**Performance Assessment**

NVM2 TASK 2: Predictive Analysis

Bader Ale

Department of Information Technology, Western Governors University

D209 Data Mining I

September 2023

Table of Contents

[Part I: Research Question 3](#_Toc146562018)

[Part II: Method Justification 3](#_Toc146562019)

[Part III: Data Preparation 3](#_Toc146562020)

[Part IV: Analysis 3](#_Toc146562021)

[Part V: Data Summary and Implications 3](#_Toc146562022)

# Part I: Research Question

For this performance assessment, our research question is: **Given certain patient characteristics, can we classify whether a patient is hypertensive or not**. Using decision trees, the goal for this data analysis is to be able to classify whether a patient is hypertensive or not taking into consideration other patient-specific variables contained in the dataset.

# Part II: Method Justification

Decision trees is a supervised learning algorithm that uses a question-based structure that poses several choices for data. Its’ name is based on the tree-like structure that it forms while creating the decision forks. Decision trees can be used for both regression and classification problems; the latter will be used for this assessment (Geeks for Geeks, 2023). The decision tree creates several nodes from a primary or “root” node and several decisions are posed by the algorithm. Depending on the answer (Yes/No, or a numerical comparison), it finds its way down the branches to the final answer. For this assessment, the last node (‘terminal node’) will contain the answer to the question, in this case whether the patient has high blood pressure or not (HighBlood\_Yes).

One of the assumptions of the decision tree classifier algorithm is that the output variable (output of last node) should be discrete, meaning a yes/no (Geeks for Geeks, 2023).

In order to run this algorithm, several packages were imported. Ther Pandas library was used to manipulate the dataset appropriately. Seaborn was used to visualize the data. Sci-kit Learn was used as the main library that contained all the machine learning modules. For example, the preprocessing module was used to split the dataset into its train and test data sets and the model\_selection module was used to import the decision trees algorithm.

# Part III: Data Preparation

In order to successfully run the model, some data preparation needs to be performed. This step, usually called data cleaning, involves looking errors, duplicates, and missing values among others, in preparation to pass the data into the model (Crabtree & Nehme, 2023).

One important step in the data preparation step is scaling. Scaling is critical for the model because it levels the scale of the variables so that the algorithm can perform better predictions (SciKit Learn, 2023).

Moreover,

# Part IV: Analysis

# Part V: Data Summary and Implications

Works Cited

Crabtree, M., & Nehme, A. (2023, July). *What is Data Analysis? An Expert Guide With Examples*. Retrieved September 2023, from Datacamp: https://www.datacamp.com/blog/what-is-data-analysis-expert-guide

*Geeks for Geeks*. (2023, August 20). Retrieved September 2023, from Decision Tree: https://www.geeksforgeeks.org/decision-tree/

SciKit Learn. (2023). *6.3. Preprocessing data*. Retrieved September 2023, from https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing