MINOR DATA DRIVEN DECISION MAKING IN BUSINESS

DATA MINING REPORT

HAN_UNIVERSITY
OF APPLIED SCIENCES

Gifty Mensah 1659945

Giang Lieu 1662109

Lecturer: Witek Ten Hove

Code: DATDRD05

Contents

Introduction

As we are both international student specializing in different fields, one being organization and change and the other being Marketing and sales, we tried to look for data set that matched either one of our specialization or interest we both had in common however, most of the data we selected were either too big or too small, while some had too much text (categorical data) and few numerical value thus we couldn't use any of them. We both came to conclusion to use heart disease data set because it was intriguing, and all of the data were numerical value making it easy to measure our data and make comparison

The data we decided to move forward with is 'Heart disease'. The heart disease dataset provided on Kaggle contains various attributes related to heart health and disease. Understanding the features in the dataset gave us insights into factors associated with heart disease and help in predicting the likelihood of someone having heart disease. Here's an overview of what the dataset typically includes:

- Age: The age of the patient.
- Sex: The gender of the patient (0 for female, 1 for male).
- Chest Pain Type (CP): The type of chest pain experienced by the patient. This feature may include categories such as typical angina, atypical angina, non-anginal pain, or asymptomatic.
- Resting Blood Pressure (Trestbps): The resting blood pressure of the patient in mm Hg.
- Serum Cholesterol (Chol): The serum cholesterol level of the patient in mg/dl. Fasting Blood
- Sugar (Fbs): Indicates whether the fasting blood sugar level of the patient is greater than 120 mg/dl (1 for true, 0 for false).
- Resting Electrocardiographic Results (Restecg): The resting electrocardiographic results of the patient, which can indicate normal, abnormal ST-T wave, or left ventricular hypertrophy.
- Maximum Heart Rate Achieved (Thalach): The maximum heart rate achieved by the patient.
- Exercise-Induced Angina (Exang): Indicates whether the patient experienced exercise-induced angina (1 for yes, 0 for no).
- ST Depression Induced by Exercise Relative to Rest (Oldpeak): ST depression induced by exercise relative to rest, a measure that could indicate coronary insufficiency.
- Slope of the Peak Exercise ST Segment (Slope): The slope of the peak exercise ST segment, which could indicate the severity of coronary artery disease.
- Number of Major Vessels (Ca): The number of major vessels colored by fluoroscopy.
- Thalassemia (Thal): A blood disorder that affects the production of hemoglobin. This feature may include categories such as normal, fixed defect, or reversible defect.
- Target: Indicates whether the patient has heart disease (1 for presence, 0 for absence).

By analyzing this dataset, we gained an insight into the relationship between various factors (age, gender, chest pain type, blood pressure, cholesterol levels, etc.) and the presence of heart disease. We developed predictive models to assess the risk of heart disease in individuals based on their characteristics and medical history using KNN and NB model.

We mainly focused on target being the dependent variable as it is what detects whether there is a presence of heart disease in the individual and it is what we aim to explain or predict.

Furthermore, when it comes to modelling perspectives, the model is trained using other variables such as age, sex, cholesterol levels and etc. to predict the value of the target variable.

Literature Review

Heart disease, also known as cardiovascular disease (CVD), remains a leading cause of morbidity and mortality globally. Detecting heart disease early is crucial for effective management and prevention of adverse outcomes. In this section we will explore various aspects of heart disease detection, including its formation, risk factors, and diagnostic methods, drawing insights from recent studies and established literature.

Heart disease includes a variety of disorders that impact the heart and blood arteries, such as arrhythmias, heart failure, and coronary artery disease. Heart disease is frequently multifaceted, resulting from the intricate interaction of underlying medical disorders, lifestyle factors, and genetic susceptibility.

Heart disease development is significantly influenced by genetic predisposition. Numerous genetic variations linked to lipid metabolism, inflammation, and endothelial function have been found to be associated with an elevated risk of CVD (McPherson et al., 2016). Nonetheless, the pathophysiology of heart disease is also significantly influenced by environmental variables, including obesity, smoking, physical inactivity, and poor diet (Benjamin et al., 2019).

Detection of heart disease:

Timely intervention and risk stratification are contingent upon the early identification of cardiac disease. Heart disease is detected using a variety of methods, including modern imaging modalities, biomarker evaluations, and non-invasive screening procedures.

Non-invasive Screening procedures:

In the early diagnosis of cardiac disease, non-invasive screening procedures like electrocardiography (ECG) and echocardiography are essential. While echocardiogram offers useful information on the structure and function of the heart, electrocardiography (ECG) is frequently used to detect irregular heart rhythms and conduction abnormalities (Dweck et al., 2019).

Advanced Imaging Modalities:

These technologies help diagnose myocardial infarction and coronary artery disease by providing detailed visualization of the coronary arteries and myocardial tissue. Examples of these technologies include cardiac magnetic resonance imaging (MRI) and coronary computed tomography angiography (CCTA) (Greenwood et al., 2016).

Lastly, heart disease identification is a complex procedure that includes diagnosing risk factors, determining the pathophysiology of the condition, and using a variety of diagnostic techniques. Early identification makes prompt management and intervention easier, which eventually improves patient outcomes and lessens the burden of cardiovascular disease.

Business Understanding

The business understanding phase is the initial stage that forms the basis of any project involving data mining or machine learning. Gaining a thorough grasp of the business opportunity or problem that the project seeks to solve is the main goal of this phase. Below is a summary of what the Business understanding is

Identify Business Objectives:

- The primary objective is to improve the detection, management, and prevention of heart disease.
- Reduce the incidence of heart-related complications and mortality rates.
- Enhance patient outcomes and quality of life through timely intervention and personalized treatment strategies.
- Optimize healthcare resources and reduce healthcare costs associated with heart disease management.

Assess Situation:

- Review existing healthcare practices and protocols related to heart disease diagnosis, treatment, and prevention.
- Analyze historical data on patient demographics, medical history, risk factors, and outcomes to identify trends and patterns.
- Understand the current challenges and limitations in heart disease detection, risk assessment, and treatment decision-making.

Determine Data Mining Goals:

- Develop predictive models to assess the risk of heart disease in individuals based on their health.
- Ensure data quality, completeness, and accuracy to facilitate reliable analysis and model development.

Data Understanding

This stage entails gathering, characterizing, investigating, and confirming the quality of the data to make sure it is appropriate for your research. The dataset for this analysis was sourced from Kaggle. After downloading the dataset from Kaggle, we loaded it into a Google Colab notebook for analysis.

The dataset consists of 1025 entries with 14 attributes, each of which represents a patient. These features include medical measurements or assessments (e.g., type of chest pain, cholesterol level) in addition to patient-specific information (e.g., age, sex). A synopsis of the initial several entries is provided below:

- Age: Patients' age varies from 29 to 77 years.
- Sex: Encoded as 1 (male) or 0 (female).
- Chest Pain Type (cp): Ranges from 0 to 3, indicating different types of chest pain.
- Resting Blood Pressure (trestbps), Cholesterol (chol), Fasting Blood Sugar (fbs), and other heart-related measurements are included.
- Target: Indicates the presence of heart disease, with 1 meaning presence and 0 meaning absence.

Data Types

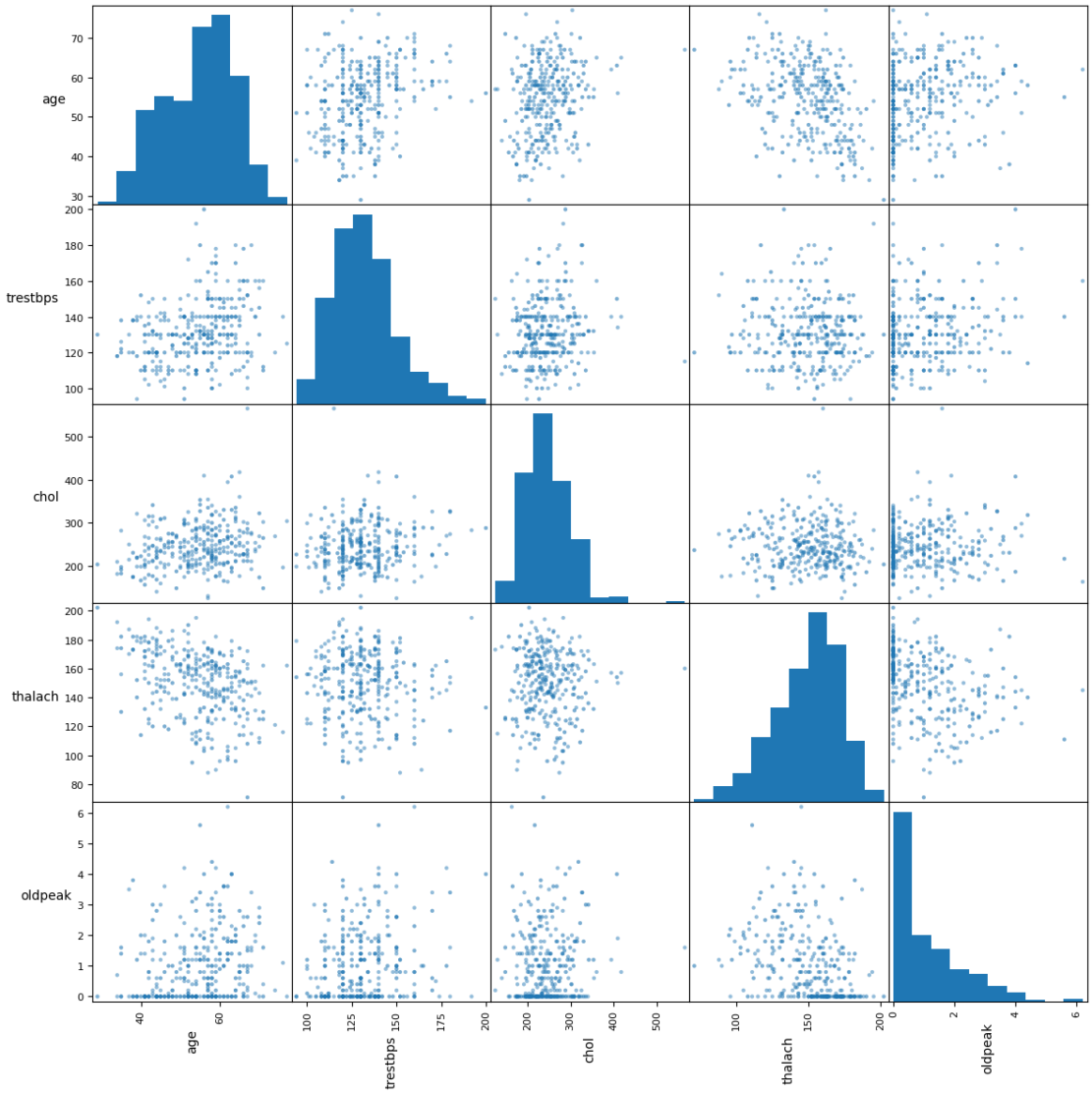
The "Non-Null Count" value that corresponds to the total number of entries for each feature shows that there are no missing values in any of the columns. This will make the data preparation process easier and is a great indication of the completeness of the data.

13 columns in the dataset are integers, and the dataset's predominant data type is integer (int64). These comprise quantitative measures (e.g., age, trestbps, chol) and categorical variables stored as integers (e.g., sex, cp, fbs).

Oldpeak is a floating-point number that shows how much of an ST depression is brought on by exercise in comparison to rest.

Statistical Overviews

The average age of the patients in the dataset is roughly 54 years old.
There is variation among the individuals in their cholesterol levels (chol), which average about 246 mg/dL with a standard deviation of roughly 51.59 mg/dL.
The maximal heart rate attained, or thalach, ranges from 71 to 202 bpm, with an average value of 149 bpm.
A patient's heart condition and exercise tolerance can be inferred from variables such as the kind of chest pain (CP), exercise-induced angina (exang), and the slope of the peak exercise ST segment (slope).



This scatter matrix was done to give us insights into the relationships between pairs of variables in our dataset. Since our dataset is based on age, resting blood pressure, cholesterol levels, and ST depression induced by exercise relative to rest, the scatter matrix will show us how these factors relate to each other.

Age distribution

The histogram of age suggests that the dataset includes a wide range of ages but is slightly skewed towards middle-aged and older individuals, which is a common demographic for heart disease studies

Resting blood pressure (trestbps)

The distribution of resting blood pressure appears to be roughly normal with a slight right skew. There is no immediately obvious pattern or correlation with age when visualizing the scatter plot between age and trestbps.

Cholestrol (chol)

The cholesterol levels also seem to be normally distributed with a right skew. The scatter plots do not show any clear linear relationship between cholesterol levels and age or resting blood pressure.

Maximum heart rate achieved (thalach)

The histogram for thalach shows a left-skewed distribution, indicating that higher maximum heart rates are more common in this dataset. The scatter plot of thalach versus age shows a potential negative correlation, suggesting that younger individuals tend to have higher maximum heart rates, which is consistent with physiological expectations.

ST depression induced by exercise relative to rest (oldpeak)

The oldpeak variable has a right-skewed distribution with many values clustered around zero. There doesn't seem to be a strong linear relationship between oldpeak and other variables, although there's a slight increase in variability of oldpeak with age.

Correlations

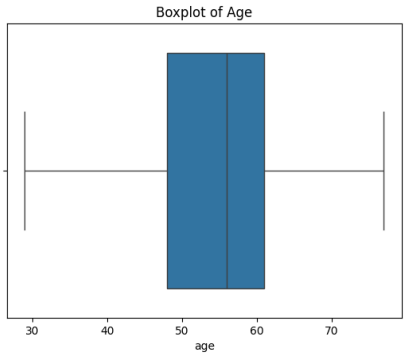
In general, the scatter plots do not show strong linear relationships between the variables, which suggests that if any relationships exist, they might be non-linear or might require more sophisticated statistical methods to uncover. Also, the presence of any clusters or distinct groups is not immediately apparent from the scatter plots.

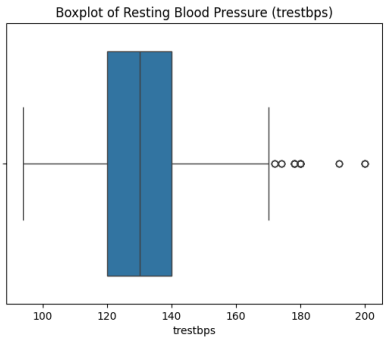
Outliers

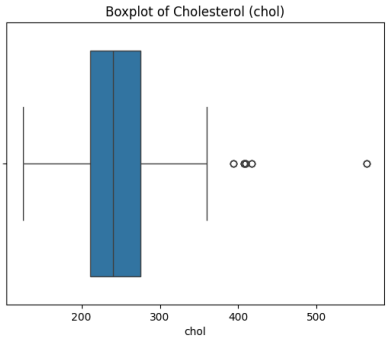
There seem to be some potential outliers, particularly in the cholesterol and oldpeak variables, which could be important when considering risk factors for heart disease.

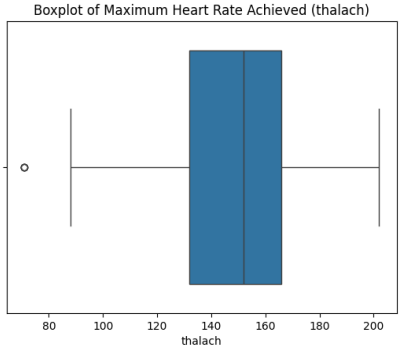
As we prepared our second step, we also concentrated on identifying and managing outliers. We decided to make boxplots for the quantitative characteristics in our dataset, including age, maximal heart rate attained (thalach), cholesterol levels (chol), resting blood pressure (trestbps), and ST depression brought on by activity in comparison to rest. IQR is especially helpful in identifying outliers. It computes the interquartile range (IQR), which is the range between the data's first (Q1) and third (Q3) quartiles. Any data points outside the range of Q1 - 1.5IQR to Q3 + 1.5IQR are therefore classified as outliers. When we looked at the boxplots for each characteristic, we saw that every attribute had outliers according to the IQR outlier detection method, except for age. It is imperative to identify these outliers as they have the potential to distort statistical studies.
Here's a summary of the outliers identified in the other variables of the dataset:

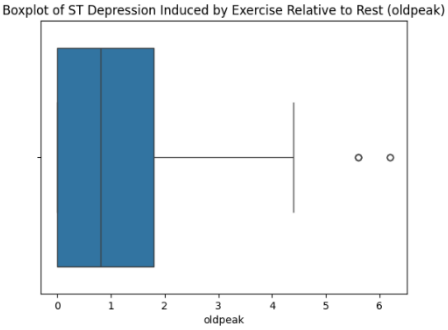
- Resting Blood Pressure (trestbps): 30 outliers identified, with values outside the range of 90 to 170 mm Hg.
- Cholesterol (chol): 16 outliers identified, with values outside the range of 115 to 371 mg/dl.
- Maximum Heart Rate Achieved (thalach): 4 outliers identified, with values outside the range of 81 to 217 bpm.
- ST Depression Induced by Exercise Relative to Rest (oldpeak): 7 outliers identified, with values outside the range of -2.7 to 4.5.











The decision to deal with these outliers should be based on considerations like extremely high or low values for trestbps, chol, or thalach might indicate specific health conditions or measurement errors, depending on the context. Similarly, unusual values for oldpeak could be significant for certain analyses, especially those focusing on exercise-induced cardiac events.

Given the context of health conditions and the potential significance of outliers in the dataset, we'll proceed with the next steps of preprocessing while retaining the outliers.

Data Preparation

Naïve Bayes & Random Forest Classifier

Preprocessing

We distinguish between categorical and numerical aspects in health data by recognizing their distinct characteristics. This separation enables us to use appropriate preprocessing approaches specific to each variable type, improving the accuracy and effectiveness of our study.

We divided our features into categories and numerical groups in order to account for the variety of data kinds. Along with numerical features like 'age', 'trestbps' (resting blood pressure), and 'chol' (cholesterol levels), categorical features like'sex', 'cp' (chest pain kind), and 'thal' (thalassemia) were found. This division is essential because it enables us to improve model accuracy by applying the proper preprocessing methods to each category of input.

With the help of ChatGPT, we used the ColumnTransformer to simplify the preprocessing workflow, assigning RobustScaler to numerical features and OneHotEncoder to categorical features. RobustScaler was chosen on purpose because it is less susceptible to data outliers, which is a prevalent issue in health datasets where outliers might reflect significant, non-noise anomalies. The OneHotEncoder is used to convert categorical information into a format suited for machine learning models, allowing the algorithm to correctly interpret these variables without assuming an ordinal relationship when none exists. We performed preprocessing on the dataset while carefully dismissing the ‘target’ variable —which denotes the presence or absence of heart disease— to enable supervised learning.

We optionally transferred the processed features back into a DataFrame for improved visibility and simplicity of analysis. This confirms that the data transformation was effective and enables us to look at the descriptive statistics of the processed data, providing insights into the outcomes of our preprocessing procedures. In order to validate the preprocessing and guarantee that the data is appropriately ready for any additional analysis or model training, this step is essential.

KNN

For data preparation we looked at variable we wanted to predict which is ‘target’ in the info() function and noticed that the variable is coded as a numerical value (integer). We needed it to be a categorical type thus we had to transform it from the integer type to category.

By examining these statistics using the function ‘cleandDF and ‘propDiag’, we can make informed decisions about data preprocessing steps, such as handling class imbalances, scaling features, or selecting appropriate distance metrics, to prepare your dataset effectively for KNN classification or regression tasks.

Normalization

We performed normalization function for KNN for the preprocessing step as it aims to scale the features to similar range, typically between 0 and 1 or -1 and 1. This is beneficial for KNN ad it

is sensitive to the scale of features. When we applied normalization to the heart disease dataset, it helped to ensure that features with different scale contribute equally to the learning process, preventing certain features from dominating due to them having bigger scale.

The result show values that ranges between 0 and 1 indicating that features of the heart disease dataset have been scaled proportionally and consistent within this range which can facilitate the training and evaluation of machine learning models on the heart disease dataset.

Modelling

Following the division of our dataset, we proceeded to the modeling phase

Training & Test Splits

We built a thorough grid of parameters with ChatGPT's help in order to investigate the effects of various training-test ratios and random seed settings on model efficacy.

Our approach comprised a thorough search over a preset grid of parameters, created especially to assess how different training-test ratios and random seed settings affected the performance of the model. The following is how the parameter grid was defined:

- Test Sizes: To investigate how varying amounts of data allotted for training versus testing might affect the model's capacity to generalize, a series of ratios [0.1, 0.15, 0.2, 0.25, 0.3] was used. This range helped us comprehend the trade-offs between training on more data and the advantages of a larger test set for validation by enabling us to evaluate performance across a spectrum from a minimal to a more substantial test set.
- Random States: To evaluate the model's sensitivity to data shuffles, a set of seeds [0, 42, 100] was selected. Our results are robust and not only the consequence of a specific data split that could unintentionally favor our model, thanks to the use of numerous seeds.
- Iterative Approach for Parameter Tuning: We used a grid search-like method by iterating over all possible combinations of test_size and random_state. This was selected to provide us with additional control over the procedure and to clearly show how each parameter affects the performance of the model. We were able to identify the ideal parameters that increased the accuracy of our predictions on the test data since the iterative procedure made sure we thoroughly investigated the parameter space.

The main objective was to determine which combination of test_size and random_state parameters gave our model the best accuracy. To guarantee that only the most efficient configuration was kept, we updated the optimal settings every time a higher accuracy was attained. We can systematically explore the parameter space with this simple yet effective direct comparison method.

1. KNN

K-Nearest Neighbors (KNN) is a machine learning algorithm commonly used for classification tasks, such as predicting whether a patient has heart disease based on various features in the dataset. KNN is a non-parametric algorithm, meaning it makes no assumptions about the underlying data distribution. This flexibility allows it to perform well in situations where the data may not follow a specific statistical distribution, which is common in healthcare datasets like the heart disease dataset. KNN is relatively easy to understand and implement, making it a good choice for beginners or when interpretability is important. The algorithm's concept of classifying data points based on their proximity to neighboring points is intuitive and straightforward to grasp.

A KNN model with low bias but high variation can be made more flexible by using fewer neighbors such as 5. However, using additional neighbors, ten or fifteen can result in a smoother decision border, which may increase bias while decreasing variance. The five neighbors achieve a balance between variation and bias. Thus, causing us to stick to 5 neighbors.

For us to achieve close to 100% accuracy we incorporated an optimum K value function to help us identify the K value that will help the model we're creating to give us 100% accuracy. After incorporating the function, it showed that the optimum K value should be 1 in order to achieve 100% accuracy. Though it said k value of 1 we decided to stick with 5 as I previously mentioned that it helps reduce risk of overfitting. It also helps improve robustness to noise and a better balance between bias and variance.

2. Naïve Bayes

We selected the Gaussian Naive Bayes algorithm for its simplicity and effectiveness in handling the mix of categorical and numerical variables present in our data. Initiation of the Naive Bayes classifier is achieved through the `GaussianNB()` method. This model is particularly adept at working with continuous data, assuming the data for each label follows a Gaussian (normal) distribution. This assumption aligns well with our heart disease dataset, where several numerical features, through appropriate scaling or transformation, can be made to closely mimic a normal distribution. Notably, Gaussian distributions are a suitable approximation for features such as 'age', 'trestbps' (resting blood pressure), and 'chol' (cholesterol level), among others, making the Gaussian Naive Bayes model a fitting choice for our analysis.

The results from our parameter tuning process for the Naive Bayes classifier:

- It is indicated that the optimal results were obtained by dividing the dataset into 10% for testing and the remaining 90% for training.
- The random_state parameter ensures reproducibility of the results by controlling the randomness of the train-test split. This precise split of the data was most advantageous for the model, as indicated by the value of 100 that produced the maximum accuracy.

These settings led to the highest achieved accuracy of approximately 88.35%.

3. Random Forest Classifier

Because it effectively handles binary and multi-class categorizations and makes use of the Random Forest ensemble technique, we chose the RandomForestClassifier. This technique reduces the likelihood of overfitting while increasing the accuracy of the model by building multiple decision trees and combining their predictions. Alongside `test_size` and `random_state`, we also consider `n_estimators` to identify the best parameters for the Random Forest model.

- [50, 100, 150] N_estimators: The `n_estimators` values of [50, 100, 150] were chosen for the Random Forest model to explore the balance between performance and computational efficiency. Increasing the number of trees can improve accuracy and reduce overfitting, but beyond a certain point, gains are minimal compared to the added computational cost. This range allows us to assess performance improvements within a realistic computational budget.

The results from our parameter tuning process for the Random Forest classifier:

- The model performs best when 10% of the dataset is used for testing and 90% is used for training, using a test_size parameter of 0.1. Reproducibility of the split is ensured by setting random_state to 0, which removes performance variability caused by data shuffling. This particular seed produced the best model performance, showing advantageous data partitioning, when combined with the test size.
- The Random Forest's selection of 50 trees (n_estimators = 50) shows that a comparatively small number of trees was adequate to successfully capture the underlying patterns in the data. This supports the idea of diminishing returns as model complexity increases by indicating that, at least for this dataset, adding more trees to the model does not always result in increased accuracy.

These settings led to the highest achieved accuracy of approximately 100%

Evaluation

We assess our model's performance using multiple indicators, mainly a classification report and a confusion matrix. We can examine not only the overall accuracy but also the model's performance in predicting each class (presence or absence of heart disease) thanks to the confusion matrix, which provides a visual and numerical representation of the model's predictions compared to the actual labels. The classification report provides a more detailed analysis of the model's advantages and disadvantages by breaking down the precision, recall, and F1-score for each class. These indicators are essential for assessing the performance of our model, particularly in the medical setting where false positives or negatives can have serious consequences.

Our attempts to investigate data prediction, particularly with the Naive Bayes (NB), Random Forest (RF), and K-nearest neighbors (KNN) models, have produced genuinely astonishing results. We followed a rigorous procedure, precisely guided by ChatGPT's amazing support, to unleash the full power of our models through numerous parameters and setups.
Across careful parameter tuning and iteration across various configurations, we were able to have a better understanding of the ways in which each element affected the predictive capacity of our models. This degree of control and precision raised our predictions' accuracy and gave us more faith in the dependability of our findings.

1. KNN

When applied to the heart disease dataset, the K-Nearest Neighbors (KNN) regression model would yield predictions for a continuous target variable linked to heart disease. KNN classification is more frequently utilized since the heart disease dataset usually consists of classification tasks (predicting the presence or absence of heart disease).

KNN regression model would make predictions by averaging the target values of the k nearest neighbors in the feature space. The predicted continuous outcome could provide valuable insights into the patient's heart health status and help guide clinical decision-making and treatment planning. With an accuracy of 77%, the KNN model developed for this case study is thought to be operating well. Healthcare institutions can use this accurate model as a reference to forecast if patients will have heart disease.

The result concluded at the end of the KNN regression model is shown through confusion matrix, it shows that:

The top-left cell (43) represents the number of instances where the model correctly predicted that the patient has heart disease (positive class) and the actual label was also positive. False Positive (FP): The top-right cell (15) represents the number of instances where the model incorrectly predicted that the patient has heart disease (positive class) when the actual label was negative (no heart disease). False Negative (FN): The bottom-left cell (8) represents the number of instances where the model incorrectly predicted that the patient does not have heart disease (negative class) when the actual label was positive. True Negative (TN): The bottom-

right cell (37) represents the number of instances where the model correctly predicted that the patient does not have heart disease (negative class) and the actual label was also negative.

From the confusion matrix, we can derive various performance metrics such as accuracy, precision, recall, and F1-score, which provide a more comprehensive understanding of the model's performance.

The result provided for the classification model performance shows that precision is 0.84 for class 0 (class 0 represents the absence of heart disease) and for class 1 (represents presence of heart disease) the precision is 0.71. This result conveys that among the instances predicted as not having heart disease, 84% of them are predicted correct while the instances predicted as having heart disease, 71% of them were predicted correct.

The recall measures the proportion of true positive predictions among all actual positive instances in the heart disease dataset. For class 0, the recall is 0.74 and for class 1 the recall is 0.82. This insinuates that the model correctly identifies 74% of the instances with absence of heart disease and 82% of the instances with presence of heart disease.

The F1-score is the harmonic mean of precision and recall and provides a single metric that balances both precision and recall. It ranges from 0 to 1, where 1 indicates perfect precision and recall. For class 0, the F1-score is 0.79, and for class 1, the F1-score is 0.76.

2. Naïve Bayes

As the presented findings demonstrate, the Naive Bayes model performs admirably in detecting people who have heart disease while also doing a great job of identifying those who do not. The model's precision and recall measures nearly coincide across both classes, supporting its high efficiency in overall disease prediction, as evidenced by its accuracy score of 88.35%.

In terms of heart disease prediction, the model shows a precision of 0.90 and a recall of 0.90, demonstrating a high degree of accuracy in detecting actual positive instances along with a low number of false negatives. This implies that the model has a high degree of reliability when it comes to identifying heart disease, which is important for treatment planning and early intervention. The model appears to retain a balanced approach to disease detection, not overly biased toward false positives or negatives, based on equal precision and recall.

The precision and recall for predictions of no disease are both 0.87, demonstrating the model's consistent ability to correctly identify people who do not have heart disease and reduce the number of false alarms. In order to avoid needless stress and medical procedures for those who are misdiagnosed, this balance is crucial in medical diagnostics.

Together, the classification report and confusion matrix show a refined model that successfully balances the needs of sensitivity and specificity. The model guarantees that a considerable majority of disease cases are detected, with 52 true positives and only 6 false negatives in terms of disease detection. In addition, it guarantees that there are as few people mistakenly diagnosed with the illness as possible, with 39 true negatives and a smaller number of 6 false positives.

This delicate balance is especially crucial in the field of medical diagnostics, where a false positive might cause needless worry and medical intervention, while a false negative can have

very high consequences and possibly result in an untreated condition. The model's performance demonstrates how useful it may be in the early diagnosis of heart disease and provides a solid basis for future improvements to maximize its prediction power.

3. Random Forest Classifier

The classification report and the confusion matrix both demonstrate the remarkable performance of the Random Forest model. With no false positives or false negatives, a perfect classification is revealed by the confusion matrix, consisting of 48 true positives (disease correctly recognized) and 55 true negatives (no disease accurately detected). This demonstrates an exceptional level of specificity and sensitivity, or recall, both scoring 1.00, or 100%, which is an uncommon result in predictive modeling. We're utterly surprised by our Random Forest model's faultless performance, which it achieved with ChatGPT's help and 100% accuracy.

These results are supported by the classification report, which verifies a precision of 1.00 for each class. This suggests that every case the model projected to be disease-related was, in fact, a case of disease (precision for class 1), and that every case expected to be non-disease-related was, in fact, true (precision for class 0). Furthermore, the recall of 1.00 for both classes indicates that the model can accurately and flawlessly identify every real occurrence of both classes.

Cross-validation Evaluation

There were first concerns after the model correctly classified every example in the test set with a 100% accuracy rate, prompting us to ask ChatGPT for help in order to verify this flawless score, a cross-validation assessment was carried out with cross_val_scores function.

It is impressive to see the results of the cross-validation process on the Random Forest model with 50 trees. The accuracy scores for each of the five folds are contained in the array [1.0, 1.0, 1.0, 1.0, 0.98536585]. The model achieved perfect accuracy (100%) in four of the five folds, and around 98.54% accuracy in the fifth fold.

The accuracy scores throughout the folds have a very low standard deviation of around 0.00585. The model appears to be stable and insensitive to the specific subset of data it was trained on, as seen by the little fluctuation in its performance over the several cross-validation folds.

We decided to proceed with this conclusion, keeping in mind the model's demonstrated stability and precision, given the remarkable performance of the Random Forest model, as evidenced by nearly flawless accuracy throughout cross-validation folds.

Deployment

The properties of the dataset, model performance metrics, computing needs, and interpretability are only a few of the elements that must be taken into consideration while choosing the optimal model for heart disease diagnosis. Here are some pros and cons of each model and factors to take into account when determining whether one is appropriate for the heart disease dataset situation:

1. KNN

Advantages: KNN is an easy-to-understand algorithm. It may capture intricate relationships in the data and makes no assumptions about the distribution of the underlying data. For datasets with nonlinear decision boundaries, it might work well.

Cons: Because KNN stores and searches through the whole training dataset for each prediction, it can be computationally expensive, particularly for large datasets. When dealing with sparse or high-dimensional data, it could not work well.

Considerations: Choosing the right number of neighbors (k) has a significant impact on how well a KNN performs. To get the best results, this hyperparameter must be properly adjusted.

2. Naïve Bayes

Advantages: The probabilistic classifier Naive Bayes is easy to use and computationally efficient. It is capable of handling both numerical and categorical features and performs well with high-dimensional data. When working with relatively tiny datasets, it is especially effective.

Cons: In real-world datasets, Naive Bayes may not hold true because it presupposes that features are conditionally independent given the class label. Complex interactions between features might not be captured by it.

Considerations: Despite its simplicity, Naive Bayes can perform surprisingly well, especially on text classification tasks. However, its performance may suffer if the independence assumption is violated.

3. Random Forest Classifier

Advantages: Handles high-dimensional data well

- Provides feature importance scores.
- Offers robustness and improved generalization performance

Cons: Training time can be longer, especially with large datasets

- May require tuning of hyperparameters to optimize performance

Considerations: Compared to certain other methods, Random Forest exhibits less sensitivity to hyperparameters and typically performs well across a wide variety of datasets. Compared to other models, it is less prone to overfit and can capture intricate feature relationships.

The best classification for this dataset: Random Forest classification

Considering the nature of the heart disease detection dataset, which typically involves a moderate-sized dataset with both numerical and categorical features, as well as potential nonlinear relationships, Random Forest classification is often a good choice. It provides robust performance, handles various types of features effectively, and is less prone to overfitting. However, it's essential to experiment with different models and evaluate their performance using appropriate metrics to determine the best approach for a specific dataset and problem domain.

Here are some reasons why random forest classification is the best use for this heart disease dataset:

- Robustness against Overfitting: Random Forest models are less likely to overfit compared to individual decision trees. By combining the predictions of several trees, Random Forest can efficiently handle noise and unpredictability in the dataset and generalize well to new data.
- Handling of Nonlinear correlations: Random Forest can capture nonlinear interactions between variables, which is useful for heart disease identification, as it may include complex correlations.

In terms of performance metrics, the Random Forest model achieved an accuracy score of 88.35%. For heart disease prediction, the model shows a precision of 0.90 and a recall of 0.90, demonstrating a high degree of accuracy in detecting actual positive instances along with a low number of false negatives. This implies that the model has a high degree of reliability when it comes to identifying heart disease, which is important for treatment planning and early intervention.

The precision and recall for predictions of no disease are both 0.87, demonstrating the model’s consistent ability to correctly identify people who do not have heart disease and reduce the number of false alarms. This balance is crucial in medical diagnostics to avoid needless stress and medical procedures for those who are misdiagnosed.

Conclusion

Suggestions

Here are some measures that hospitals should take into account when using either one of the classifications talked about above:

Firstly, they should consider the data quality and preprocessing. By ensuring that the dataset used for training the models is of high quality, with accurate and reliable data, the result the models will produce will be positive. Furthermore, they should thoroughly perform data preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features. By doing this, it will improve the performance of the models. To add on, hospitals should identify relevant features that may contribute to heart disease detection and include them in the analysis. By feature selection it can help prioritize important features and reduce dimensionality. They should also consider engineering new features or transforming existing ones to acquire additional information that may be relevant for heart disease diagnosis.

Furthermore, hospitals should consider the interpretability of the models (making it easier to read and understand) especially in a clinical setting where transparency and explainability are crucial. Interpretability techniques such as features importance analysis, partial dependence plots and Shapley additive explanation values can help understand the models' predictions and provide insights into decision making process.

Hospitals should engage in continuous monitoring and improvement. Continuously monitoring the performance of the models over time will help them identify places they need to focus on or need better work so that the model works effectively, they should also update them as new data is added. They should also incorporate feedback from healthcare professionals and patients to iteratively improve the models and adapt them to changing clinical needs and practices.

Lastly, hospitals should consider ethical considerations such as fairness, bias, and transparency in model development and deployment, especially when making decisions that may impact patient care. By following these suggestions, hospitals can effectively leverage machine learning models for heart disease detection while ensuring their reliability, interpretability, and ethical use in clinical practice.

Personal Reflection

Gifty Mensah

Working on this project has definitely opened my eyes to a whole new world of data analytics. At first, I thought excel was an excellent data analytical tool to use however after this project it has showed me that there are more than just function in excel but also in python which gives a broader and deeper understanding of analyzing large sums of data. However, I won't say it was quite an easy course, most definitely not. It was interesting and complicating to understand at times, however it was very interesting embarking on this journey. If it wasn't for ChatGPT I would say that this course would have been impossible for me to understand. ChatGPT really broke down the code and why they are used which made it very useful and easier to understand.

I was intrigued by the models that were created every time a code is entered on python. It made it less boring and intimidating because it's not all about just seeing the code but also seeing the model you create which makes you proud of your work. I would definitely say that I enjoyed it very much. I'm proud of how far I've come with knowing nothing on coding and leaving the class with some knowledge of coding that I can use in real time bases.

Giang Lieu

The journey through the CRISP model report has been incredibly personal and enlightening for me. Working with real-world data, delving into the complexities of dataset preparation, and developing code solutions to solve complicated challenges have not only broadened my technical skills but also impacted my perspective of the data analytics ecosystem.

Through many tries and errors, I've uncovered the actual heart of data analysis: it's more than just crunching statistics; it's about comprehending the story concealed inside the data. Each problem presented an opportunity for progress, encouraging me to try new ideas and improve my techniques. What's truly astounding is ChatGPT's crucial involvement in this trip. Its advice and insights have proven invaluable, not only streamlining processes but also inspiring fresh ideas and viewpoints that I had not before considered.

As I look back on this trip, I realize how far I've come both professionally and emotionally. The concepts and abilities I've learned have not only prepared me for the difficulties that lie ahead, but they've also sparked a newfound interest in data analytics. Moving forward, I feel empowered and eager to apply what I've learned to new projects and endeavors. This trip has not only expanded my grasp of data analytics, but it has also given me the confidence and determination to keep pushing the boundaries and make meaningful contributions in this sector.

APA references

- Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., ... & Virani, S. S. (2019). “Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation,” retrieved from https://www.ahajournals.org/doi/10.1161/CIR.0000000000000659

- Dweck, M. R., Bularga, A., Hahn, R. T., Bing, R., Lee, K. K., Chapman, A. R., ... & van Beek, E. J. R. (2019). “Global evaluation of echocardiography in patients with COVID-19. European Heart Journal-Cardiovascular Imaging” retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7337658/

- Greenwood, J. P., Ripley, D. P., Berry, C., McCann, G. P., Plein, S., Bucciarelli-Ducci, C., ... & Croisille, P. (2016). “Effect of care guided by cardiovascular magnetic resonance, myocardial perfusion scintigraphy, or NICE guidelines on subsequent unnecessary angiography rates: the CE-MARC 2 randomized clinical trial. Jama” retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5329750/

- McPherson, R., Tybjaerg-Hansen, A., & CRP Studies Collaboration. (2016).” Genetics of coronary artery disease. Circulation Research” retrieved from: https://www.ahajournals.org/doi/full/10.1161/circresaha.115.306566

Appendix

Our dataset and codes created for the project can be accessed on Github via the provided link https://github.com/Liz283/DataMiningAssignment/tree/main

The source for our datasets can be accessed via https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset