****

**Emerald International College**

**Department Of Data Science**

**Business Intelligence Project**

**Project title: The Analysis of Obesity Status According to the Individuals' Social and Physical Activities** **Using Artificial Intelligence Techniques.**

**Group 8 Members**

**Name of the Students:** **Id number**

Hagos W/senbet DS-0063/15

Melkamu Adigo DS-0065/15

Wassye Yitages DS-0099/15

Wondwossen Alemu DS-0087/15

Zelalem Aysheshim DS-0086/15

**October, 2024**

**Addis Ababa, Ethiopia**

1. **Introduction**

A severe, long-term condition, obesity is influenced by both genetic and environmental factors. It is characterized as having an unhealthy quantity of bodily fat or tissue. Eating habits, social circumstances, and psychological issues are the primary risk factors for obesity. Globally, obesity is a serious health issue that affects people of all ages. Over 2 billion individuals are obese or overweight globally at the moment. Obesity can be avoided, according to research. In this project, artificial intelligence methods were used to identify individuals at risk of obesity by using the obesity dataset collected by Koklu, N., & Sulak, S.A .from Kaggle. To analyze the survey data, four commonly used artificial intelligence methods in literature, namely Artificial Neural Network, K Nearest Neighbors, Random Forest and Support Vector Machine, were employed after pre-processing.

1. **Methods and Techniques**

This section offers a comprehensive overview of the dataset employed in the project, the artificial intelligence techniques utilized, and the performance criteria used to evaluate these models. The application was developed by applying the programming language Python and its accompanying libraries.

**Obesity Dataset:**

|  |  |
| --- | --- |
| **Table 1:Obesity Dataset** | |
| **Attributes** | **Values** |
| Sex | 1. Male (712) |
| 2. Female (898) |
| Age | Values in integers |
| Height | Values in integers (cm) |
| Overweight/Obese Families | 1. Yes (266) |
| 2. No (1344) |
| Consumption of Fast Food | 1. Yes (436) |
| 2. No (1174) |
| Frequency of Consuming Vegetables | 1. Rarely (400) |
| 2. Sometimes (708) |
| 3. Always (502) |
| Number of Main Meals Daily | 1. 1-2 (444) |
| 2. 3 (928) |
| 3. 3+ (238) |
| Food Intake Between Meals | 1. Rarely (346) |
| 2. Sometimes (564) |
| 3. Usually (417) |
| 4. Always (283) |
| Smoking | 1. Yes (492) |
| 2. No (1118) |
| Liquid Intake Daily | 1. amount smaller than one liter (456) |
| 2. Within the range of 1 to 2 liters (523) |
| 3. In excess of 2 liters (631) |
| Calculation Of Calorie Intake | 1. Yes (286) |
| 2. No (1324) |
| Physical Exercise | 1. No physical activity (206) |
| 2. In the range of 1-2 days (290) |
| 3. In the range of 3-4 days (370) |
| 4. In the range of 5-6 days (358) |
| 5. 6+ days (386) |
| Schedule Dedicated to Technology | 1. Between 0 and 2 hours (382) |
| 2. Between 3 and 5 hours (826) |
| 3. Exceeding five hours (402) |
| Type of Transportation Used | 1. Automobile (660) |
| 2. Motorbike (94) |
| 3. Bike (116) |
| 4. Public transportation (602) |
| 5. Walking (138) |
| Class | 1. Underweight (73) |
| 2. Normal (658) |
| 3. Overweight (592) |
| 4. Obesity (287) |

**Artificial Intelligence Models**

Four distinct artificial intelligence techniques were utilized in this project, and a description of each approach is provided in the subsequent sections.

**Artificial neural networks (ANN)**

Artificial neural networks (AI), in which basic functions such as the ability to generate new data from data collected by the brain by learning, remembering, and generalizing by imitating the learning path of the human brain, are performed by computer software [2].

**Support Vector Machine (SVM)**

Support Vector Machines (SVMs) are a widely employed and highly effective machine-learning technique for data classification [3]. SVMs are supervised learning models based on statistical learning theory, which involve learning algorithms that analyze the data for classification and regression analysis.

**K Nearest Neighbors (KNN)**

K Nearest Neighbor (KNN) is a non-generalizing or sample-based learning algorithm, also referred to as a "lazy learning" algorithm, that gained popularity in statistical data analysis during the early 1970s. Unlike other algorithms, KNN does not concentrate on constructing an internal model by storing all training data samples in n-dimensional space [3].

**Random Forest (RF)**

Random Forest is an ensemble learning approach proposed and developed by Breiman to solve classification and regression problems [4, 5]. This algorithm creates many decision trees in general structure and combines them to get the best result.

1. **Results**

This section presents experimental results obtained by training and testing artificial intelligence methods (ANN, SVM, KNN, and RF). In the subsequent section, the classification outcomes of the models employed in the investigation are outlined.

**ANN Architecture-Based Classification Results**

Table 2 presents the classification values obtained from the confusion matrix of the ANN model. Based on these values, the accuracy rates for the underweight, normal, overweight, and obesity categories were calculated. The results indicate an overall classification success of 73.29%, as shown in Figure 9. However, the confusion matrix also reveals that the overweight and obesity categories were less accurately differentiated.

Table 2. ANN confusion matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANN** | | **Predicted Class** | | | |
| (4 x 4) | | Underweight | Normal | Overweight | Obesity |
| **Actual Class** | Underweight | 7 | 8 | 2 | 1 |
| Normal | 1 | **110** | 16 | 0 |
| Overweight | 0 | 21 | **77** | 28 |
| Obesity | 0 | 0 | 9 | **42** |

When Table 2 is examined, it is seen that 5 data in the Underweight class, 107 in the Normal class, 83 in the Overweight class, and 34 in the Obesity class are correctly classified. When TP, TN, FP, and FN data are examined, the most successful class is observed to be Underweight. It has been determined that the Normal, Overweight, and Obesity classes are highly mixed with each other. The reason is that, the training of the models is not fully realized due to the data imbalance between the classes. Classes are confused with each other since the data of the classes are very close to each other [1].

**KNN Architecture-Based Classification Results**

Upon analyzing the classification values in Table 3, the accuracy rates of the underweight, normal, overweight, and obesity classes were computed for the KNN model. Figure 9 demonstrates a classification success of 72%. However, it is evident from the confusion matrix in Table 3 that the differentiation performance of the overweight and obesity classes is inadequate.

Table 3. KNN Confusion matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | | **Predicted Class** | | | |
| (4 x 4) | | Underweight | Normal | Overweight | Obesity |
| **Actual Class** | Underweight | **5** | 10 | 2 | 1 |
| Normal | 1 | **107** | 17 | 2 |
| Overweight | 0 | 27 | **83** | 16 |
| Obesity | 0 | 2 | 15 | **34** |

Upon examining Table 3, it can be observed that the number of correctly classified data is 5 in the Underweight class, 107 in the Normal class, 83 in the Overweight class, and 34 in the Obesity class for the KNN model. Analysis of the TP, TN, FP, and FN data reveals that the most successful class is Obesity. However, it is evident that the classification success of the Underweight class is the lowest due to the scarcity of data.

**RF Architecture-Based Classification Results**

After analyzing the classification results of the RF model, as shown in Table 8, the accuracy rates of underweight, normal, overweight, and obesity classes were calculated. Figure 9 illustrates that the RF model achieved the highest classification success rate of 87% accuracy. However, the confusion matrix in Table 4 indicates that the normal and overweight classes had low performance in terms of differentiation.

Table 4. RF Confusion matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RF** | | **Predicted Class** | | | |
| (4 x 4) | | Underweight | Normal | Overweight | Obesity |
| **Actual Class** | Underweight | **8** | 9 | 1 | 0 |
| Normal | 0 | **118** | 9 | 0 |
| Overweight | 0 | 13 | **109** | 4 |
| Obesity | 0 | 0 | 7 | **44** |

Upon reviewing Table 4, it can be observed that 8 data in the Underweight class, 118 in the Normal class, 109 in the Overweight class, and 44 in the Obesity class are classified accurately. Upon analyzing the TP, TN, FP, and FN data, it can be deduced that the Normal class has the highest classification success. Due to the limited amount of data, the Underweight class has the lowest classification success.

**SVM Architecture-Based Classification Results**

The classification results for the SVM model are presented in this section, with the corresponding confusion matrix given in Table 5. The accuracy rates of the underweight, normal, overweight, and obesity classes were calculated using this matrix, and it was found that the model achieved a classification success rate of 72%, as shown in Figure 9. However, the confusion matrix in Table 5 indicates that the SVM model had difficulty differentiating between the overweight and obesity classes.

Table 5. SVM Confusion matrix

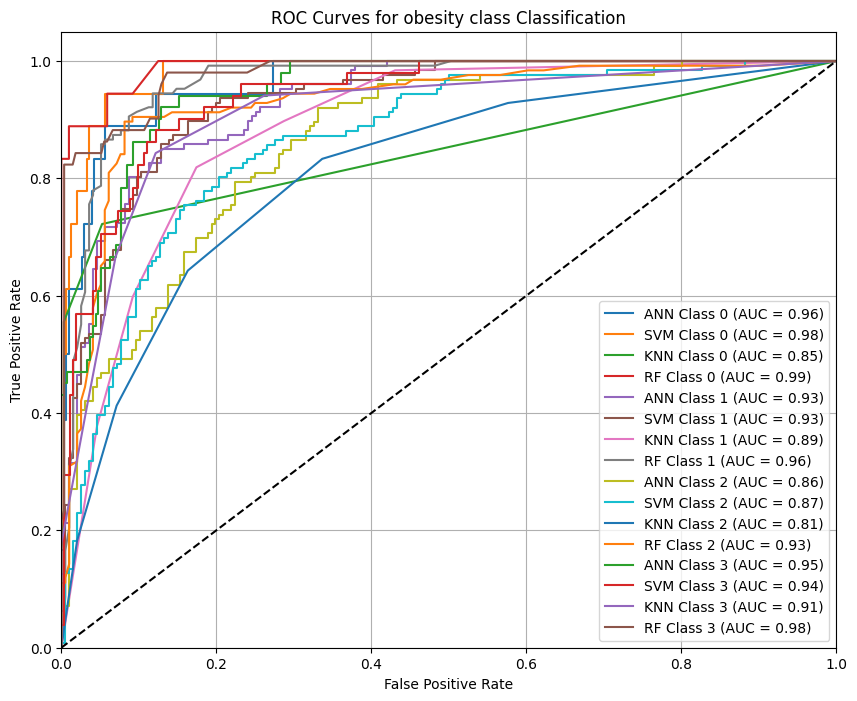
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SVM** | | **Predicted Class** | | | |
| (4 x 4) | | Underweight | Normal | Overweight | Obesity |
| **Actual Class** | Underweight | **7** | 10 | 1 | 0 |
| Normal | 2 | **108** | 15 | 2 |
| Overweight | 0 | 31 | **80** | 15 |
| Obesity | 0 | 3 | 12 | **36** |

Upon analyzing Table 5, it is observed that the SVM model accurately classified 7 instances of the Underweight class, 108 instances of the Normal class, 80 instances of the Overweight class, and 36 instances of the Obesity class. A closer look at the TP, TN, FP, and FN data indicates that the Normal class had the highest classification performance. However, due to limited data availability, the Underweight class demonstrated the lowest classification success.

**Results of All Classification Models**

The accuracy, precision, recall, and F1 Score values of each model were obtained using the confusion matrix data, and the results and graph for all models are shown in Figure 1. Upon analyzing the outcomes presented in Figure 1, it was revealed that the RF model has the highest classification success rate (87%), whereas the KNN model has the lowest (71%). In a similar vein, the RF model exhibits the highest metric values beyond classification success, whereas the KNN model displays the lowest values. ROC curves were obtained to conduct a thorough analysis of the model's performance, and their results are shown in Figure 2 for all models.

**Figure 1. Performance metrics were obtained for ANN, SVM, KNN, and RF**



**Figure 2. ROC curves of all models**

Upon examining the ROC curves presented in Figure 2, it can be observed that the RF model achieved the most successful learning performance while the KNN model had the least successful learning performance.

1. **Conclusions**

In recent years, obesity has spread quickly over the world's continents. Obesity in people can lead to a variety of disorders. Because of this, it is essential to assess the risk of obesity by routine monitoring and to adopt the appropriate safety measures. Recently, there has been a greater requirement for artificial intelligence techniques to carry out these tasks precisely and rapidly. In this project, a dataset of 14 features and four classifications was used to determine an individual's obesity status using a variety of machine learning approaches. For the obesity data set we get from Kaggle, classification procedures were carried out utilizing ANN, SVM, KNN, and RF techniques. With 87% classification accuracy, the RF model outperformed the KNN model, which had a 71% accuracy rate. The SVM model had a 72% classification success rate, whereas the ANN model had a 73% classification success rate. The dataset's size should be expanded in order to boost classification success. Additionally, distributing data evenly among classes will improve classification success. Several machine learning techniques can be used to improve classification success. Additionally, optimization or feature selection techniques can be used to identify the features that work best for classification, which can improve classification speed and success.

1. **Reference**

[1] Koklu, N., & Sulak, S. A. (2024). Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities. *Sinop Üniversitesi Fen Bilimleri Dergisi*, *9*(1), 217-239.

[2] Kim P. (2017). *Matlab Deep Learning*. Springer.

[3] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*, 1-21. <https://doi.org/10.1007/s42979-021-00592-x>.

[4] Moreno, L. A., & Rodriguez, G. (2007). Dietary risk factors for development of childhood obesity. *Current Opinion in Clinical Nutrition & Metabolic Care*, *10*(3), 336-341. <https://doi.org/10.1097/MCO.0b013e3280a94f59>.

[5]Breiman L. (2001). Random forests. *Machine Learning, 45*, 5-32. <https://doi.org/10.1023/A:1010933404324>.