Heart Disease Prediction with Machine Learning Models

2023-05-05

Ali Suhail
21072712

Yie Nian Chu
21061710

Edgaras Levinskas
21065305

1 Abstract

Heart diseases are hard to diagnose due to the large amount of data involved. Many features such as age, gender, heart rate, blood pressure etc. affect the chances of someone having a heart attack or else. In this experiment we used machine learning (ML) algorithms to train on an existing patients' dataset. Then use the algorithms to predict if a person is at a high risk of a heart attack in the future.

We approached this problem by first analysing the dataset, checking for outliers, and removing erroneous data. Once the dataset was processed, we used it to train algorithms such as Support Vector Classifier (SVC), Logistic Regression (LR), Random Forest (RF) and Hard Voting Classifier (HVC). Then, using testing dataset, we verified the results and compared each algorithm according to their precision, accuracy, and recall scores.

2 Introduction

According to CDC (CDC, 2023) the leading cause of death in U.S. is heart diseases. Heart diseases include but are not limited to heart attacks, arrhythmia, valve failures etc. By analysing people's features such as age, gender, heart rate and glucose we can try to predict if a person is at risk of having heart disease. However, due to the amount of data that needs to be analysed, doing so by hand would take weeks if not months. This is where ML algorithms shine: they train on large amounts of existing data quickly and then can use their gained knowledge to make predictions on new inputs. It is believed that many fields in medicine could greatly benefit from having ML algorithms incorporated into diagnosing and treating patients. On the other hand, there are no algorithms that are 100% accurate, therefore they should never be used as replacements for doctors and nurses but rather as a second opinion/assistant tool.

After training various ML algorithms and comparing their results, we have found that they can reliably predict heart attacks in a given dataset. This is strong evidence that ML algorithms can be used in medical fields to assist doctors and nurses to make diagnosis and treatment faster and more reliable.

3 Related Work

To predict different types of disease, such as heart disease, breast cancer, and diabetes, many researchers have employed various types of ML algorithms, such as LR, SVC, and even Artificial Neural Network.

Singh and Kumar's paper "Heart disease prediction using machine learning algorithms" (2020) claims that a variety of ML methods, including SVM, Decision Tree (DT), LR, and K-Nearest Neighbour (KNN), are utilised to predict heart disease using a given dataset. According to the results, KNN algorithm had the best accuracy (87%) and SVM had the second-highest accuracy (83%). Lower accuracy was reached by the DT and LR, at 79% and 78%, respectively.

In addition, Jindal's study (2021) employed KNN, LR, and RF classifiers to predict heart disease, and the features they had included were age, sex, and various forms of chest discomfort, resting blood pressure, serum cholesterol, and others. With those features, the results they obtain from the models differ from Singh and Kumar's paper. KNN and LR outperform RF classifier in the prediction of the patient's heart disease diagnosis, with results equal to 88.5%.

Other than that, Sharma's study (2020) states that in order to predict heart disease, they employed SVM, DT, Naive Bayes (NB), a supervised ML method based on the Bayes' Theorem, and RF classifier. There are 1025 cases in their dataset, which they obtained from the UCI Repository, and 14 characteristics are taken into account in the study. According to their findings, SVM, which had an accuracy of 98%, and RF, which had an accuracy of 100%, outperformed NB and DT, which had accuracy rates of 88.9% and 96.6%, respectively.

The articles demonstrate that ML techniques have enormous promise for cardiac disease prediction. However, different choices of algorithms and features can significantly affect how accurate the predictions are. As shown by the studies mentioned earlier, accuracy rate can vary from 78% to 100%. It is also crucial to keep in mind that the datasets utilised in those research varied in size and other aspects, which might have an impact on how well the algorithms performed.

Following the analysis of numerous studies, we decided to implement SVC, RF, LR, and HVC and see how they would perform with the dataset we had selected as they have a high accuracy range in predicting heart disease, ranging from 88.5% to 100%. Even though KNN was used in two of the papers to predict the likelihood of developing heart disease, we made the decision not to apply KNN when creating our model. It is because KNN is sensitive to outliers, and our dataset has a lot of them, especially for the two features CK-MB and Troponin. Inaccurate predictions from KNN would result from outliers in our dataset, which would lower accuracy.

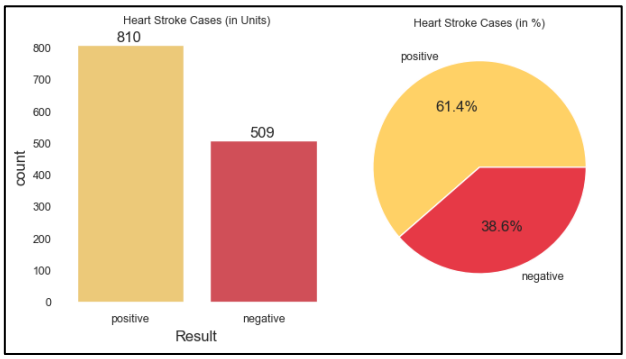


Figure 1 Heart Stroke Cases Distribution in Dataset

Furthermore, we discovered that the dataset contains an uneven number of heart stroke cases (Figure 1), which would have an impact on KNN's performance because it favours the majority class when making predictions.

4 Data

4.1 Collection

Data collection for research must abide by ethical guidelines to ensure that participants are not harmed and that their privacy remains secure. For instance, while collecting medical data, it must be guaranteed that patients have the option to decline or provide their written consent. To preserve the privacy of participants, the data must be kept private and anonymized. According to University of Kurdistan Hewler (UKH, 2023), they are accredited by the Accreditation Service for International Schools, Colleges, and Universities (ASIC). Therefore, we believe UKH's data collection would be morally and legally acceptable.

After examining several datasets, we have decided to use a heart attack dataset that is available online on Mendeley Data (Rashid, Tarik A. 2022). It was chosen because the dataset was collected in 2019 by Dr. Tarik Ahmed Rashid, an Iraqi professor in the Department of Computer Science and Engineering at the University of Kurdistan Hewler. Moreover, it is sufficiently large and comprehensive.

These are the 9 characteristics taken into account:

Attribute	Description	Mean
Age	Patient's age	56.192
Gender	Patient's Gender (0 – female, 1 – male)	0.6596
HeartRate	Patient's heart rate	78.337
SBP	Systolic Blood Pressure	127.171
DBP	Diastolic Blood Pressure	72.269
BloodSugar	Patient's blood sugar	146.634
CKMB	Creatine Kinase-myoglobin binding	15.274
Troponin	Patient's troponin complex	0.361
Result	Output (positive – 1, negative – 0)	0.614

Table 1 Attributes

Table 1 shows the mean value of each attribute. There are 1319 entries in total. The dataset includes a multi-class variable and binary classification. The multi-class variable shows the presence or absence of heart disease with a value of 1 and 0 respectively.

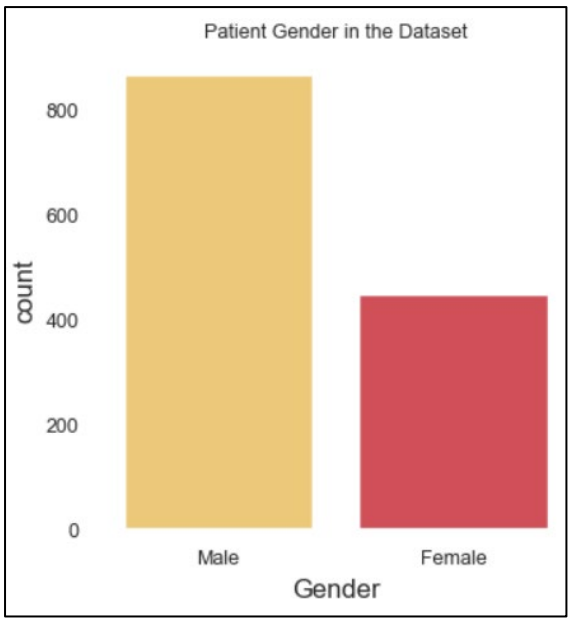


Figure 2 Gender Distribution in Dataset

Unfortunately, as seen in Figure 2, the gender distribution in the dataset is unbalanced. A heart disease classification model's prediction accuracy might potentially be impacted by this, especially if the predictions are biased. The model might be more accurate in predicting male patient's results as the dataset comprises more male patients than female patients. Nevertheless, we believe that this dataset is thorough and has sufficient amount data points.

4.2 Pre-processing

Data preparation procedures are used to deal with the dataset's erroneous values. For each feature, we produced a histogram as well as a box and whisker plot (Appendix 1) to better comprehend the dataset. The graphs demonstrate that there are various numbers of outliers in each characteristic. In order to deal with the outliers, we established a range for each feature based on the graphs so that it would only eliminate those anomalous outliers or else it would result in the removal of all the data.

There are 1319 patient records altogether in the dataset before the outliers are removed. 48 records were then taken out of the dataset because they had erroneous values, leaving 1271 data points for our models to be trained and tested (Appendix 2).

4.3 Scaling

Since some of the features still contain a significant number of outliers (Appendix 2), we decided to use MinMax Scaler to scale them since it works best when the scale of the feature values is known and stable, and there are no notable outliers. It also keeps the original distribution's shape and the relationships between the data points.

5 Methods

We used Scikit-learn which is a valuable library for our ML models as it offers tools for data pre-processing, model selection, and evaluation. It is also widely used and well-documented, making it easy to implement and replicate algorithms. Once we process and scale the dataset, it can be used to train various ML algorithms. After experimenting with four models: SVC, LR, RF and HVC, we found that using 20% of the data as testing and 80% as training produced the best results overall. Moreover, we made sure to stratify the test and training datasets on results column. That ensures an equal amount of positive and negative results in both sets, thus making the algorithm less biased towards one or the other.

We also tried several combinations of hyperparameters for each model using a trial and error approach. This involved testing various combinations of hyperparameters to determine the best set for each model. We did this to ensure that our models were performing at their optimal levels and to achieve the highest accuracy possible.

5.1 Logistic Regression

We used LR as it is a simple and efficient algorithm that can handle both linear and nonlinear relationships between the features and the target variable (Result). Additionally, LR has the advantage of providing interpretable results, which can help in understanding the contribution of each feature to the final heart attack prediction.

5.2 Support Vector Classifier

SVC is a beneficial algorithm to use with a heart attack dataset that has 8 features due to its ability to handle high-dimensional data and nonlinear relationships between variables. It also has a strong ability to handle noisy and complex data, making it a good choice for medical datasets that have a lot of variation and uncertainty.

5.3 Random Forest

RF may be helpful with a heart attack dataset because they can handle high-dimensional datasets with numerous traits, and they are especially useful when

the features have complicated and nonlinear interactions with the target variable. They are also less prone to overfitting than other ML methods, such as DT, because of their ability to combine the outputs of several trees and decrease model variance.

5.4 Voting Classifier – Hard

The HVC combines the predictions of multiple models, which can lead to improved performance and reduced risk of overfitting. In this case, combining SVC (Poly), LR, and RF could potentially result in a more robust and accurate model. Additionally, using multiple models with different strengths and weaknesses can help capture different aspects of the dataset like SVC (Poly) may perform well in capturing non-linear relationships between features and the target variable. Likewise, LR may perform well in handling categorical features and linear relationships, and RF may perform well in handling complex and high-dimensional datasets.

6 Experiments and Evaluation

To evaluate and analyse the heart attack dataset, we employed the 'classification_report' function, which provided us with metrics such as accuracy, precision, and recall for each model. Additionally, we used 'cross_val_score' to obtain cross-validation scores, which allowed us to assess the model's generalization performance on unseen data.

6.1 Logistic Regression

6.1.1 Parameter Tuning

These are the parameters used for the LR model.

```
1. LogisticRegression(C=1, solver='lbfgs')
2. LogisticRegression(C=100, solver='lbfgs')
3. LogisticRegression(C=1e5, solver='lbfgs')
4. LogisticRegression(C=1e10, solver='lbfgs')
5. LogisticRegression(C=1e100, solver='lbfgs')
6. LogisticRegression(C=1e5, solver='saga',
   max_iter=1000)
```

Hyperparameters choice for Logistic Regression

We observed that increasing the C value had a significant positive impact on the model's performance. However, raising the C value beyond 1e5 did not yield any further improvements in accuracy scores. We also attempted to use the saga solver (6th test case), but it resulted in convergence issues as it was not suitable for a small dataset like heart attack dataset. Therefore, the 'lbfgs' solver was chosen because it performs best for small to medium-sized data. This solver uses an optimization approach that is effective and does not require a lot of memory, making it a better option compared to other solvers.

Solver "Limited-memory Broyden–Fletcher–Goldfarb–Shanno" (lbfgs)			
Parameters	Cross-val Accuracy	F1 – Score	
C		Positive	Negative
1	72%	79%	59%
1e5	95%	95%	93%
1e10	95%	95%	93%

Table 2 Logistic Regression Results

Table 2 clearly indicates that increasing the value of C has a positive impact on the performance of LR, which is particularly advantageous for datasets that are noisy and complex such as the heart attack dataset. This is evident in the considerable increase in accuracy scores from parameter C=1 to C=1e5, which resulted in an impressive 23% improvement. Additionally, the F1-scores for positive and negative cases improved by 16% and 34%, respectively.

6.2 Support Vector Classifier

6.2.1 Parameter Tuning

These are the following hyperparameters used for the SVC model.

```
1. SVC(kernel="poly", C=1, coef0=1, degree=1)
2. SVC(kernel="poly", C=1, coef0=1, degree=2)
3. SVC(kernel="poly", C=5, coef0=1, degree=3)
4. SVC(kernel="poly", C=10, coef0=5, degree=3)
5. SVC(kernel="poly", C=20, coef0=10, degree=4)
6. SVC(kernel="poly", C=30, coef0=20, degree=5)
7. SVC(kernel="rbf", C=20, gamma=1)
8. SVC(kernel="rbf", C=1, gamma=5)
```

Hyperparameters choice for Support Vector Classifier

The 5th Hyperparameters produced the best results among all the tested cases, while the 7th parameters, which uses an RBF kernel, came in second. We also took notice of increasing the C value, coefficient and degree having a positive impact on the model's performance.

The effects of changing parameters can be seen in the table below:

Kernel "Polynomial" (poly)					
Parameters			Cross-val Accuracy	F1 – Score	
C	coef0	degree		Positive	Negative
1	1	2	76%	82%	74%
20	10	4	93%	94%	91%

Table 3 SVC Polynomial Results

Kernel "Radial Basis Functions" (rbf)				
Parameters		Cross-val Accuracy	F1 – Score	
C	gamma		Positive	Negative
1	5	71%	77%	59%
20	1	79%	82%	75%

Table 4 SVC RBF Results

The reason as to why 5th test case outperformed 7th with an RBF kernel may be due to polynomial kernels allowing a more flexible decision boundary compared to a linear kernel. Making it suitable for non-linear and complex datasets. It is also more efficient than RBF kernels, which can be computationally expensive for larger datasets. SVC (Poly) is also less vulnerable to overfitting than SVC (RBF), particularly when dealing with high-dimensional feature spaces. The polynomial kernel constrains the flexibility of the decision boundary, which prevents it from excessively bending to accommodate the noise in the data.

Like LR, increasing the value of C significantly improved the accuracy of predictions in patients by more severely penalizing incorrect classifications. However, high values of C may cause overfitting problems in the heart attack dataset.

Moreover, increasing the degree and coefficient values (Table 3) has also resulted in a significant improvement in the model's performance, with accuracy scores improving from 76% to 93%, slightly worse than the LR model, and F1-scores seeing more than 11% increase. What's more is that increasing the degree helps create a more complex decision boundary that can capture non-linear relationships between features and the target variable. Additionally, a high coefficient value places more emphasis on higher-order terms in the kernel, which increases the flexibility of the decision boundary and allows for more complex relationships between features and the outcome. However, there is a higher risk of overfitting with higher values of the coefficient in the heart attack dataset.

6.3 Random Forest

1. RandomForestClassifier(n_estimators=1, random_state=0)
2. RandomForestClassifier(n_estimators=75, random_state=0)
3. RandomForestClassifier(n_estimators=150, random_state=0)

Hyperparameters choice for Random Forest

Random Forest				
Parameters		Cross-val Accuracy	F1 – Score	
n_estimators	Random state		Positive	Negative
1	1	92%	96%	94%
75	1	98%	100%	99%
150	1	99%	100%	99%

Table 5 Random Forest Results

Table 5 shows that increasing the number of estimators in the RF model resulted in improved accuracy and F1 scores above 96%. However, the extremely high accuracy scores could indicate that the model is overfitting the data. This is a common issue with RF models, especially when the dataset is imbalanced, as is the case with the heart attack dataset where 60% of the cases are negative and male patients are twice as large as female patients.

6.4 Voting Classifier – Hard

Voting Classifier – Hard			
Parameters	Cross-val Accuracy	F1 – Score	
Estimators		Positive	Negative
RF, LR and SVM (Poly)	97%	98%	96%

Table 6 Hard Voting Classifier Results

As shown in Table 6, when combining all three models using a HVC, we achieved an impressive cross-

validation accuracy of 97% and f1-scores just under 100%. This means that our model can accurately predict 97% of all patients who may or may not suffer from a heart attack. Additionally, among patients who have suffered from a heart attack, our model can accurately predict their condition 98% of the time based on features such as age, CK-MB, Troponin, and more. Despite the RF model being prone to overfitting, the HVC was able to leverage the strengths of all three models to create an overall better predicting model.

6.5 Confusion Matrix

Confusion Matrix	Actual Positive	Actual Negative
Logistic Regression		
Predict Positive	93	7
Predict Negative	7	148
Cross-validation Score	95%	
Support Vector Classifier (Polynomial)		
Predict Positive	89	11
Predict Negative	7	148
Cross-validation Score	93%	
Random Forest		
Predict Positive	99	1
Predict Negative	0	155
Cross-validation Score	99%	
Hard Voting Classifier		
Predict Positive	95	5
Predict Negative	2	153
Cross-validation Score	97%	

Table 7 Confusion Matrix & Cross-validation Scores for each Model

The results presented in Table 7 show that RF had the best performance among the tested models, achieving an impressive accuracy score of 99% on the testing set (size 255), with only one negative case being incorrectly labelled as positive. LR and SVM (Poly) followed closely, with accuracy scores of 95% and 93%, respectively. Although the confusion matrices of SVC (Poly) and LR are similar, LR was slightly more accurate than SVC (Poly). The HVC model, which combines the strengths of all three models, achieved an accuracy of 97%. That is slightly lower than RF but with the added benefit of having very low misclassifications of people who had a heart attack as negative. This is crucial in healthcare, where reducing false negatives can potentially save lives.

7 Conclusion

In conclusion, we strongly believe that given the results of multiple ML algorithms (Table 7), they can be an invaluable tool in medicine. These algorithms could be integrated into various medical laboratories to provide second opinion or possible predictions when given patient features and history. This could also save time and cost for both hospitals and patients by speeding up diagnosis and possibly making it more reliable. Moreover, throughout our experience, we learned valuable skills such as data analysis, processing, scaling, and feature selection, as well as how to tune and compare various ML algorithms like SVC, RF, LR, and HVC.

As seen in Table 7, the RF model surpassed the other models with an accuracy score of 99%. The HVC came in second with an accuracy of 97%, outperforming both LR and SVC, which had scores of 95% and 93%, respectively. In terms of misclassifications, RF only had one false positive misclassification, which is not as serious as false negatives. HVC, on the other hand, finished in second again with seven misclassifications, two of which were false negatives. LR came next with 14 misclassifications, while SVM (Poly) came last with 18 misclassifications.

Our research discovered that the RF model is the best algorithm to use in predicting heart disease, when compared to earlier studies on heart disease prediction using ML algorithms, such as Singh and Kumar (2020) and Sharma (2020). This is in line with Jindal's study from 2021, which discovered that RF performed better than other models at diagnosing heart disease. Additionally, when we are building ML models, the dataset that we choose is one of the most crucial components. The models would be noticeably better at predicting disease if the training dataset was more precise, extensive, and constantly updated such as Electronic Health Records (EHRs) (HealthIT.gov, 2019). EHRs are files that are kept in many healthcare systems and contain detailed and comprehensive patient data. We believe that the models could be more accurate if they were to be trained on EHRs. Furthermore, using deep learning algorithms such as Convolutional Neural Networks or Artificial Neural Networks, could have the potential to increase prediction accuracy, particularly when working with complicated and sizable datasets like EHRs.

8 References

CDC (2023) *Leading Causes of Death Centers for Disease Control and Prevention.* 18 January 2023 [online]. Available from: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. (Accessed: April 29, 2023).

Singh, A. and Kumar, R. (2020) "Heart disease prediction using machine learning algorithms" *2020 International Conference on Electrical and Electronics Engineering (ICE3)* [Preprint]. Available at: https://doi.org/10.1109/ice348803.2020.9122958. (Accessed: April 29, 2023).

Jindal, H. *et al.* (2021) "Heart disease prediction using machine learning algorithms" *IOP Conference Series: Materials Science and Engineering*, 1022(1), p. 012072. Available at: https://doi.org/10.1088/1757-899x/1022/1/012072. (Accessed: April 29, 2023).

Sharma, Vijeta, et al. "Heart Disease Prediction Using Machine Learning Techniques." *IEEE Xplore*, 1 Dec. 2020, ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9362842. (Accessed: April 30, 2023).

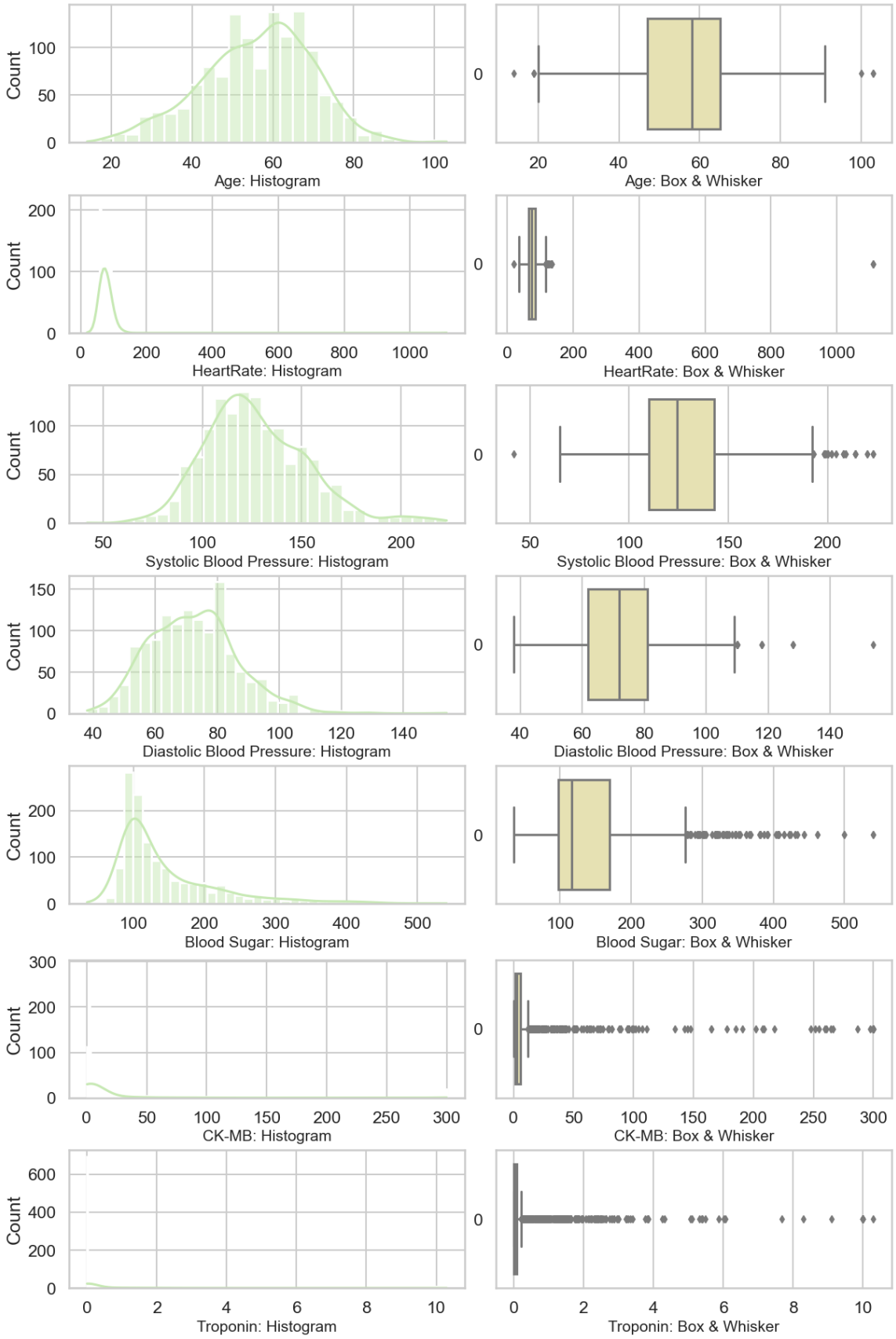
Rashid, Tarik A.; Hassan, Bryar (2022), "Heart Attack Dataset", Mendeley Data, V1, doi: 10.17632/wmhctcrt5v.1 (Accessed: April 30, 2023).

About Us (2023) *University of Kurdistan Hewlêr*. Available at: https://www.ukh.edu.krd/about-us/ (Accessed: May 1, 2023).

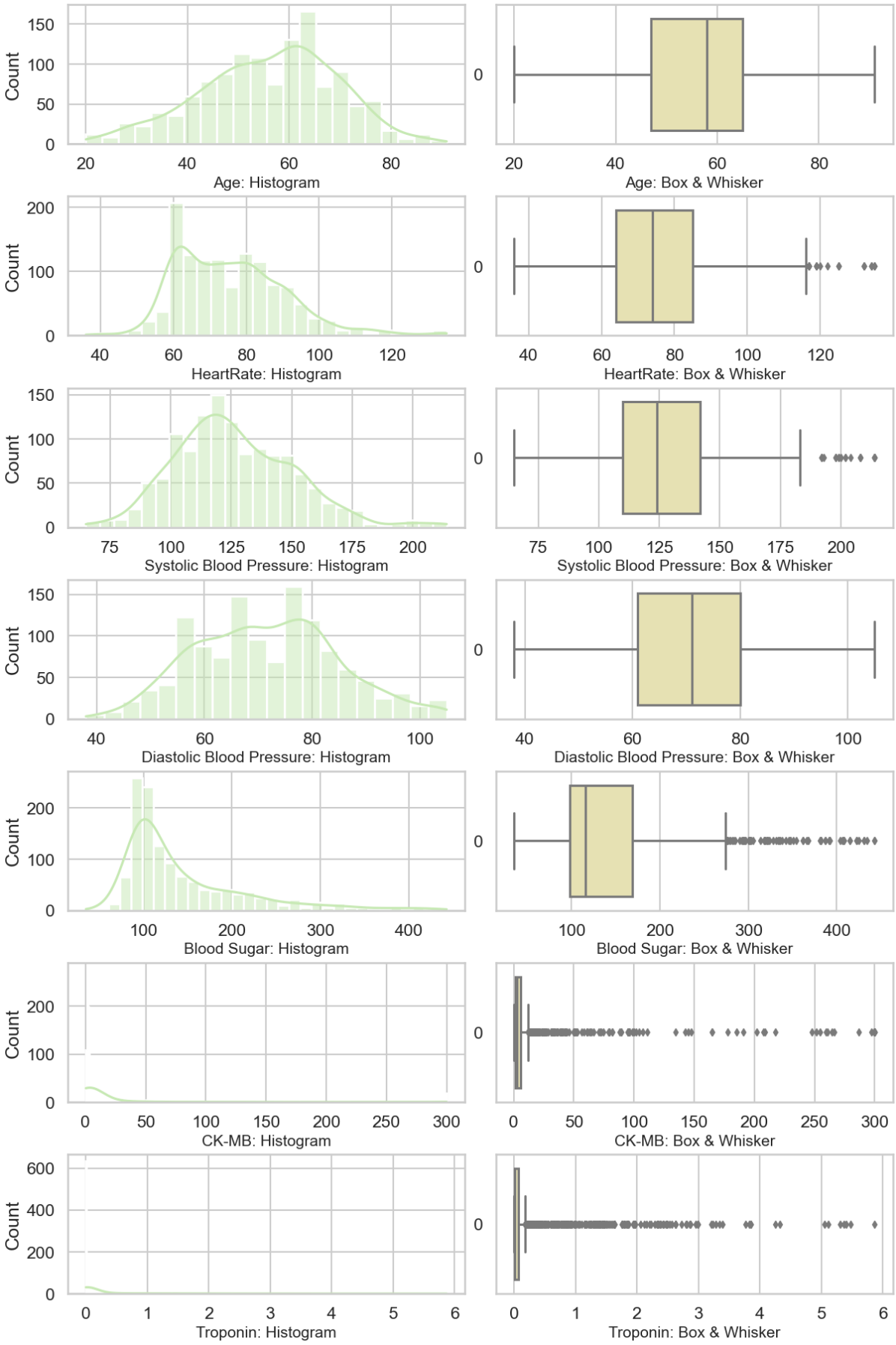
(no date) *ASIC*. Available at: https://www.asic.org.uk/ (Accessed: May 1, 2023).

What is an electronic health record (EHR)? (2019) *What is an electronic health record (EHR)? | HealthIT.gov*. Available at: https://www.healthit.gov/faq/what-electronic-health-record-ehr (Accessed: May 2, 2023)

9 Appendices



Appendix 1 Before removing outliers



Appendix 2 After removing outliers