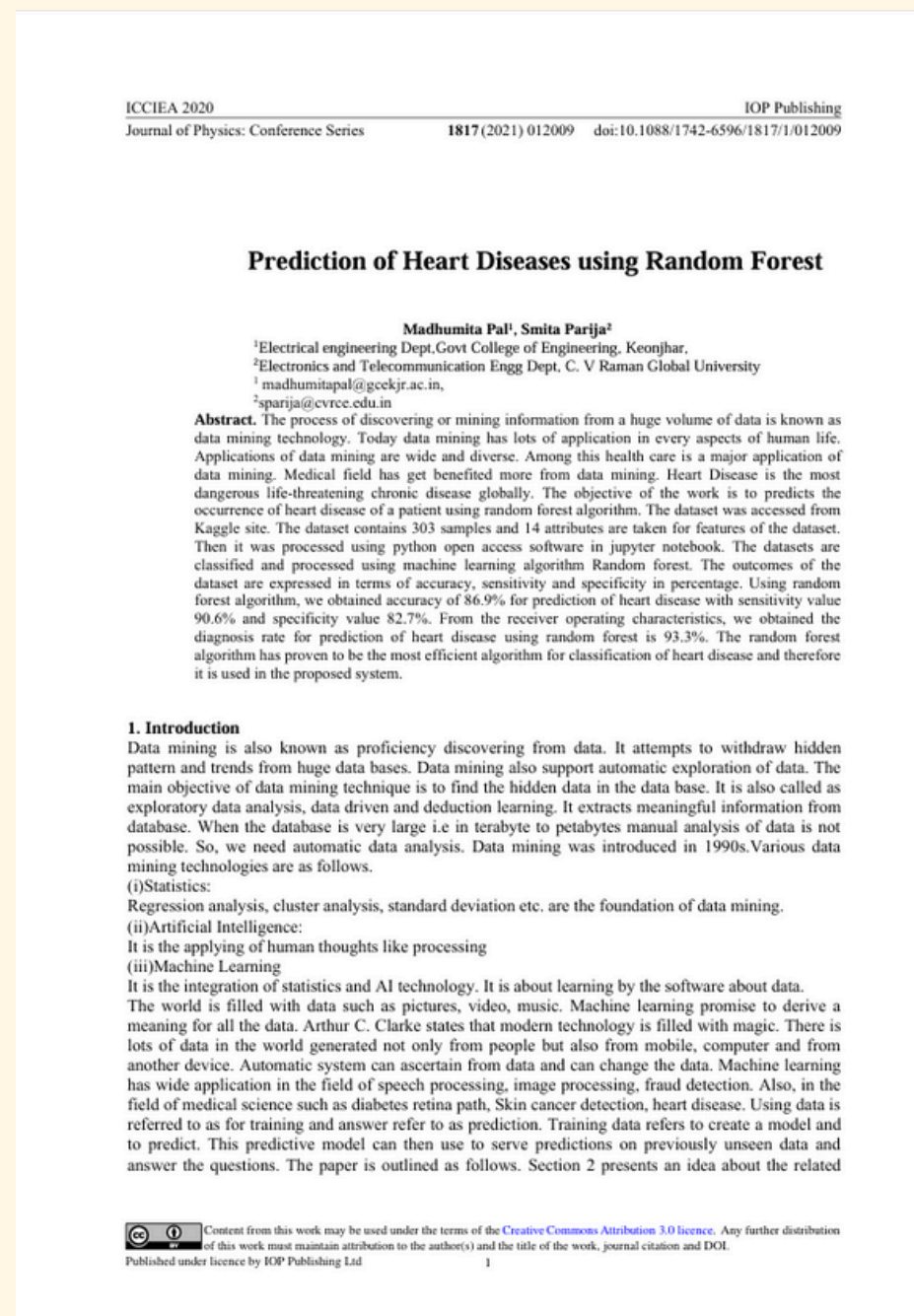# JUDUL / TOPIK

Hybrid Prediksi Model Penyakit Jangtung menggunakan Algoritma Random Forest dan K-Nearest Neighbors

Riset Informatika | Daffa Tungga Wisesa | 21081010243

Prediction of Heart Diseases using Random Forest

Madhumita Pal[1], Smita Parija[2]

[1]Electrical engineering Dept,Govt College of Engineering, Keonjhar,

[2]Electronics and Telecommunication Engg Dept, C. V Raman Global University

[1] madhumitapal@gcekjr.ac.in,

[2]sparija@cvrce.edu.in

Abstract. The process of discovering or mining information from a huge volume of data is known as data mining technology. Today data mining has lots of application in every aspects of human life. Applications of data mining are wide and diverse. Among this health care is a major application of data mining. Medical field has get benefited more from data mining. Heart Disease is the most dangerous life-threatening chronic disease globally. The objective of the work is to predicts the occurrence of heart disease of a patient using random forest algorithm. The dataset was accessed from Kaggle site. The dataset contains 303 samples and 14 attributes are taken for features of the dataset. Then it was processed using python open access software in jupyter notebook. The datasets are classified and processed using machine learning algorithm Random forest. The outcomes of the dataset are expressed in terms of accuracy, sensitivity and specificity in percentage. Using random forest algorithm, we obtained accuracy of 86.9% for prediction of heart disease with sensitivity value 90.6% and specificity value 82.7%. From the receiver operating characteristics, we obtained the diagnosis rate for prediction of heart disease using random forest is 93.3%. The random forest algorithm has proven to be the most efficient algorithm for classification of heart disease and therefore it is used in the proposed system.

1. Introduction

Data mining is also known as proficiency discovering from data. It attempts to withdraw hidden pattern and trends from huge data bases. Data mining also support automatic exploration of data. The main objective of data mining technique is to find the hidden data in the data base. It is also called as exploratory data analysis, data driven and deduction learning. It extracts meaningful information from database. When the database is very large i.e in terabyte to petabytes manual analysis of data is not possible. So, we need automatic data analysis. Data mining was introduced in 1990s.Various data mining technologies are as follows.

(i)Statistics:

Regression analysis, cluster analysis, standard deviation etc. are the foundation of data mining.

(ii)Artificial Intelligence:

It is the applying of human thoughts like processing

(iii)Machine Learning

It is the integration of statistics and AI technology. It is about learning by the software about data. The world is filled with data such as pictures, video, music. Machine learning promise to derive a meaning for all the data. Arthur C. Clarke states that modern technology is filled with magic. There is lots of data in the world generated not only from people but also from mobile, computer and from another device. Automatic system can ascertain from data and can change the data. Machine learning has wide application in the field of speech processing, image processing, fraud detection. Also, in the field of medical science such as diabetes retina path, Skin cancer detection, heart disease. Using data is referred to as for training and answer refer to as prediction. Training data refers to create a model and to predict. This predictive model can then use to serve predictions on previously unseen data and answer the questions. The paper is outlined as follows. Section 2 presents an idea about the related

RISET GAP

Penelitian sebelumnya oleh Madhumita Pal dan Smita Parija (2024) menunjukkan bahwa algoritma Random Forest dapat menghasilkan akurasi tinggi (86.9%). Namun, penelitian tersebut belum mengeksplorasi potensi penggabungan algoritma lain, seperti K-Nearest Neighbors (K-NN), untuk meningkatkan hasil prediksi.

MIND MAPPING

Dataset:
- Dataset Penyakit Jantung (Kaggle)
- 303 sampel, 14 atribut fitur.

Referensi Utama:
- Madhumita Pal, Smita Parija (2024): Menggunakan Random Forest untuk prediksi penyakit jantung.
- Artikel tentang K-NN untuk clustering dan preprocessing data medis.

MIND MAPPING

Metode Kombinasi:
- Voting Ensemble: Kombinasi output dari K-NN dan Random Forest.
- Hybrid Pipeline: K-NN sebagai preprocessing, Random Forest sebagai prediktor akhir.

METODE YG DIUSULKAN

1. **Metode Pengumpulan Data**

Data yang digunakan berasal dari kaggle yang berisi 303 sampel dengan 14 atribut yang terkait dengan penyakit jantung yang berasal dari dataset UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/45/heart+disease).

2. **Metode Eksplorasi Data**

Metode eksplorasi data yang untuk menganalisa dataset adalah matriks korelasi untuk memahami hubungan antar atribut yang ada pada dataset

METODE YG DIUSULKAN

3. **Metode Pra-pemrosesan Data**

Data diambil dan ditampilkan pada excel dengan menggunakan separator koma. Selanjutnya data diklasifikasi sebelum diproses dengan algoritma machine learning.

4. **Metode Eksplorasi Data**

Metode eksplorasi data yang untuk menganalisa dataset adalah matriks korelasi untuk memahami hubungan antar atribut yang ada pada dataset

METODE YG DIUSULKAN

5.**Metode Evaluasi Model**

Dataset dibagi menjadi 2, yaitu 80% untuk data training dan 20% untuk data training. Ukuran Evaluasi yang digunakan adalah accuracy, sensitivity, dan specificity dalam persentase.

MATRIK PENGUJIAN

1. **Matriks Korelasi:**
   - Mengidentifikasi hubungan antar fitur dataset sebelum masuk ke model.
   - Memastikan tidak ada multikolinearitas yang signifikan.
2. **Confusion Matrix:**
   - Mengukur performa model menggunakan metrik seperti True Positive (TP), False Positive (FP), True Negative (TN), dan False Negative (FN).