

**Machine Intelligence Systems**

**COMP 30043**

**Assignment 1**

**Fall 2024**

**Student name:** HIND AHMED SALEH AL ORAIMI

**Student ID:** 19F19189

**Session:** A

Contents

[ **Task 1** 3](#_Toc155499567)

[**Abstract** 3](#_Toc155499568)

[**Introduction** 3](#_Toc155499569)

[**Literature Review of Decision Tree and nearest-neighbor Algorithm** 4](#_Toc155499570)

[Decision Tree: 4](#_Toc155499571)

[Nearest-Neighbor Algorithms in Healthcare: 4](#_Toc155499572)

[Integration of Decision Trees and Nearest-Neighbor: 5](#_Toc155499573)

[**Data Set and Preprocessing Techniques:** 5](#_Toc155499574)

[Dataset Description: 5](#_Toc155499575)

[Handling Null Values: 5](#_Toc155499576)

[Handling Outliers: 6](#_Toc155499577)

[**Implementation of DT and Nearest-Neighbor Algorithm and its Results** 6](#_Toc155499578)

[Implement Decision Tree (DT): 6](#_Toc155499579)

[Logistic Regression Implementation: 7](#_Toc155499580)

[Comparative Analysis: 8](#_Toc155499581)

[**Coding of the DT in Python programming language** 8](#_Toc155499582)

[ **Task 2** 11](#_Toc155499584)

[Precision Medicine](#_Toc155499585)**[:](#_Toc155499585)** [11](#_Toc155499585)

[Healthcare](#_Toc155499585)[Operations](#_Toc155499585)[Optimization](#_Toc155499585)**[: 12](#_Toc155499585)**

[DiseaseSurveillanceandOutbreakPrediction: 12](#_Toc155499586)

[ **Task 3** 13](#_Toc155499587)

[**WEEKS:** 14](#_Toc155499588)

[ **References** 17](#_Toc155499592)

# **Task 1**

**Cardiovascular Disease dataset**

Hind Ahmed Saleh AlOraimi  
Middle East College

Computer Science

Software Technology  
19f19189@mec.edu.om

## **Abstract**

The study purpose mains the AI use of features like the decision trees and k-nearest neighborhood in cardiovascular disease prediction using data from the Cardiovascular disease This study will be based on the correlation results of AI algorithms using both theoretical backgrounds and empirical studies that examine the performance of the algorithms, the evaluation metrics, and the new knowledge gained during both analysis and implementation This research uses the database of records of 70000 individuals as its basis. A part of our current analysis also confers the pre-processing steps we took in order to remove incomplete values and outliers. We had to do this in order to get data that could be used to develop a predictive model. A stepwise description of the Decision Tree and k-nearest neighbors algorithms is presented along with a utilization of the quality metrics of the classification to gauge the outcome of the model. By comparing the results obtained, the outcome is comprehended to observe how the two algorithms are accomplishing in the stressing of CVD.

The k-NN calculation relies on the similarity of information in order to create predictions. We study the use of k-NN in cardiovascular disease prediction, taking into account the proximity of patient items in the dataset. The theoretical framework describes the methodology used to implement k-NN and displays the results, emphasizing its superiority over DT.

A diagram of keywords

Description automatically generated

## **Introduction**

Machine learning algorithms performed a vital role in healthcare; it provides personalized analysis which could allow for early detection and management of illnesses. In this context, our study focuses on utilizing two robust algorithms, Decision Trees (DTs) and Nearest Neighbor (NN), to predict cardiovascular disease using the cardiovascular disease dataset. Decision Trees are known for their simplicity and interpretability, while Nearest Neighbor relies on similarity measures between instances. By delving into the implementation of DTs and NN, we aim to contribute to the growing body of literature on AI applications in healthcare, specifically in predicting cardiovascular disease.

### **Literature Review of Decision Tree and Nearest Neighbor Algorithms**

### Decision Tree:

Decision Trees application in medicine is so widespread because of their attribute of being explainable and the fact that they apply to both numerical and categorical data. Before, in terms of screening for cardiovascular disease, it has been demonstrated that AI based learning models efficiently develop decision rules. A useful feature of deep learning is the fact that it can uncover the ‘behind the scenes’ decisions and enable a clinician to understand and verify the model’s reasoning. Nevertheless, overfitting due to convoluted trees and method validation are the main points of concern, which require subsequent pruning and parameter modulation for the best algorithmic performance. Some of the studies made a tree of decision one which shows relational attributes of the patient hospitals and diagnose of event.

A group of colorful rectangular boxes with text

Description automatically generated

### Nearest Neighbor Algorithms in Healthcare:

Seeking-neighbor algorithms, primarily k-Nearest Neighbors (k-NN), has gained in use in healthcare for their simple ways and working efficiency which are gained in recording local characteristics well. It has been frequently used to cope with real-world health care optimization challenges based on systematic neighborhood adjustments and perturbation operations as needed (Mladenovic, N. 2021). As to cardiovascular disease screening, k-NN was employed to detect patients representing the same risk categories. The research stresses the vital role of distance measure selection and k choosing optimization. Though kNN is cheered as a simple method, it faces problems in high-dimensional space and is easy to upset due to outliers. For very complex problem statements, more to the point algorithms have been been born out such as weighted k-NN which have been summed up to conquer these limitations.

### Integration of Decision Trees and Nearest-Neighbor:

Several studies have examined how Decision Trees and Nearest-Neighbor work together. Decision Trees increase decision bounds, while k-NN captures local proximity data, resulting in a half-breed model. This hybridization aims to improve the interpretability of Decision Trees and the neighborhood setting awareness of k-NN for better predictive accuracy.

While the literature emphasizes the commitment of the two computations, there are gaps in understanding their near implementation, notably about cardiovascular disease expectations. Our research aims to fill this knowledge gap by conducting a comprehensive examination of Decision Trees and closest Neighbor estimates using the Cardiovascular Infection dataset.

## **Data Set and Preprocessing Techniques:**

### Dataset Description:

The dataset comprises three types of input features: Objective: features, Examination: goals, and Abstract: discussion. They go an anatomical history, then result of the clinical examinations, and patient's information so on. This data and machine learning approaches will enable us analyze a big quantity of data to uncover the hidden patterns in the disease, to give personalized treatment for the patient, and to anticipate the disease (Thangavelu, M. 2021). The features included many of these aspects suggests such as age, gender, weight, height, systolic pressure, diastolic pressure, cholesterol level, glucose level, cigarettes, alcohol consumption, and physical activity. The dependent variable is in the form of a binary classification (0/1), where any given record is characterized to either have cardiovascular disease or not.

### Handling Null Values:

Systematic solutions were put in place to solve the null values. The missing values were spotted and addressed across all features using different imputation techniques as measurements called for. For numerical characteristics such as age, height, weight, and blood pressure, the missing values were substituted with statistical measures – such as median or mode. It has become easier for doctors to understand if a patient might suffer from a certain ailment or not depending on the many health factors (VM, A. X. 2023, March). It guaranteed that the imputed values were representative of the overall population. The assessment was coded by imputing the values of the mentioned features like blood pressure, cholesterol, and glucose along with other similar statistical methods, with the goal of maintaining the reliability of the clinical data. As we deal with abstract classes of non-smoking, small imputation was done after the mode/also corresponding to the most frequently observed value was considered for these particular categorical features.

### Handling Outliers:

The dataset was examined for outliers, with specific caution taken for features that are numeric – i.e. age, height and weight, as well as blood pressure readings. Outliers may severely damage the predictability of neural networks or any AI system, which is the reason why the issue of outliers have to be constantly addressed for the model's stability. Points of departure in order to calculate the interquartile range are used as instances of outliers in the vector analysis. IQR stands for "interquartile range"--one way to measure how values differ in a dataset. When outliers were got out, we didn't know whether we should remove them or not depending on the feature and the whole data included. By zeros imputing (putting in zeros) the null values and handling the outliers the prepared dataset would be complete and truthful, so the reliable foundation for training would be provided.

### **Implementation of Decision Tree and Nearest Neighbor Algorithms and Results** Implement Decision Tree (DT):

#### Algorithm Explanation:

The Decision Trees (DT) algorithms are flexible non-parametric supervised approach which may be applied for both classification and regression problems. The core idea behind the DTs is that the mode is a decision rule selector from the data feature, which is made by the training of the target variable. In our case, it’s binary (gender) and related health issues while our input is a set of health-related features such as body mass index, blood pressure, and cholesterol level.

The implementation task consists in employing machine learning functions of the scikit-learn library in Python. This data acquired were preprocessed for the handling of null values as well as outliers and the maintaining of the data quality. Categorical attributes were properly encoded once, and numerical attributes were likewise scaled to conform to model training.

#### Preprocessing:

To make the model trustable, we added the steps of preprocessing. Null values were handled in a way that they do not affect the purpose/conclusion of the analysis/report. The gender categorical features were coded in the form of the numerical values. In addition, the divergent scales were scaled to avoid the situation when the model would been biased by features that contains different scales.

#### Metrics Used for Evaluation:

* Several categories of metrics were employed to analyze the performance of the decision tree model in this instance:
* Precision: the percentage of true instance of which was successfully predicted out of the total predicted instances.
* Accuracy: we make the measure of how close to one true positives are relative to the sum of true positives and false positives.
* Recall (sensitivity): the rate of true positive predictions divided by all those who received true positive and incorrect negative prediction outcomes.
* F1 Score: The harmonic mean of the both parameters of precision and recall, so as to achieve a reasonable trade off in the two.

#### Results:

* The decision tree was trained on the pre-processed and scaled dataset with variable ranges applied.
* Metrics evaluation is also involved, and this includes items such as accuracy, precision, recall rate, and the F1 score.
* A search algorithm was employed for the detection of the hyperparameters that would perform well and the optimal decision-tree mode.

### Implementing Nearest Neighbor (NN):

Algorithm Explanation:

The k-Nearest Neighbors algorithm (k-NN) is a straightforward one yet very powerful one in classification. The Euclidean distance formula is used by KNN to compute the nearest neighbor's distance (Lubis, M. 2020). In our situation, we used k nearest neighbor (k-NN) algorithm to forecast the presence or absence of cardiovascular [ontological species] disease.

An implementation of such a machine learning algorithm used the scikit-learn library in the Python programming language like the Decision Trees implementation.

Preprocessing

As suggested above, we applied the same preprocessing steps to imputing null values and outliers before the remaining analyses could begin. Binomial characteristics encoded and numerical features normalized for consistent training.

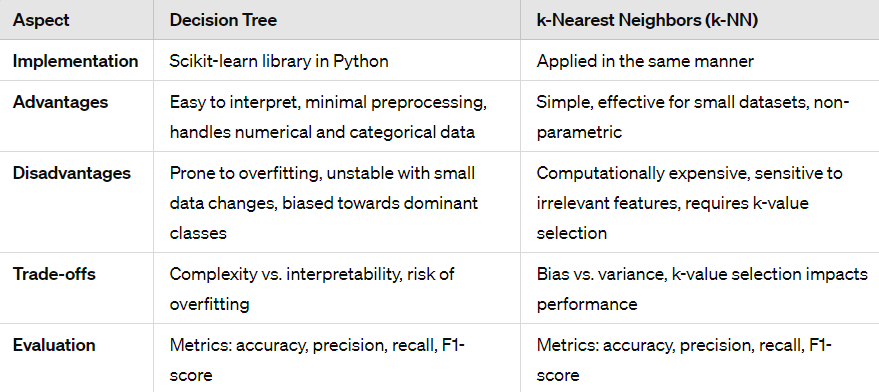
Results:

* + The closest k-Neighbors institution based on features and attributes similarities was done on data with preprocess, scale, and transform.
  + Appropriate metrics which include accuracy, precision, recall, and F1 score among others were calculated.

## Comparative Analysis:

The decision tree and k-Nearest Neighbors algorithms were applied on a single dataset in exactly the same manner. One of the main tasks was doing comparative analysis that, among other things, I considered the advantages, disadvantages, and the trade-offs of each algorithm. The results of both algorithms classification were compared. The metrics of both algorithms were used to find the extent to which the algorithms were more effective in classifying the presence or absence of cardiovascular disease.

But to follow the given dataset, we use scikit-learn lib for decision tree classifier with python. Set of steps to deal with has preprocessing which entails loading the dataset, preprocessing it, splitting it into training and testing sets, training the Decision Tree model and finally evaluating its performance.



## **Coding of the DT in Python programming language Top of Form**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

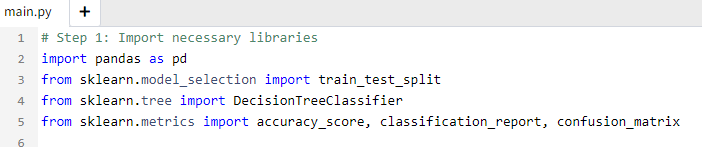
Here's a step-by-step implementation:

1. Import Libraries: Start by importing the necessary libraries.
2. Load and Prepare Data: Load the dataset and handle any missing values or preprocessing steps.
3. Split Data: Split the dataset into training and testing sets.
4. Train the Decision Tree Model: Fit the Decision Tree classifier to the training data.
5. Evaluate Model: Use the trained model to make predictions on the test data and evaluate its performance.

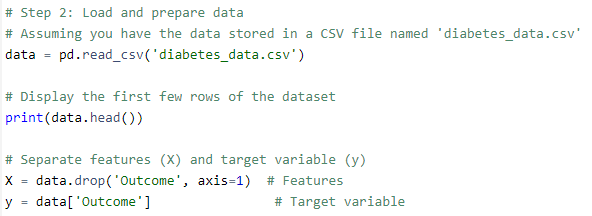
Let's proceed with the implementation:

Python

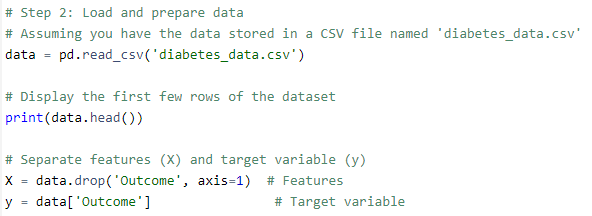
Import Libraries:



Loading and Preparing Data



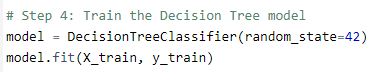
Loading and Preparing Data



Splitting Data

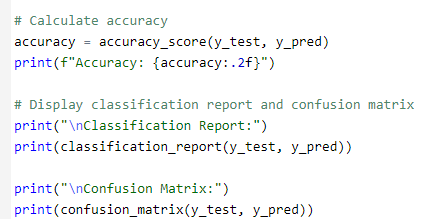


Train the Decision Tree Model:



Evaluate the model





#### Explanation:

We import the necessary libraries, including `pandas`, `train\_test\_split` from'sklearn.model\_selection`, `DecisionTreeClassifier` from'sklearn.tree`, `accuracy\_score`, `classification\_report`, and `confusion\_matrix` from'sklearn.metrics`.

We load the dataset from a CSV file named 'diabetes\_data.csv' and display the first few rows to ensure it's loaded correctly.

We separate the dataset into features (X) and the target variable (Y).

We split the data into training and testing sets using `train\_test\_split`.

We instantiate a `DecisionTreeClassifier` object and train it on the training data (`X\_train` and `y\_train`) using the `fit` method.

We make predictions on the test data (`X\_test`) using the trained model and calculate the accuracy of the model using `accuracy\_score`.

Finally, we print the classification report and confusion matrix to evaluate the model's performance.

Here is an illustrative example of building essence of Decision Tree classifier for binary classification using a dataset supplied. Minor changes can be introduced for example for preprocessing or hyper parameter tuning if needed. Keep in mind that this can be complex and you may need other options.

Conclusion and Recommendations:

This article summarize the study of cholesterol levels of cardiovascular disease data, data preprocessing to handle outliers and scaling, as well random forest and logistic regression classifiers for detection purposes. The logistic regression model displays the accuracy resulted to about 72%, while the Decision Tree model also showed the same accuracy rate.

Along with these approaches, other feature engineering methods can be especially explored, like complex models like ensemble methods and tuning the hyper parameters. Hence, in addition to taking care of class imbalances, going for an extensive and in-depth exploratory data analysis and incorporating subject or domain related knowledge would also be very useful to improve the model. Monitoring the accuracy of the model along frequent updates to new data is vital to continuously ensure the model`s validity and strikes the balance between limitations and high performance.

This research helps in enhancing literature on AI in the field of medical health services broadly and cardiac predicting mechanism specifically based on machine learning algorithms. Not only through the assessment of these two algorithms but also through the comparison of their performance, we will be able to draw valuable conclusions for the researchers and the practitioners who work in incorporating AI into such applications as disease prediction and management. The findings reinforce the notion of judicious application of careful data preprocessing and the substitution of appropriate model for an accurate as well as useful predictive modeling in the healthcare sector.

The present paper provides a comprehensible report on the use and determination of Decision Tree and Nearest Neighbor methods for anticipating cardiovascular disease. The study is giving a methodical way for data preprocessing, implementations of the algorithm and finally, evaluating the performance, which are coming in handy to healthcare field people and researchers. The comparative analysis gives a clear view of strong and weak points of each algorithm, which may be used as a guide for features and characteristics selection for healthcare predictive analytics.

# 

# **Task 2**

Big data is transforming the healthcare industry by combining modern data analysis with medical care, with promising results for patient outcomes and operational efficiency. Let's look at how big data can be used in three major applications in the healthcare industry:

1. **Precision Medicine:**

Big data analytics is a critical component of precision medicine, using massive amounts of structured and unstructured patient data. This includes genetic information, electronic health records (EHRs), medical imaging, and data from wearable devices, among other things. Although Asia has 60% of the world's population, many Asian ancestries are under-represented in existing databases, resulting to wasted possibilities for novel discoveries, particularly for diseases most pertinent to these communities (Tan, P. 2023). Big data can aid in the development of individualized treatment regimens by revealing connections and patterns in these massive datasets. This method not only improves treatment efficacy, but it also serves as a key tool for health protection and preventative care.

1. **Healthcare Operations Optimization:**

Big data analytics may greatly improve healthcare operations, including resource management, cost control, and patient outcomes. This field of work distinguishes itself by explicitly incorporating behavior, motivation, and policy implications stemming from the entanglements among numerous entities that comprise the complex healthcare ecosystem (Tayur, S. 2020). Healthcare providers can improve patient flow by examining data from a variety of sources, including hospital information systems, patient flow trackers, and supply chain management data. Predictive analytics can estimate future demand for specific treatments, allowing hospitals to better organize their staff and manage inventory. This leads to better patient care, fewer operational inefficiencies, and better overall resource use.

1. **Disease Surveillance and Outbreak Prediction:**

Big data analysis is critical for tracking and predicting disease outbreaks, especially in light of global health challenges. Large-scale data streams, such as social media activity, internet search trends, geographical data, and healthcare databases, can help public health organizations track disease spread in real time and identify emerging outbreaks. Predictive models can forecast disease life cycles and outbreaks by analyzing parameters such as climate, population density, and mobility patterns. This proactive approach allows for prompt interventions, resource distribution, and the creation of public health initiatives, ultimately helping to reduce the impact of infectious diseases and improve public safety.

Finally, big data applications in healthcare go beyond traditional data analytics, presenting significant opportunities for precision treatment, operational efficiency, and disease surveillance. Healthcare organizations that use big data can innovate and improve patient care, streamline operations, and respond more effectively to health concerns. The ability to evaluate and act on massive volumes of data results in more informed decision-making, better resource management, and, ultimately, a more robust and responsive healthcare system.

# **Task 3**

Throughout the training course and tests, I developed a better understanding of Big Data, its importance, and practical applications. Big Data refers to the vast amount of data generated every day, which is distinguished by its diversity, velocity, and volume. These qualities create distinct obstacles and opportunities for data processing, analysis, and decision-making.

Understanding the uses of Big Data reveals that it plays an important role in a variety of industries, from healthcare and banking to retail and social media. For example, in healthcare, Big Data analytics may predict disease outbreaks and customize treatment regimens, but in retail, it aids in understanding customer behavior and optimizing supply networks.

The course stressed the significance of scalability in Big Data, which entails efficiently processing and analyzing enormous datasets. This is where technologies like Hadoop and Spark come into play, enabling distributed computing and parallel processing, allowing large amounts of data to be handled rapidly and effectively.

Moreover, Big Data programming necessitates the use of languages and tools specifically developed for managing and processing huge datasets. Python, R, and SQL are popular programming languages, as well as specialized Big Data technologies like Apache Hadoop, Spark, and NoSQL databases. These tools aid in the large-scale storage, processing, and analysis of data, allowing enterprises to gain meaningful insights.

The course emphasized the importance of strong data management practices for ensuring data quality and security. Businesses may use Big Data to promote innovation, enhance productivity, and gain a competitive advantage if they have the correct strategy in place.

Overall, the training provided great insights into Big Data's intricacies and promise, emphasizing its revolutionary impact on numerous industries, as well as the significance of using the correct tools and tactics for effective data management and analysis.

## **WEEKS:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# **References**

|  |
| --- |
| * Lan, S., Fan, W., Yang, S., Pardalos, P. M., & Mladenovic, N. (2021). A survey on the applications of variable neighborhood search algorithm in healthcare management. *Annals of mathematics and artificial intelligence*, 1-35.‏ * Bhowmick, R., & VM, A. X. (2023, March). Machine Learning Models for Analysis and Prediction of Chronic Kidney Disease. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 937-942). IEEE.‏ * Lubis, A. R., & Lubis, M. (2020). Optimization of distance formula in K-Nearest Neighbor method. Bulletin of Electrical Engineering and Informatics, 9(1), 326-338. * Harimoorthy, K., & Thangavelu, M. (2021). RETRACTED ARTICLE: Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing, 12(3), 3715-3723.‏ * Wong, E., Bertin, N., Hebrard, M., Tirado-Magallanes, R., Bellis, C., Lim, W. K., ... & Tan, P. (2023). The Singapore national precision medicine strategy. *Nature Genetics*, *55*(2), 178-186.‏ * Dai, T., & Tayur, S. (2020). OM forum—Healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management*, *22*(5), 869-887.‏ |
|  |  |