CSE 841 Project Proposal
Heart Disease Prediction and Factors Analysis

Asadullah Hill Galib

2 November 2020

1 Overview

This project aims at predicting heart disease effectively with consideration of performance measures and significant factors/attributes analysis. So, it addresses two aspects: how well can we predict heart disease and how the factors influence heart disease prediction. Several machine learning models and data mining techniques will be employed here for prediction and feature analysis. UCI repository: Cleveland database will be used here.

2 Data

The Cleveland dataset (UCI, 1990) will be used in this study which is a public dataset and can be found in the University of California Irvine (UCI) Machine Learning Repository [1].

The dataset contains 303 instances of real patient data. There are 14 attributes:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar ¿ 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest

11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
14. target: 0= less chance of heart attack 1= more chance of heart attack

3 Scientific Question

This project addresses the following scientific questions:

1. How well can we predict heart disease using machine learning techniques?
2. How the factors/attributes influence heart disease prediction?

Several prior works address these scientific questions. Nahar et al. [2, 3] investigate some computational intelligence techniques in the detection of heart disease using six well-known classifiers for the Cleveland dataset. It identifies heart disease risk factors. Medhekar et al. [4] use the Naive Bayes classifier in heart disease prediction. Batii et al. [5, 6] propose a Hybrid Naïve Possibilistic Classifier (HNPC) for heart disease detection from the heterogeneous data. Kumar et al. [7] predict heart disease using an advanced fuzzy resolution mechanism. Shah et al. [8] extract high impact features using Probabilistic Principal Component Analysis (PPCA) for heart disease prediction. Amin et al. [9] identify significant features for heart disease prediction using data mining techniques. Uyar et al. [10], Fida et al. [11], and Gokulnath et al. [12] employ a genetic algorithm-based technique for prediction. Rani et al. [13] use neural networks for heart disease prediction.

All of the research works addressed the first scientific question. Nahar et al. [2, 3], Shah et al. [8], and Gokulnath et al. [12] also addressed the second scientific question. All the mentioned works use the Cleveland dataset.

4 Implementation

This project will be implemented in Python on the Jupyter Notebook platform.

Several machine learning techniques like Logistic Regression, Multiple Linear Regression, Decision Tree, Random Forest, KNN, Support Vector Machine, etc. will be employed in this project for predicting heart disease. For analyzing the influences of factors/attributes several feature selection and data mining techniques will be used, like principal component analysis (PCA), mutual information gain, tree-based feature importance, recursive feature elimination, correlation, ANOVA test, etc.

The author of this project also wants to employ the genetic algorithm in any aspect of this project (like in feature analysis probably). Existing research [10, 11, 12] motivates him in doing so. However, he is not sure about how the

genetic algorithm can be used in this context. Also, the author also wants to employ neural network or ensemble learning-based techniques if there is plenty of time. This last implementation planning is optional for now.

5 Optional Reading: Other Project Ideas

Primarily, the author has finalized the project after analyzing the following ideas. Those ideas are also being attached here for future convenience.

5.1 N-Queen Problem using Genetic Algorithm, Hill Climbing, and K-Beams

- **Data:** Data can be simulated here
- **Scientific Question:** How do the local search heuristics perform in N-queen problem?
  - Several publications addressed this question.
- **Implementation:** Genetic Algorithm, Hill Climbing (with and without Random Start), and K-Beams will be evaluated.

5.2 Predicting Medical Insurance Cost

- **Data:** A public dataset (found in Kaggle). There are 7 attributes:
  1. age: age of primary beneficiary
  2. sex: insurance contractor gender, female, male
  3. bmi: Body mass index, providing an understanding of body
  4. children: Number of children covered by health insurance / Number of dependents
  5. smoker: Smoking
  6. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest
  7. charges: medical cost
- **Scientific Question:** How can we predict medical insurance cost using machine learning techniques?
  - Tkachenko et al. [14] addressed this question.
- **Implementation:** Logistic Regression, Multiple Linear Regression, Decision Tree, Random Forest, KNN, Support Vector Machine, etc. will be evaluated.

5.3 Classifying and/or Clustering Amazon Best Selling Books

- **Data:** A public dataset (found in Kaggle). The dataset includes Amazon's Top 50 bestselling books from 2009 to 2019. There are 550 instances and 7 attributes:

  1. Name
  2. Author
  3. Review Count
  4. User Rating
  5. Price
  6. Year
  7. Genre

- **Scientific Question:** How can we classify and/or cluster Amazon best selling books using machine learning techniques?

  - Maity et al. [15] addressed this question.

- **Implementation:** Logistic Regression, Multiple Linear Regression, Decision Tree, Random Forest, KNN, Support Vector Machine, etc. will be evaluated.

References

[1] Dua, D., Graff, C., and Detrano, R. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Nahar, J., Imam, T., Tickle, K. S., Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96-104.

[3] Nahar, J., Imam, T., Tickle, K. S., Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093.

[4] Medhekar, D. S., Bote, M. P., Deshmukh, S. D. (2013). Heart disease prediction system using naive Bayes. Int. J. Enhanced Res. Sci. Technol. Eng, 2(3).

[5] Baati, K., Hamdani, T. M., Alimi, A. M. (2014, August). A modified hybrid naive possibilistic classifier for heart disease detection from heterogeneous medical data. In 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR) (pp. 353-358). IEEE.

[6] Baati, K., Hamdani, T. M., Alimi, A. M. (2013, December). Hybrid naive possibilistic classifier for heart disease detection from heterogeneous medical data. In 13th International Conference on Hybrid Intelligent Systems (HIS 2013) (pp. 234-239). IEEE.

[7] Kumar, A. S. (2013). Diagnosis of heart disease using Advanced Fuzzy resolution Mechanism. International Journal of Science and Applied Information Technology, 2(2), 22-30.

[8] Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., Hussain, S. A. (2017). Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis. Physica A: Statistical Mechanics and its Applications, 482, 796-807.

[9] Amin, M. S., Chiam, Y. K., Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[10] Uyar, K., İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science, 120, 588-593.

[11] Fida, B., Nazir, M., Naveed, N., Akram, S. (2011, December). Heart disease classification ensemble optimization using genetic algorithm. In 2011 IEEE 14th International Multitopic Conference (pp. 19-24). Ieee.

[12] Gokulnath, C. B., Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 22(6), 14777-14787.

[13] Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626.

[14] Tkachenko, R., Izonin, I., Kryvinska, N., Chopyak, V., Lotoshynska, N., Danylyuk, D. (2018). Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure. In IDDM (pp. 170-179).

[15] Maity, S. K., Panigrahi, A., Mukherjee, A. (2018, August). Analyzing social book reading behavior on goodreads and how it predicts amazon best sellers. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 211-235). Springer, Cham.