EECS 4404/5327 Project Part 5 Report

Abstract

Our application is a basic screening heartrate abnormality detector. It determines if the individual's heartrate is normal or abnormal based on a set number of input features. The design compromises of multiple features including age, resting blood pressure, and cholesterol levels.

The techniques we have chosen are multiple logistic regression and SVM because we have defined this as a classification problem. The ML algorithm's outputs will be either a 1 or 0 indicating abnormality or normality of the individuals on the dataset.

The result will be 2 confusion matrices (1 logistic regression, the other SVM) on whether the heartrate is abnormal or normal, based on whether their heartrate falls below or above the algorithm's decision boundary.

This will all be incorporated into a GUI design. With the algorithm's predicative capability, it is possible for individuals who have been screened as having a higher chance of heartrate abnormality to seek further assistance and potential diagnoses.

Introduction

1. What is your application?

We are developing a machine learning model that predicts whether a patient is at risk of heart disease based on multiple parameters such as age, old peak (ST depression from electrocardiogram) etc.

2. What are the assumptions/scope of your project?

Due to limitations in time our project was limited to available opensource data sets through online resources. This limited our available options and our group had to settle with data that originated from different areas around the globe. Some areas of origin for the data include Switzerland and Hungary. This sets the scope and the assumption that the user is an average person living in a first-world country. Additionally, by setting such a scope we become unable to develop the model for a specific geographical region.

3. Justify why is your application important?

The importance of our application cannot be overstated as it concerns itself with the health and safety of the general public.

Life and health are very fickle matters that can abruptly change without a moment's notice. That is why it is best to know and catch issues early. With the help of our application, users can get a comprehensive guide based on their personal biodata to predict whether they are at risk of heart disease.

This application is not only more feasible compared to our initial project, but also more accessible which is inline with one of our goals of ease of accessibility. This also plays a role into the importance of our application, the ease of accessibility.

4. Similar applications

Our application is similar to existing screening questionnaires but our application is more accurate as it is tailored to the individual through the use of accurate values as well as advanced machine learning techniques/algorithms

Our application is also set apart from others as it includes a fully functioning GUI for ease of access and use towards any possible users.

5. Adjustments to part 1

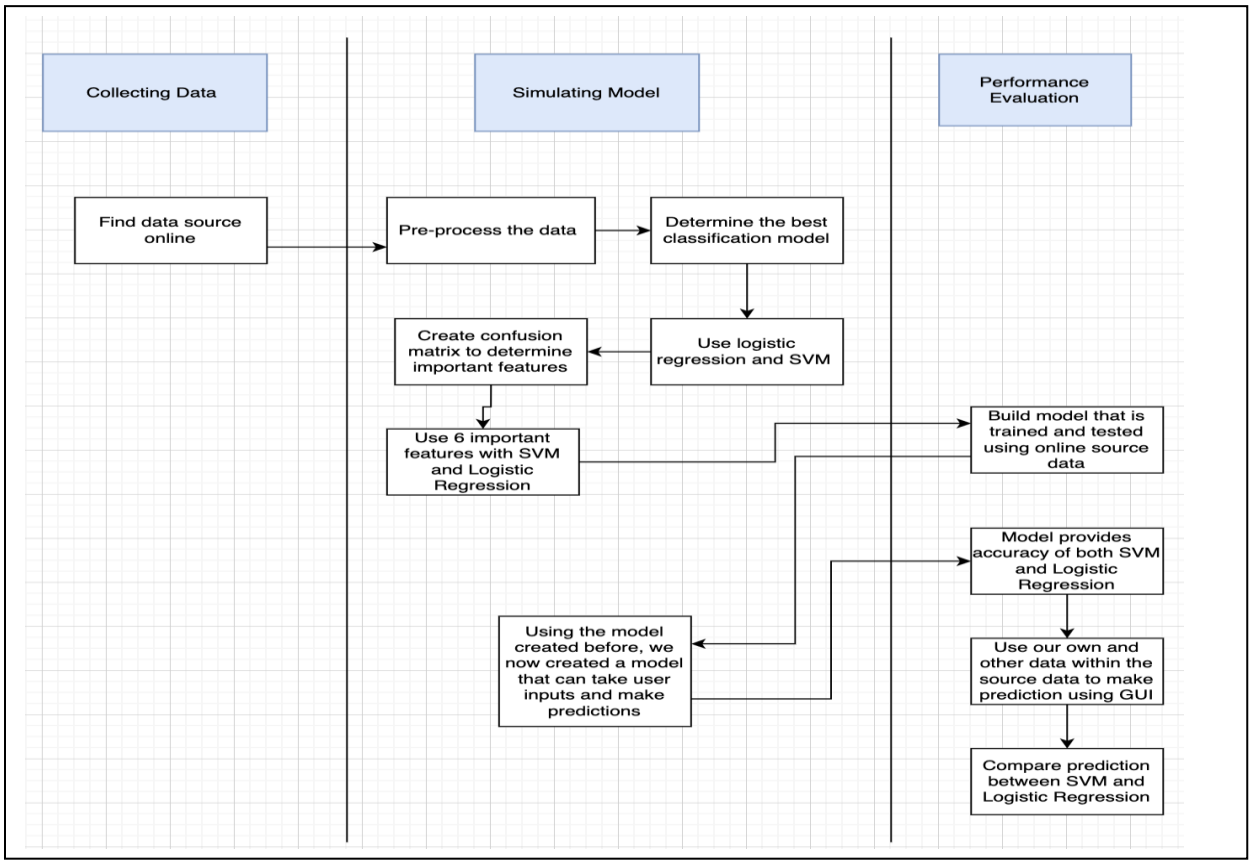
The scope of the problem has changed significantly from our initial project pitch in part 1.

In part 1, the application was supposed to be a constant heart monitor that predicted/notified the user of heart abnormality by checking for anomalies in the heart rate graph. However, our current project functions as a static heart disease risk checker through input features such as age, old peak, etc.

This change in scope and application caused our problem to categorically change from regression to classification. Thus, leading to a change of technique from multiple linear regression to logistical regression as well as SVM.

Methodology

1. Design/pipeline



2. Dataset

Our dataset is a csv file with a multitude of features including age, sex, chest pain type, and resting blood pressure. The dataset was obtained through online resources [1] after thorough research and investigation [2][3][4]. This dataset was chosen in comparison to others due to the sizable amount of accurate data with relevant supporting information.

We preprocessed and parsed through all the data to use a select number of relevant features through the use of multiple libraries. The features were chosen in advance after thorough research and investigation to determine relevant metrics pertaining to heart abnormalities [2][3][4]. The most notable libraries used were 'pandas' and 'sklearn'. The pandas library was used for importing the dataset and basic data manipulation. Multiple features from sklearn were then used to do more of the heavy lifting of preprocessing the data. These features include the SimpleImputer, ColumnTransformer and StandardScaler. The SimpleImputer library was used to fill in missing data using the mean of the relevant values. The ColumnTransformer was used as an encoder to the pandas import. The StandardScaler was used to standardize all the data for ease of use within the training and models. These processes were all crucial for different reasons but ultimately it was to standardize the data and allow the data to smoothly be used to train the models without much hinderances.

3. Model training

The machine learning technique that was ultimately used was the multiple logistical regression and SVM technique. Due to changes within the scope and goal of the project the technique used shifted away from the original multiple linear regression technique as it no longer suited the project. This was due to the very nature of the project categorically changing from what was previously perceived as a regression problem to a classification problem. The SVM technique was used as verification and validation towards the multiple logistical regression model.

The inputs are age, resting blood pressure, cholesterol, blood sugar when fasting, maximum heartrate, and old peak.

The output is the heart disease prediction.

The decision to change came about after much deliberation to create a more achievable project that further falls inline with the goals of the project. Project part 2 did not have much of an influence on the overall project.

4. Prediction

The model predicts the output from the information extracted from the GUI. The user inputs the features in questions into the text fields within the GUI which are then extracted and inputted into the model to make a prediction.

Results

1. Evaluation

Group Survey:

On a scale of 1-10 how satisfied were you with the product?

Jethro: 7

Eric: 8

Inderpreet: 8

Chaudry: 7

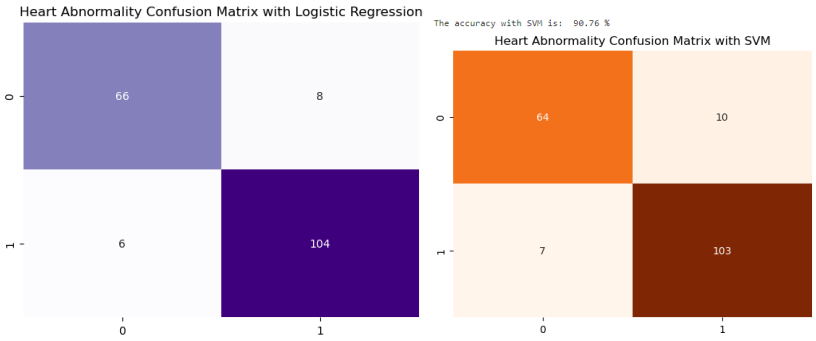
Was the prediction accurate?

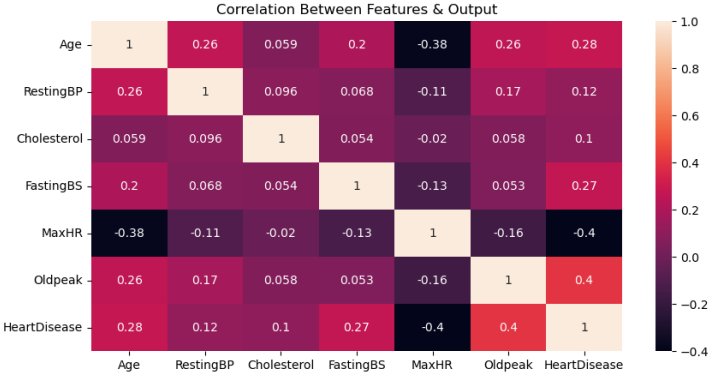
Overall, we all agreed the prediction was accurate as both logistic regression and SVM had a over

90% accuracy rating.

Survey END.

The data is going to be analyzed and represented through the use of a graph to visualize the feedback information. Dataset is parsed and separated into tuples using a column transformer. Our multiple logistical regression model was about 1.63% higher than the trial SVM model. The metric used is accuracy. We have included pictures of the GUI near the bottom of the report (due to box formatting issues).





2. Results

   The current results of our algorithm are formatted to display a confusion matrix, depicting the number of false positives, false negatives, true positives and true negatives. This information provides great insight to how correct our information is, how we can take a step back and improve the accuracy of our classification model.

   Also, we create an output of a correlation matrix to help us understand the affects of the features we have chosen to help build on our accuracy. From the results, we see that MaxHR is closer to negative 1 when correlated to heart disease, and we can take this information and improve algorithm in the future steps.

Discussions

1. Implications

The results obtained from our model demonstrate a commendable accuracy rate, with its efficacy in identifying heart abnormalities. With a low number of false positives and negatives, the model proves to be a reliable resource that has strong potential for real-life settings.

The key factors crucial to the project's success rate include precise feature selection and a meticulously balanced dataset. The feature selection optimized the predictive capabilities of the model, while the dataset reinforced its learning and ability to induce findings with high accuracy.

2. Strengths

The strength of our algorithm is the ability to use multiple different methods. Currently we are using Linear SVM and Logistic Regression. This proves to be useful because we know in machine learning no one algorithm is obsolete. The use of multiple algorithms on one data set can show how each method interprets the data and can provide different results. This in turn can help our situation where the dataset consists of many features.

Additionally, the 2 algorithms also work very well for the binary classification task that we have: whether a patient has heart abnormality (1) or not (0).

3. Limitations

The limitation within our designs lies within the features that are required in order to get a truly accurate screening. Some of these features/inputs are also not readily available information that the general public has (i.e., their cholesterol levels)

Although our model has many features, there are likewise just as many that are missing due to insufficient data (especially to unquantifiable data) and due to project time constraints.

4. Future directions

   In the future we would like to pursue adding more features to this model to fine tune the model and improve its accuracy. We did not have enough time to collect data online and thus training and testing was done on a limited set of data. Though our model was 92.39% accurate we expect much more data for training before the model would achieve high level accuracy in real life.
   Alternatively, we would also like to pursue other classification techniques.
   We are also interested in whether this application has potential for commercial use. By using it commercially we can possibly use it for collecting data from a wider variety of sources and perform additional training of the model. We can also perform data analytics of heart disease based on different geographical locations, ethnicities, regions etc. However, this can lead to concerns such as privacy and cost.

Additional Questions

1. What are the feedbacks that you found useful from the peer evaluation?

We received much feedback that we deemed useful. The feedback mainly pertained to including more data and more features such as cholesterol, exercise levels, smoking status etc. as well as better explaining the data preprocessing step. Other feedback included adding more visuals and diagrams to show results of the process.

Among the feedback received the points that were addressed were including visuals as well as the feedback to better explain the preprocessing step. Finding more data especially with additional features remained elusive due to time constraints and limited resources.

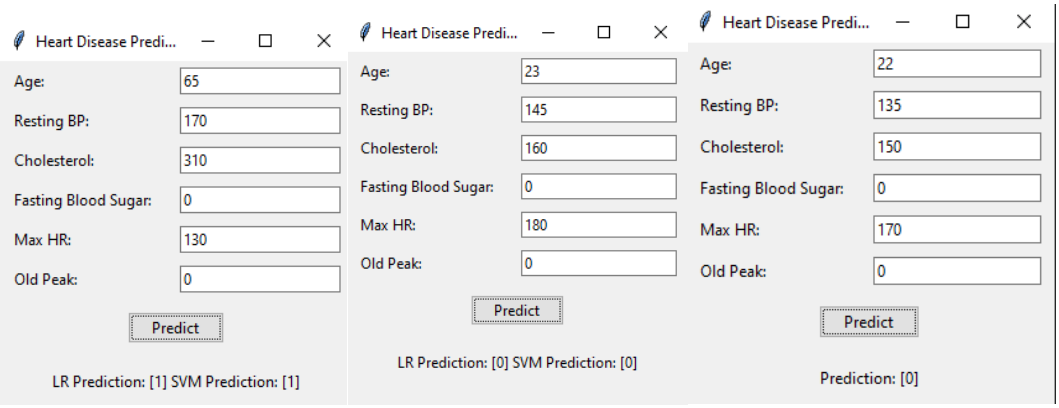
2. What changes did you made based on the feedback from peer evaluation?

The changes made included added visuals for better understanding of the project as well as more in depth explanation in regards to processing the data.

In accordance with the feedback the design/pipeline was changed into a diagram to make the information more digestible. Visuals were also added into the report for better understanding and proof. These visuals include the confusion matrices.

The process of preprocessing the data was better explained to give a more in-depth understanding of the process. The imputer, transformer, as well as the scaler were mentioned and explained. The reason for choosing the features was also explained.

Pictures of GUI:



References

1. [reference 1] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data
2. [reference 2] https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979
3. [reference 3] https://theheartfoundation.org/2018/11/02/your-heart-rate/
4. [reference 4] https://www.canada.ca/en/public-health/services/diseases/heart-health/heart-diseases-conditions/prevention-heart-diseases-conditions.html
5. [reference 5] https://pandas.pydata.org/docs/getting_started/index.html
6. [reference 6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
7. [reference 7] https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html#:~:text=Applies%20transformers%20to%20columns%20of,form%20a%20single%20feature%20space.
8. [reference 8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
9. [reference 9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
10. [reference 10] https://scikit-learn.org/stable/modules/svm.html
11. [reference 11] https://scikit-learn.org/stable/modules/model_evaluation.html
12. [reference 12] https://scikit-learn.org/stable/model_selection.html
13. [reference 13] https://seaborn.pydata.org/tutorial/introduction.html
14. [reference 14] https://scikit-learn.org/stable/index.html