk of CHD according to numerical variables



- **Age groups ranging from 51 to 63 are more affected by CHD and age groups ranging from 32 to 37 are less affected by CHD.**

10-year risk of CHD according to numerical variables(Continued)

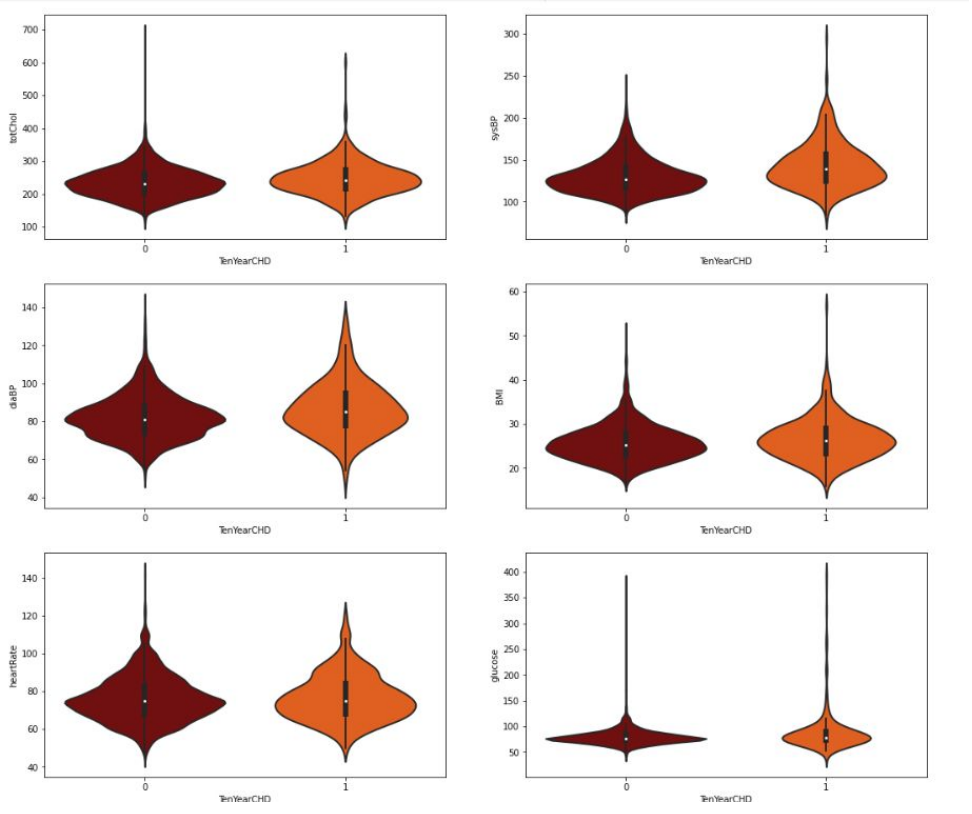


- On average, patients who do not smoke daily are more affected by CHD. On the contrary, these patients are least affected by CHD.
- Patients who smoke 20 cigarettes daily on average are more affected by CHD.
- On average, patients who do not consume cigarette daily are at high risk of CHD, indicating that there are factors other than smoking that causes CHD.

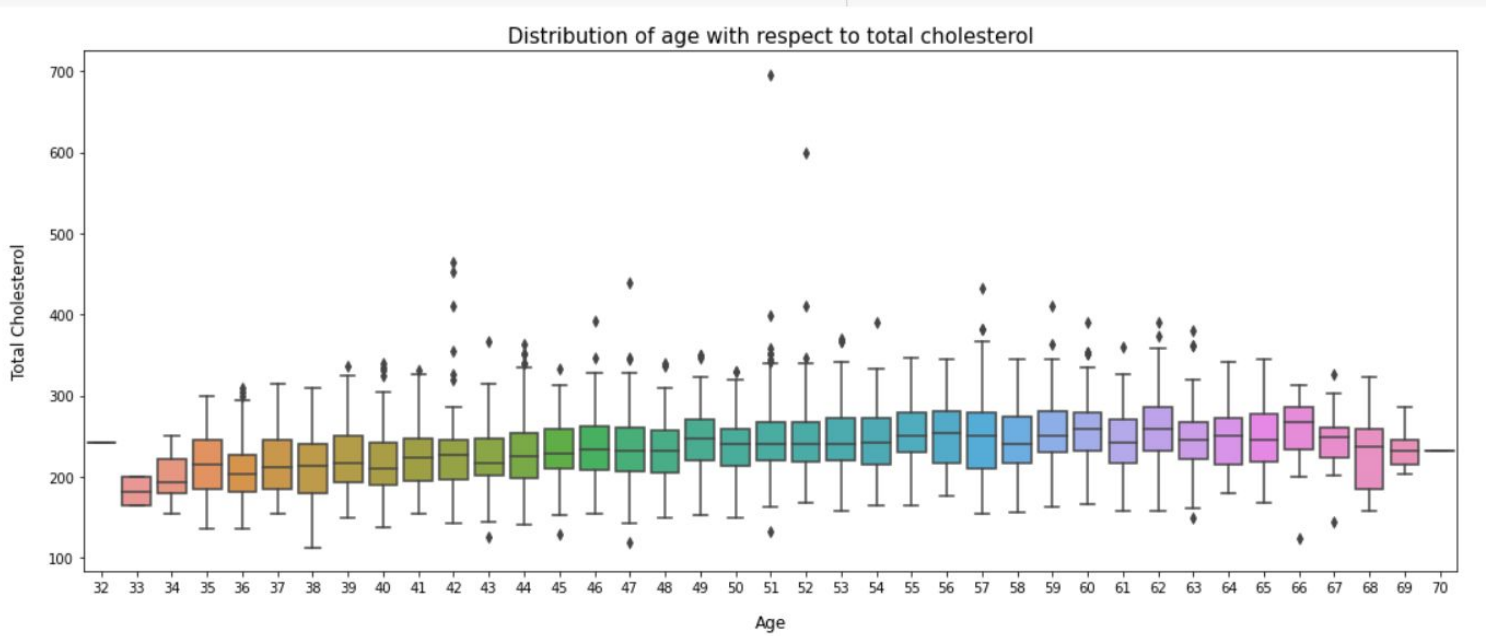
10-year risk of CHD according to numerical variables(Continued)

The majority portion of numerical variables affected by the target variable lie in the following ranges:

- **totChol : 150 to 350**
- **sysBP : 100 to 175**
- **diaBP : 65 to 110**
- **BMI : 17 to 32**
- **heartRate : 55 to 100**
- **glucose : 50 to 110**



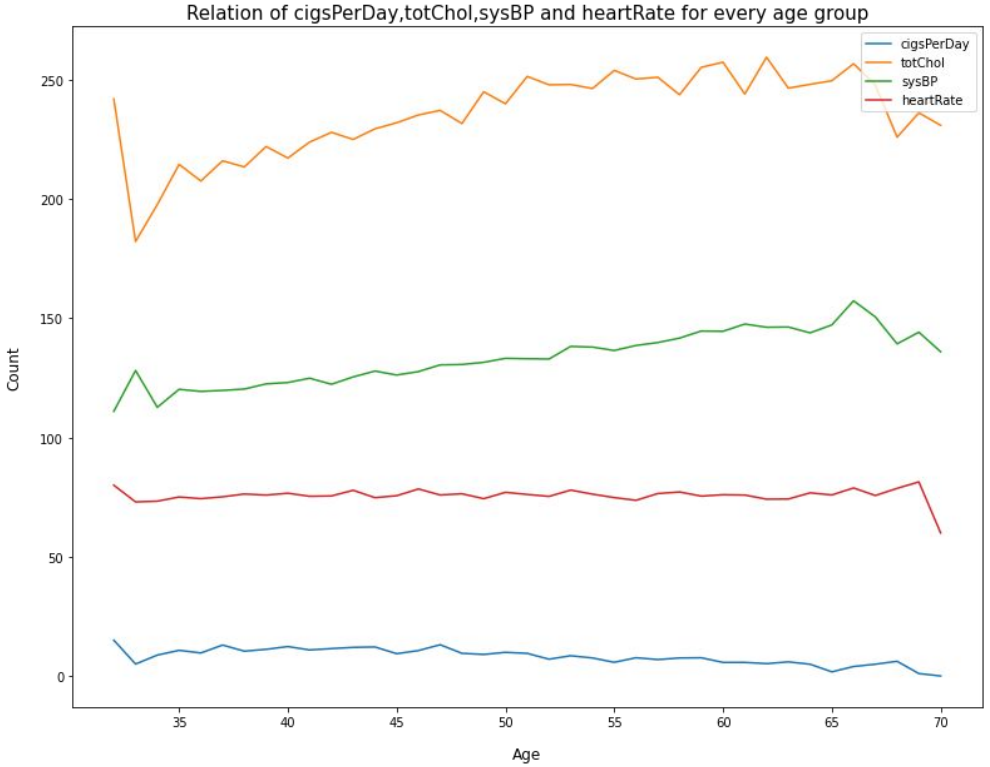
Relation between age and total cholesterol



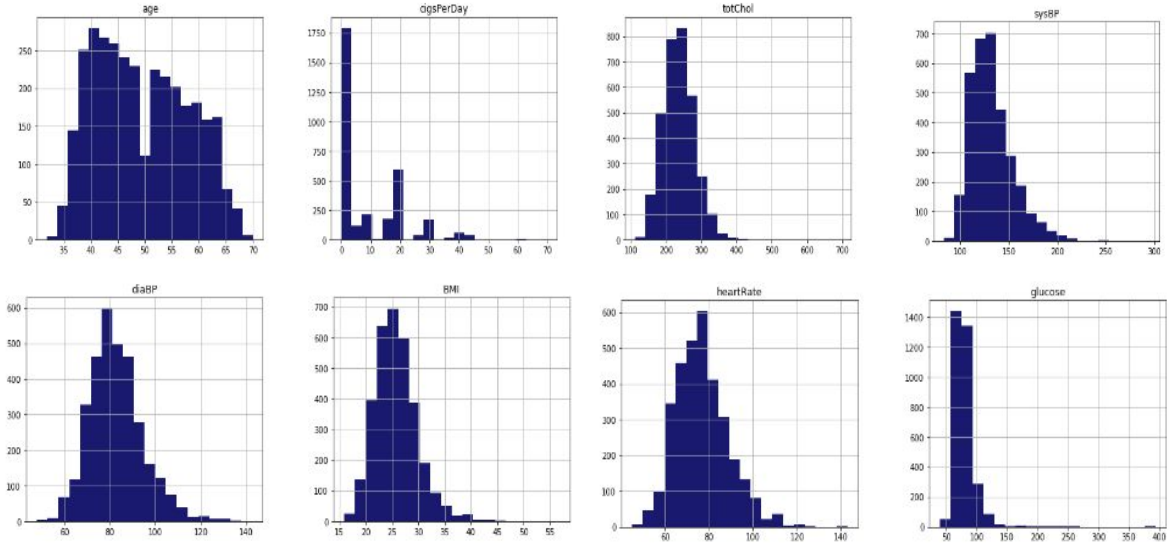
- From the above observation, we can infer that the boxplots are shifted in an upward manner representing that cholesterol level increases with the increase in age.

Relation among cigsPerDay, totChol, sysBP, and heartRate based on age.

- The variable cigsPerDay and heartRate have a fairly parallel relationship with age but slightly decrease with the increase of age.
- Total cholesterol level and systolic blood pressure of the patient are moving in an increasing manner w.r.t. age.



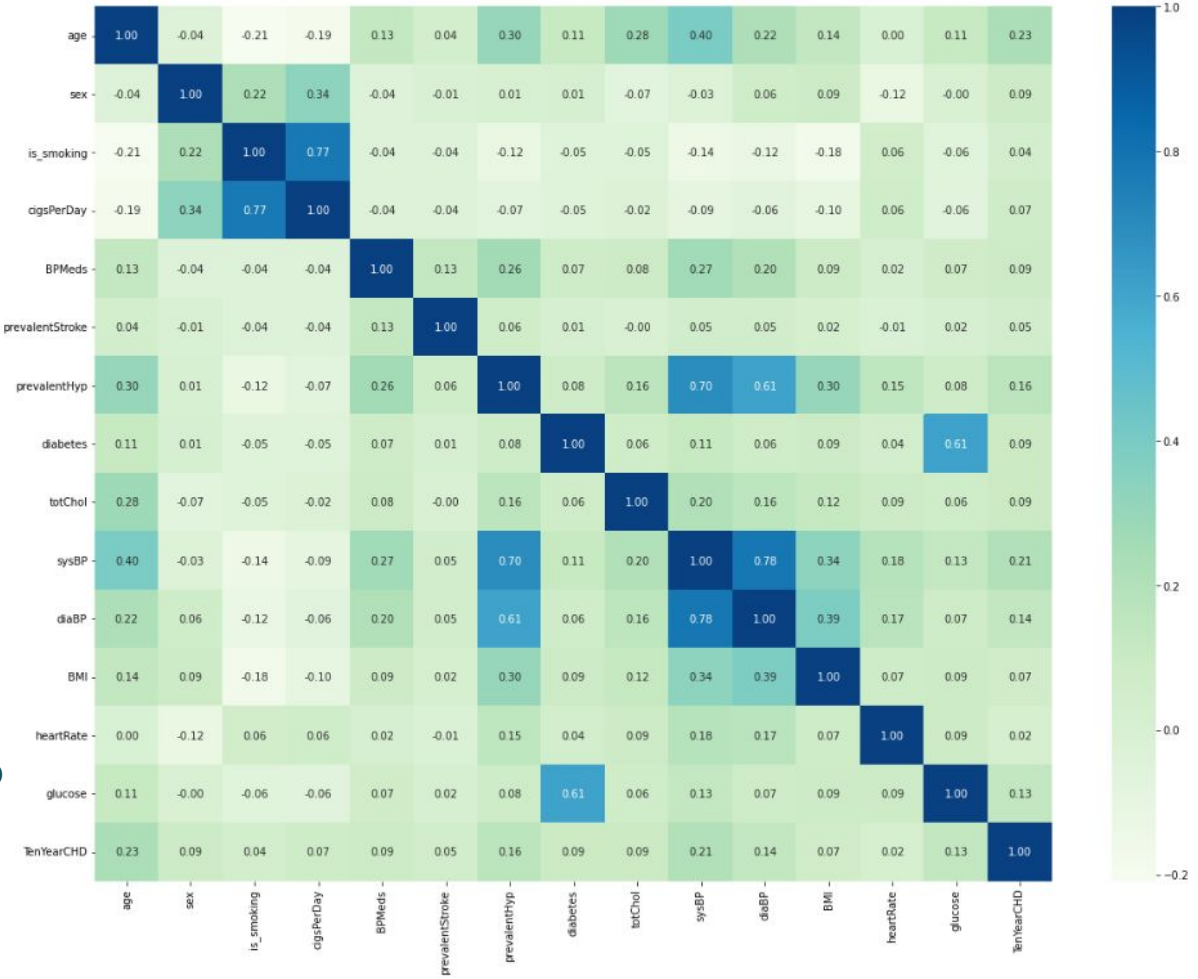
Frequency distribution of numerical variables using histogram



- All numerical variables follows almost normal distribution except 'age', 'cigsPerDay', and 'glucose'.
- The variables 'age', 'cigsPerDay', and 'glucose' shows irregular distribution.

Heat map

- All variables are positively correlated with the target variable.
- The variables cigsPerDay and is_smoking are more positively correlated.
- The variables sysBP and diaBP are more positively correlated.
- The variable prevalentHyp is more correlated with sysBP.



Model Implementation

The following algorithms will be evaluated:

- ❑ **Logistic Regression**
- ❑ **Support Vector Machines**
- ❑ **KNeighbors Classifier**
- ❑ **Decision Tree Classifier**
- ❑ **Random Forest Classifier**
- ❑ **Gradient Boosting Classifier**
- ❑ **Extra Trees Classifier**
- ❑ **XGB (Extreme Gradient Boosting)**

Model Implementation(Continued)

	Name	Train Time	Accuracy	Precision	Recall	f1 Score	ROC-AUC Score
0	Logistic Regression:	0.027925	0.678636	0.673684	0.690647	0.682060	0.740727
1	SVC:	5.384045	0.734291	0.715947	0.775180	0.744387	0.821320
2	KNeighbors Classifier:	0.009378	0.783662	0.722772	0.919065	0.809184	0.878948
3	Decision Tree Classifier:	0.035230	0.815978	0.796954	0.847122	0.821273	0.816034
4	Random Forest:	0.767087	0.903950	0.901610	0.906475	0.904036	0.968980
5	Gradient Boosting:	1.005097	0.823160	0.840607	0.796763	0.818098	0.904850
6	Extra Trees Classifier:	0.480533	0.921005	0.914894	0.928058	0.921429	0.977260
7	XGB Classifier:	0.322304	0.807899	0.820225	0.787770	0.803670	0.901690

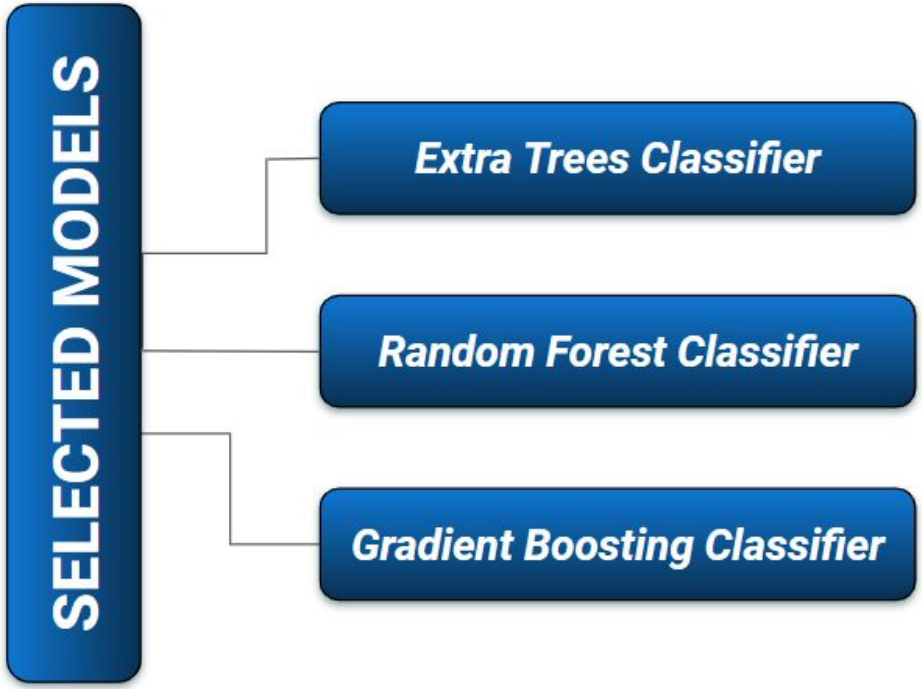
On the basis of validation set performance:

- **The KNeighbors classifier and SVC take the least and most training time, respectively.**
- **Extra trees classifier and logistic regression show the highest and lowest performance, respectively.**
- **Ensemble models are performing well on the validation set.**

Best performing models

Out of 8 different models, three best performing models were selected to be trained using hyperparameter tuning with GridSearchCV.

- Extra Trees Classifier
- Random Forest Classifier
- Gradient Boosting Classifier



Extra-Trees Classifier

The extremely randomized (or extra) trees classifier is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its classification result. After hyperparameter finetuning using GridSearchCV, we observed:

1. Best tuned parameters are 'max_depth' = 50, 'max_features' = 'auto', 'n_estimators' = 112.
2. For this model, the accuracy score is 0.92.
3. The precision and recall for a patient having a 10-year risk of coronary heart disease (CHD) are 0.92 and 0.93, respectively, whereas for non-CHD patients are 0.93 and 0.92, respectively.
4. The f1-score for a patient having a 10-year risk of coronary heart disease (CHD) and for a non-CHD patient is 0.92.
5. The data tested for the CHD patient is 556 and for the non-CHD patient is 558.

Random Forest Classifier

A random forest is an ensemble method capable of performing both classification and regression tasks with the use of multiple decision trees and a technique called bagging. After hyperparameter finetuning using GridSearchCV, we observed:

1. Best tuned parameters are 'max_depth' = 50, 'max_features' = 'auto', 'n_estimators' = 92.
2. For this model, the accuracy score is 0.91.
3. The precision and recall for a patient having a 10-year risk of coronary heart disease (CHD) is 0.91, whereas for non-CHD patients it is 0.91.
4. The f1-score for a patient having a 10-year risk of coronary heart disease (CHD) and for a non-CHD patient is 0.91.
5. The data tested for the CHD patient is 556 and for the non-CHD patient is 558.

Gradient Boosting Classifier

Gradient boosting classifier is a group of machine learning algorithms that combine many weak learning models together to build a strong predictive model. After hyperparameter finetuning using GridSearchCV, we observed:

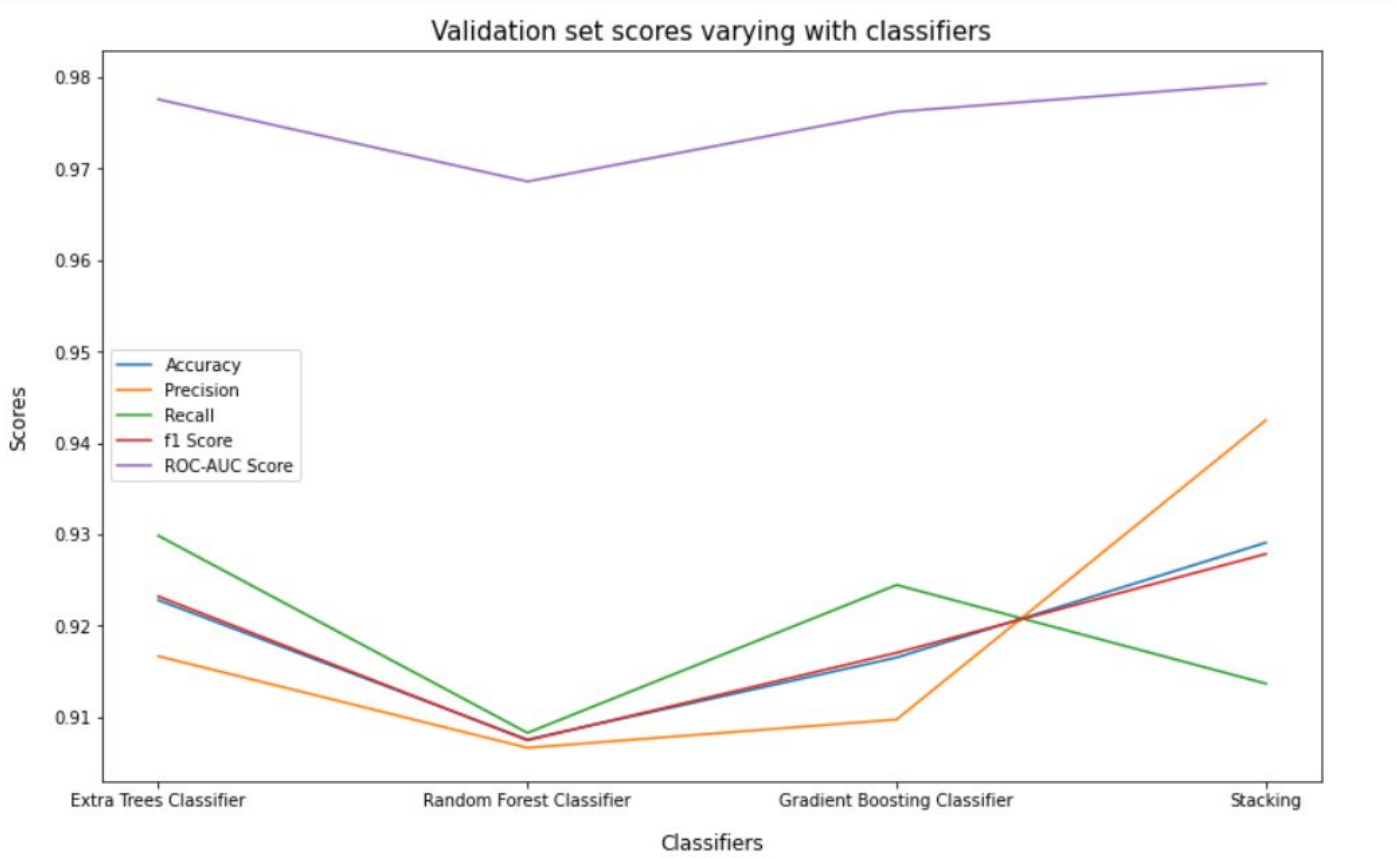
1. Best tuned parameters are 'max_depth' = 50, 'max_features' = 'sqrt', 'n_estimators' = 100.
2. For this model, the accuracy score is 0.92.
3. The precision and recall for a patient having a 10-year risk of coronary heart disease (CHD) are 0.91 and 0.92, respectively, whereas for non-CHD patients are 0.92 and 0.91, respectively.
4. The f1-score for a patient having a 10-year risk of coronary heart disease (CHD) and for a non-CHD patient is 0.92.
5. The data tested for the CHD patient is 556 and for the non-CHD patient is 558.

Stacking

Stacking allows the users to train multiple models to solve similar problems, and based on their combined results, it builds a new model with improved performance. For further model improvement we performed stacking on selected models.

1. For this model, the accuracy score is 0.93.
2. The precision and recall for a patient having a 10-year risk of coronary heart disease (CHD) are 0.94 and 0.91, respectively, whereas for non-CHD patients are 0.92 and 0.94, respectively.
3. The f1-score for a patient having a 10-year risk of coronary heart disease (CHD) and for a non-CHD patient is 0.93.
4. The data tested for the CHD patient is 556 and for the non-CHD patient is 558.

Comparing models performance



Comparing models performance (Continued)

	Classifier	Accuracy	Precision	Recall	f1_score	Roc_Auc_Score	Best Hyperparameters
0	Extra Trees Classifier	0.922801	0.916667	0.929856	0.923214	0.977602	{'max_depth': 50, 'max_features': 'auto', 'n_e...
1	Random Forest Classifier	0.907540	0.906643	0.908273	0.907457	0.968607	{'max_depth': 50, 'max_features': 'auto', 'n_e...
2	Gradient Boosting Classifier	0.916517	0.909735	0.924460	0.917038	0.976242	{'max_depth': 50, 'max_features': 'sqrt', 'n_e...
3	Stacking	0.929084	0.942486	0.913669	0.927854	0.979334	NA

- **Accuracy, precision, f1 score, and ROC-AUC score are high for stacking.**
- **The recall is high for the extra trees classifier.**

Why it is important to predict CVD risk?

- Cardiovascular disease has the highest death rate in the world due to its lethal and chronic nature.
- In 2019, an estimated 17.9 million people died from CVDs, accounting for 32% of all global deaths. 85% of these deaths were caused by heart attack and stroke.
- By addressing behavioral risk factors, it is possible to prevent most cardiovascular diseases like tobacco use, poor diet, obesity, physical inactivity, and excessive use of alcohol.
- It is critical to detect the disease as early as possible so that management with counselling and medicine can begin. Its early prediction can help people to change their lifestyles.

Conclusion

- Out of 3 selected models, the extra trees classifier was found to be the best performing.
- Further feature reduction in any of the selected models didn't add to the scores.
- Stacking, on the other hand, surpassed the performance of selected models.
- The final model had 11 features.
- The best algorithm to use for this dataset is stacking, as the best results for the validation set are given by stacking with an accuracy of 0.93 and a ROC-AUC score of 0.98.
- Stacking predicted that 539/1114 patients have a 10-year risk of future CHD and 575/1114 patients don’t. The precision, recall, and f1-score for a patient having a 10-year risk of future CHD are 0.94, 0.91, and 0.93, respectively, whereas for non-CHD patients are 0.92, 0.94, and 0.93, respectively.


THANK YOU