Using Machine Learning for Disease Diagnosis

Eren Kızılırmak 210704025

Abstract

Early and accurate diagnosis of diseases is important for effective treatment. Machine learning (ML) offers the potential to improve the speed and accuracy of disease diagnosis by analyzing data from patients and hospitals, such as medical history, symptoms, and lab results. I've utilized 7 machine learning models to evaluate their performances. The performance metrics contain two criteria ROC-AUC and cross validation score. Based on the experimental results, the conclusion is that the SVC and XGB classifiers were the best overall among the used dataset and can be used for diagnosing heart disease.

Keywords

Artificial Intelligence, Machine Learning, Medical Diseases, Performance

1. Introduction

Machine learning (ML) is one of the main fields of artificial intelligence (AI) in which different areas like statistics, computer science, and mathematics are integrated to propose intelligent systems that mimic human behaviors (D. Jain & Singh, 2020; Liu et al., 2019). This integration is used to build and extract a model by learning from the past experiences data (large and complex datasets) to predict feature data. Such a model may mimic the human behaviors analog to the human brain neurons and human decision-making rules (Harper, 2005; D. Jain & Singh, 2020). Using ML in this sector will enhance the health and welfare of the people and will help in decreasing the error rates in medical diagnosis and procedures. The ML is usually integrated with medical device systems to help them enhancing their diagnosis results and increasing stability and accuracy (Abbas et al., 2020).

2. Literature Review

In this work, Heart disease classification will be addressed. This disease has been worked on previously by many researchers. The related works on this disease and dataset is reviewed in below ;

(Kahtan et al., 2018) designed a fuzzy, bunch of if statements, logic system for the detection and the diagnosis of the heart disease. Five variables were selected, which are age of patients, blood pressure, cholesterol and blood sugar. The system was implemented using Java.

(Nassif et al., 2018) implemented SVM, Naive Bayes and KNN algorithms and used feature selection techniques to detect Heart Disease. This showed SVM was best in terms of accuracy, precision, recall and specificity.

(Sundaraman et al., 2015) compared Naive Bayes, J48 decision trees and an artificial neural network(ANN) and saw that the Naive Bayes algorithm performed the best.

However, none of these previous work have used XGB and implemented PCA for feature generation. Also the classification algorithms used in each did not used ROC-AUC (Receiver operating characteristic (ROC)’s area under the curve) for metric.

The results presented in this paper show how feature engineering with PCA combined with feature set provided an improvement in model performance and XGB performs the best.

3. Material

3.1 Dataset

The dataset used in this study was provided by Janosi et al. (1988). The original Cleveland dataset contains 303 rows and 14 columns. There are no null nor duplicate rows.

Table 1. Dataset attributes description with their values

Attributes	Values and Description
Age	29 - 77
Sex	1=Male; 0=Female
cp *(Type of chest pain experienced by patient)*	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain
trestbps *(patient's level of blood pressure at resting)*	94 – 200 (mm/Hg)
chol *(Serum cholesterol)*	126 – 564 (mg/dl)
fbs *(Blood sugar levels on fasting > 120)*	1: true, 0: false
estecg *(Resting ECG)*	Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach *(Maximum heart rate)*	71 – 202
exang *(Exercise induced angina)*	1 = yes; 0 = no
oldpeak *(ST depression induced by exercise relative to rest)*	0 – 6.2
slope *(The slope of the peak exercise ST segment)*	Value 1: up-sloping Value 2: flat Value 3: down-sloping
ca *(The number of major vessels)*	0 – 4
thal *(A blood disorder called thalassemia)*	3 = normal; 6 = fixed defect; 7 = reversible defect
target *(Heart disease diagnosis)*	0 = Absence 1 = Precense

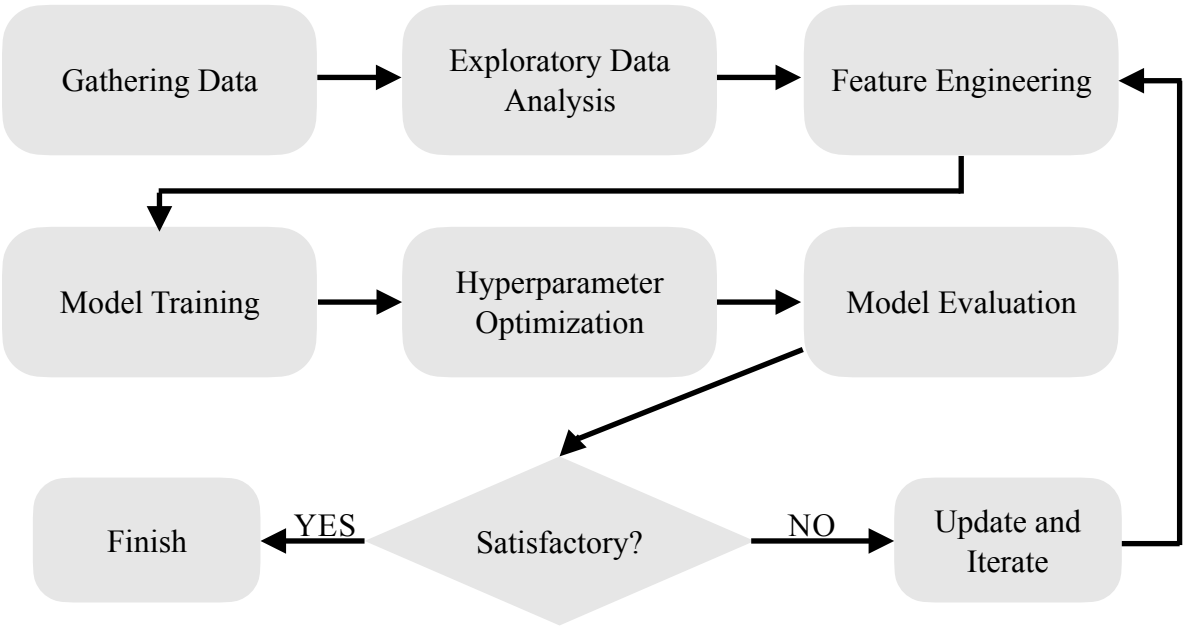
4. Methodology

Methodology for this project is shown briefly in the diagram below the section;

4.1 Gathering Data

Data has been gathered from UCI Machine Learning Repository (Janosi et al., 1988). through Google Dataset Search (*Dataset Search*, n.d.).

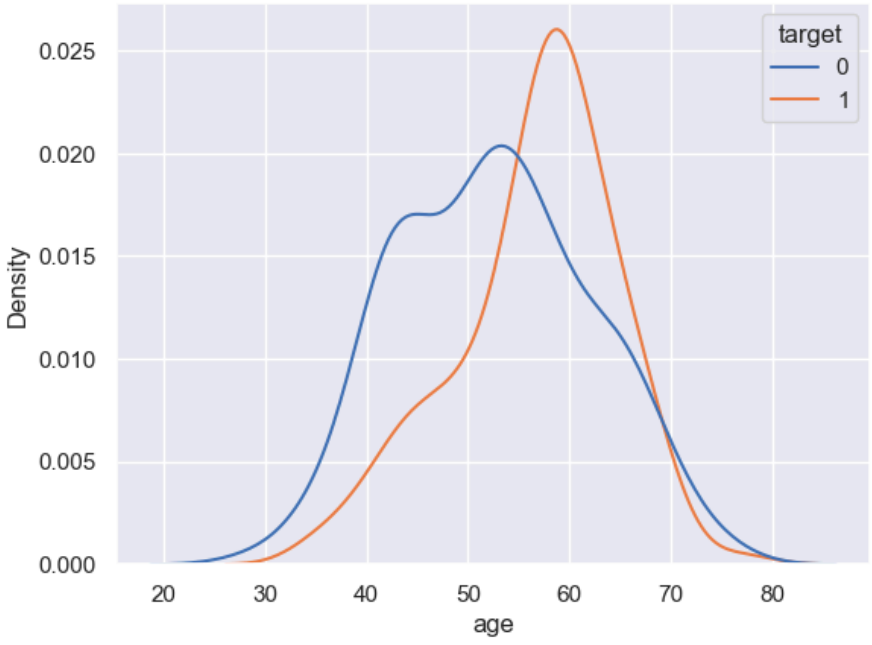
Figure 1. Methodology Diagram



4.2 Exploratory Data Analysis

In this part, data has been analyzed in terms of null variables, duplicates, count, outliers, density, and distribution comparison.

Figure 2. Density Diagram



for example from figure 2, it is seen that old people tend to have heart attacks more often than the average.

All features have been analyzed and visualized and some of them explained. Visualizations and code can be found on the attached file.

4.3 Feature Engineering

Binning is used as a technique for feature engineering which is similar to Fuzzification (Kahtan et al., 2018) and was effective in that paper. Binning is firstly used for age, trestbps and chol features. But, binning trsetbps and chol features was just lowering models' performance drastically. Thus, binned age feature was decided to be kept.

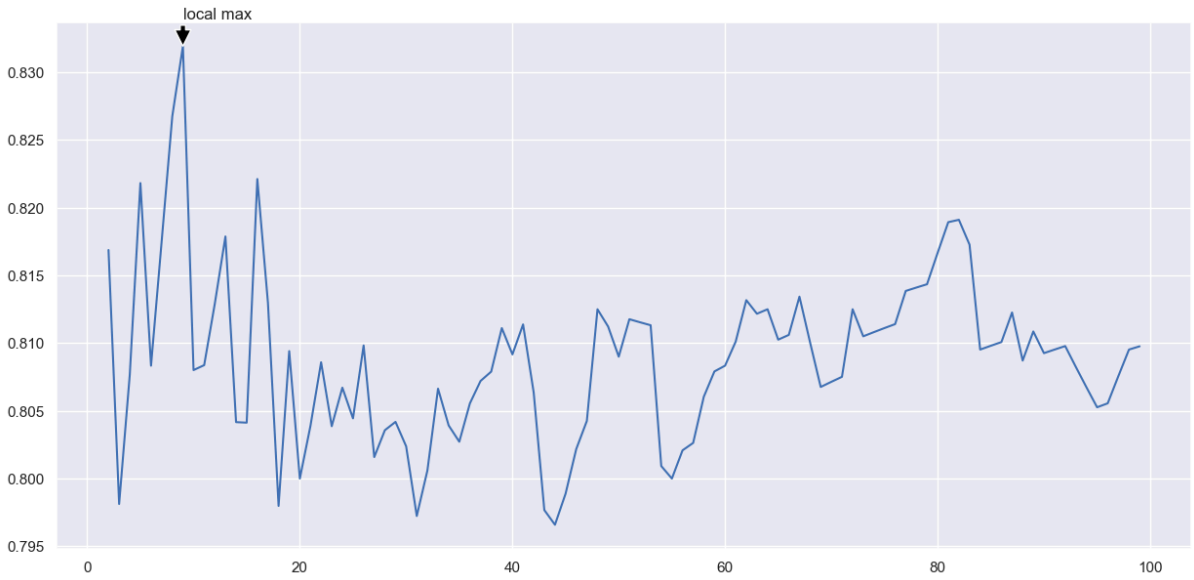
Then, Principal Component Analysis (PCA) reduced dimensionality of the data in 5 projections and is combined with the pre-PCA (before component analysis) data.

Finally, Encoding and Scaling is applied.

4.4 Model Training

SVM, ANN, DTC, RFC, XGB, KNN, ADA-BOOST models are trained on the preprocessed data. Their cross validation scores are measured and the number of KFolds required in each is determined for best score.

Figure 3. Best Kfold split graph for SVM



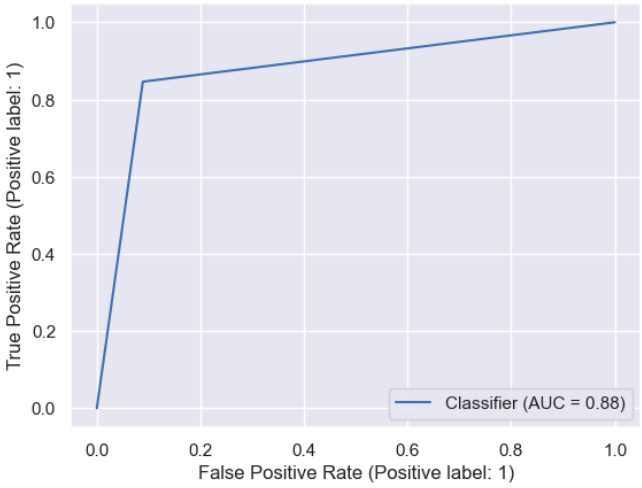
Results

The models are evaluated by ROC-AUC metric and cross validation score. My results cannot be compared to the other works' results because ROC-AUC wasn't calculated.

Table 2. Result Table

MODEL	CROSS VALIDATION	ROC-AUC
SVM	0.833	0.840
ANN	0.816	-
DTC	0.769	0.746
RFC	0.833	0.829
XGB	0.825	0.878
KNN	0.805	0.798
ADA-BOOST	0.819	0.820

Figure 4. ROC-AUC graph of XGB



Conclusion

CHD is a major cause of death worldwide, including Turkey. This project proposed that XGB model enhanced with PCA combined feature set is useful to diagnose the presence of CHD and help patients get the treatment immediately to longer their lifespan. Moreover, implementation of machine learning in health care is needed especially considering the human made errors and advances in medical technology.

My approach has it's own weaknesses one of them is, PCA is already used to reduce dimensionality to reduce overfitting but if i combine with more features, overfitting may appear because data will have more columns than it already had.

Potential direction for future research is the investigation of alternative dimensionality reduction techniques other than PCA which may mitigate the risk of overfitting when combining multiple features.

Appendix

Attached code and visualizations with explanations to the zip file. file's name is main_project.ipynb. Jupyter notebook needs to be installed to view it. for convenience, i've uploaded it on GitHub https://github.com/eren9677/case_studies/blob/main/main_project.ipynb .

References

Kahtan, H., Zamli, K. Z., Fatthi, W. N. A. W. A., Abdullah, A., Abdulleteef, M., & Kamarulzaman, N. S. (2018). Heart Disease Diagnosis System Using Fuzzy Logic. *Proceedings of the 2018 7th International Conference on Software and Computer Applications*. https://doi.org/10.1145/3185089.3185118

Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. *UCI Machine Learning Repository*. https://doi.org/10.24432/C52P4X

Jain, D., & Singh, V. (2020). A novel hybrid approach for chronic disease classification. *International Journal of Healthcare Information Systems and Informatics*, *15*(1), 1–19. doi:10.4018/IJHISI.2020010101

Harper, P. R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy (Amsterdam)*, *71*(3), 315–331. doi:10.1016/j.healthpol.2004.05.002 PMID:15694499

Abbas, S. A., Rehman, A. U., Majeed, F., Majid, A., Malik, M. S. A., Kazmi, Z. H., & Zafar, S. (2020). Performance Analysis of Classification Algorithms on Birth Dataset. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 102146–102154. doi:10.1109/ACCESS.2020.2999899

Nassif, A. B., Mahdi, O., Nasir, Q., Talib, M. A., & Azzeh, M. (2018). Machine Learning Classifications of Coronary Artery Disease. 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). doi:10.1109/isai-nlp.2018.8692942

Sundaraman, S., & Kakade, S. CLINICAL DECISION SUPPORT FOR HEART DISEASE USING PREDICTIVE MODELS.

Dataset search. (n.d.) Google Dataset Search https://datasetsearch.research.google.com