Cardiac Analysis

**CISC 5800**

**Machine Learning**

**Fordham University**

**Juan “River” Reyes-Veloz, Ciara Serpa**

**Introduction**

Heart illness has received much interest in medical research, among other life-threatening illnesses. According to the CDC, About 659,000 people in the United States die from heart disease each year. Identifying heart illness is a difficult task that can provide an automated prognosis of the patient's heart health, allowing for more successful therapy. The signs, symptoms, and physical exam of the patient are commonly used to diagnose cardiac disease. Factors like smoking, high cholesterol, a family history of cardiovascular disease, obesity, hypertension, and a lack of physical activity are higher risks of heart disease. By accurately diagnosing these patients early on, physicians are more adequately able to provide their patients with targeted and aggressive treatments and lifestyle changes.

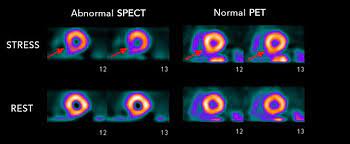
Providing high-quality services at reasonable prices is a crucial problem for healthcare institutions like hospitals and medical clinics. Quality service includes correctly diagnosing patients and providing appropriate therapies. There are both categorical and numerical data in the accessible heart disease database. These records are cleaned and screened before further processing to remove extraneous data from a database.  The suggested method can extract exact hidden information from a historic heart illness database, i.e., patterns and associations related to heart disease.

There are several open sources for gaining access to patient information. Studies may be undertaken to see if various computer technologies can be utilized to correctly diagnose people and detect this condition before it becomes fatal. Machine learning is now widely acknowledged to have a significant role in the healthcare system. To analyze the state and categories or forecast the outcomes, we may utilize a variety of machine learning models for better results.

**Dataset**

This study was conducted using a publicly accessible dataset for heart disease from the University of California Irvine. The dataset describes diagnosing cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature patterns were created for each patient. The CLIP3 algorithm was used to generate classification rules from these patterns. The CLIP3 algorithm generated rules that were 77.0% accurate (as compared with cardiologists' diagnoses). There are 265 entries in the dataset, separated into two groups: training (70%) and testing (30%). Which has 44 features, column vectors, and one column for labels that either the heat is normal or abnormal. Some machine learning methods were also employed to train and assess the data.

A SPECT scan is a form of medical imaging that shows how blood flows to tissues and organs. This allows it to be used to diagnose medical issues like seizures, spinal tumors, stress fractures and strokes. When used specifically for heart disease, it locates areas of the heart muscle that have inadequate blood flow compared with areas that have normal flow. Inadequate blood flow may mean that coronary arteries are narrowed or that a heart attack has occurred.



Understanding how SPECT imaging and examinations work allows computer programmers and data scientists to better understand how to use and manipulate the data in order to find patterns and relationships.

**Methodology**

In this project, we used different machine learning models for prediction and training. The two most successful were Naive-Bayes Classifier and Support Vector Machine (also referred to as SVM)

1. **Support Vector Machine**

The Support Vector Machine, or SVM, is a common Supervised Learning technique that may be used to solve both classification and regression issues. However, it is mainly utilized in Machine Learning for classification challenges.

The SVM algorithm aims to find the optimum decision boundary for categorizing n-dimensional spaces into categories so that additional data points may be readily placed in the proper category in the next. A hyperplane is a name for the optimal choice boundary.

The extreme vectors that assist in creating the hyperplane are chosen via SVM. Support vectors are extreme instances, and the method is called a Support Vector Machine. Consider the picture below, which shows how a decision boundary is used to classify two separate groups.

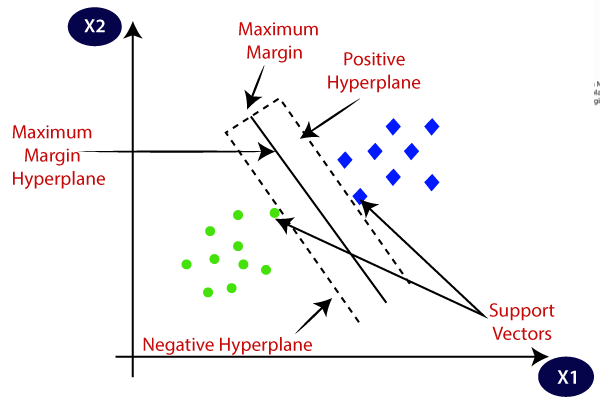


Figure 01: SVM representation

1. **Naïve Bayes classifier**

The Bayes Theorem-based probabilistic machine learning method Naive Bayes is employed in a wide range of classification problems. The Naive Bayes method is a statistical machine learning technique used for a wide range of classification applications. Filtering spam, categorizing documents, and predicting sentiment are common uses. The name is derived from the work of Rev. Thomas Bayes. And why is it referred to as 'Naive'? The term naïve refers to the assumption that the characteristics that make up the model are unrelated to one another. Changing one variable has no direct impact on the value of the other elements employed in the algorithm. Naive Bayes appears to be a simple, robust algorithm. Then why is it well-liked? This is since NB has a considerable edge. Because that is a probabilistic model the Programme may be quickly developed and predictions generated.

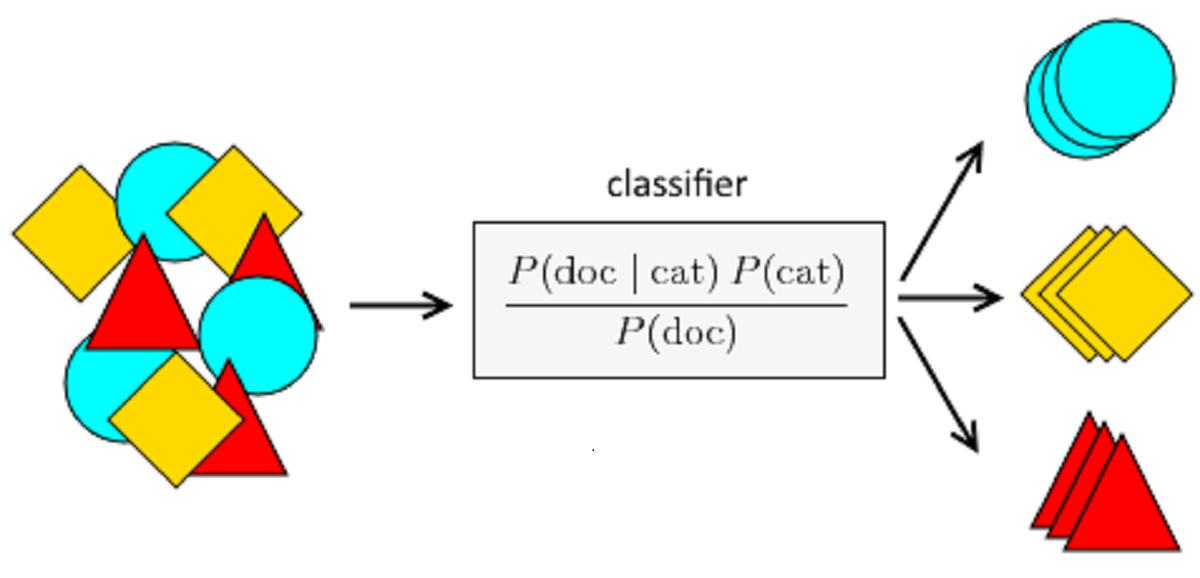


Figure 02: Naïve Bayes classifier

Implementations of both models are given below.

1. Naïve Bayes

First of all the dataset was imported and then cleaning the data was done to make it suitable for the machine learning model, which is Naïve Bayes. The whole process can be seen in the following figure.





Figure 03: Implementation for Naïve Bayes

As seen in the above figure, the input data comes for preprocessing. The dataset checks to remove the null values and then sends it to the classifier, which is Naïve Bayes; in Naïve Bayes, we keep all the values as default. After fitting it on the train data, the model checks on the test data, and the final prediction comes, which is either the class belongs to Normal or Abnormal hearts.

1. Support vector Machine

For the support vector machine classifier, the whole process is shown in figure 04. Data come for preprocessing. After cleaning the data, it passes through the SVM for classification and prediction in binary classes, which are Normal and Abnormal heart.



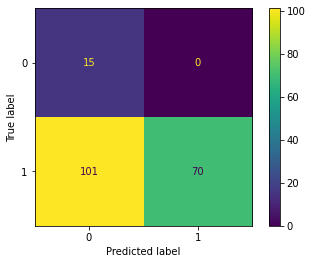
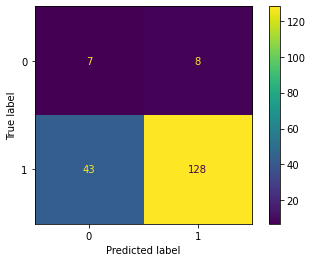
Figure 04: Classification process for SVM

For Svm, we further expand our experiment to take some other argument values. We took three combinations of arguments which are as below:

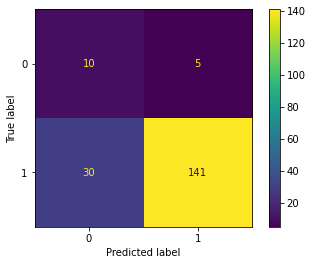
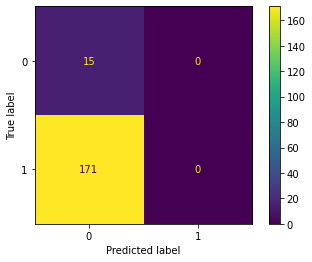
1. When the kernel value is linear, and the C value is 10,
2. When the kernel value is RBF, C value is 940, and gamma value is 0.004,
3. When the kernel is Gaussian, C is 940 and gamma is 0.004

**Results**

This section includes the results for the random forest and SVM with different arguments. To check the model performance, the Confusion matrix is an important metric. A confusion matrix is an evaluation metric used to check the model performance. It has different values of true positive, true negative, false positive, and false negative. Confusion matrices for both models are as follows.

1. (c)

1. (d)

Figure 04: (a) For Naïve Bayes while (b)(c) and (d) for SVM with different C and kernel values.

Apart from the confusion matrix to check the model's performance, there is a need to check the classification reports. A classification report is a built-in function in sklearn with different metrics such as precision-recall and F1 score. Let’s have a brief discussion about them.

**The precision**

Precision metric is the ratio of True Positives to all Positives in its most basic form. That'd be the percentage of patients we properly identify as having heart problems out of all who have it for our issue statement.

Precision = True Positive/ True positive + False positive

**Recall**

The recall measures how well our model detects True Positives. As a result, recall informs us how many patients we accurately recognized as having heart disease out of all those who have it.

Recall = True Positive/ True Positive + False positive

**Accuracy**

Now, we'll look at one of the most basic measures: Accuracy. The ratio of the overall number of right forecasts to the overall number of predictions is known as accuracy.

Accuracy= True Positive + True negative/ True positive +False positive+ True negative +False negative

**F1 score**

This F1 score is the harmonic mean between recall and precision values and the formula is given below

F1 score= 2\* Precision \* Recall/Precision + Recall

These values for each model can be seen in the following table:

Table 01: Evaluation Parameters values for different models

| S.NO | Models | Average Precision | Average Recall | Average  F1 score | Accuracy |
| --- | --- | --- | --- | --- | --- |
| 1 | Naïve Bayes | 56% | 70% | 52% | 46% |
| 2 | SVM (1) | 54% | 61% | 52% | 73% |
| 3 | SVM (2) | 61% | 74% | 62% | 81% |
| 4 | SVM(3) | 0.08% | 1% | 15% | 0.08% |

As can be seen in the above table we took the average value of each parameter and with SVM we put 1 2 and 3. Where 1 mean version 1, 2 mean version 2, and 3 mean version 3. These versions are different because of different C and kernel values, which are shown in the code section in the Jupyter notebook.

**Conclusion**

From the results, it can be seen that the SVM with version 2 has the best accuracy in all models, so this model can be used for further deployment. By choosing this deployment method, many patients could be assisted in their treatment plan.

Works Cited

Centers for Disease Control and Prevention. [Underlying Cause of Death, 1999–2018](http://wonder.cdc.gov/ucd-icd10.html). CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.

Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. [Heart disease and stroke statistics—2021 update: a report from the American Heart Association](https://doi.org/10.1161/CIR.0000000000000950)

[external icon](https://doi.org/10.1161/CIR.0000000000000950). *Circulation.* 2021;143:e254–e743.

Fryar CD, Chen T-C, Li X. [Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010[PDF-494K]](https://www.cdc.gov/nchs/data/databriefs/db103.pdf). NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.

Centers for Disease Control and Prevention, National Center for Health Statistics. [About Multiple Cause of Death, 1999–2019](https://wonder.cdc.gov/mcd-icd10.html). CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2019. Accessed February 1, 2021.

Heron, M. [Deaths: Leading causes for 2017](https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_06-508.pdf). National Vital Statistics Reports;68(6). Accessed November 19, 2019.

*SPECTF Heart Data Set*. UCI Machine Learning Repository: SPECTF heart data set. (n.d.). Retrieved May 5, 2022, from http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart