**OBESITY LEVEL PREDICTION USING SVM AND KNN**

**Abstract -** This study focuses on classifying obesity levels in individuals from Mexico, Peru, and Colombia, using a dataset of 17 attributes. The NObesity variable categorizes data into seven classes. The dataset's unique composition, with 77% synthetically generated and 23% directly collected, adds complexity to the analysis. The paper employs data cleaning, exploratory data analysis, visualization, and hypothesis testing, aiming to optimize classification accuracy. Comparative analysis between SVM and KNN algorithms highlights improved accuracy. This research contributes insights into obesity classification factors, incorporating statistical rigor through hypothesis testing.

***Keywords:* SVM, Classification, KNN, Machine learning, Obesity**

1. **INTRODUCTION**

Obesity is a critical health concern worldwide, presenting multifaceted challenges in healthcare. Influenced by an intricate interplay of genetic, environmental, and lifestyle factors, obesity contributes significantly to various health issues, including cardiovascular diseases, diabetes, and overall diminished quality of life. Understanding and addressing this complex problem necessitate advanced analytical approaches, particularly in the era of healthcare data abundance.

This project centers on applying supervised learning methods, specifically Support Vector Machines (SVM) and k-Nearest Neighbors (KNN), to tackle the intricate challenge of classifying obesity levels in individuals from Mexico, Peru, and Colombia. These algorithms are trained on a diverse set of features encompassing aspects such as eating habits, physical condition, and genetic predispositions.

Aims of the Project:

The project unfolds through a systematic approach to data analysis. The preprocessing phase involves critical steps such as data cleaning, addressing missing values, and removing redundant features to ensure the

dataset's integrity and relevance. Exploratory data analysis encompasses univariate, bivariate,

and multivariate analyses to unravel intricate relationships among features. Visualization techniques are employed to gain a comprehensive understanding of the data distribution.

Hypothesis testing adds a statistical dimension to the exploration, providing valuable insights into population characteristics. Scaling the data ensures uniformity, facilitating more effective model training. Feature engineering involves a meticulous feature selection process, refining the dataset for optimal model performance.

The core of the project lies in modeling the data using SVM and KNN algorithms, accompanied by extensive hyperparameter tuning to enhance predictive accuracy. Comparative analysis between these models serves to identify strengths and weaknesses, guiding future applications in healthcare data analysis.

In navigating through these comprehensive steps, the project aims to not only build accurate models for obesity classification but also contribute methodologically to the broader field of healthcare analytics. The insights gained from this study are expected to inform healthcare practitioners and researchers, ultimately advancing our understanding of obesity-related factors, and fostering more targeted intervention strategies.

1. **LITERATURE REVIEW**

In this section, five papers are discussed related to the topics and the algorithms used.

This study aims to employ machine learning algorithms for predicting obesity risk, utilizing a diverse dataset of over 1100 individuals spanning various age groups. Nine prominent algorithms, including k-nearest neighbor (k-NN), random forest, logistic regression, and support vector machine (SVM), were applied. The Logistic Regression Algorithm demonstrated the highest accuracy at 97.09%, while gradient boosting exhibited the lowest accuracy of 64.08%, along with inferior metric values. The research provides insights into obesity risk levels and algorithmic performance, enhancing understanding in this health-related domain [1]. This paper explores BMI levels and employs multiple machines learning algorithms, including Logistic Regression, Naïve Bayes, Decision Tree, XGBoost, and KNN, to predict obesity using patient data. Achieving higher accuracy with Logistic Regression and Decision Tree, the future direction involves leveraging these algorithms to recommend diseases associated with obesity [2]. This study aims to employ machine learning techniques to create a predictive model for identifying individuals with overweight or obesity. Various algorithms, such as Multilayer Perceptron (MLP), Support Vector Machine (SVM), and others, were evaluated, with Random Forest (RF) and Logistic Regression (LR) showing the highest accuracies at 95.78% and 95.22%, respectively. The calibrated model has practical implications for physicians in the early detection and management of obesity-related diseases [3]. This study employed three machine learning models to estimate obesity metrics, optimizing with four Hyperparameter Optimization (HPO) models, highlighting Optuna's efficiency. The deployed models, integrated with Python Flask and a website framework, included a dashboard, customizable diet plans, and an IoT-based weighing scale. Future efforts aim to enhance accuracy, address dataset limitations, and introduce computer vision for improved predictions[4].This study employs computational intelligence techniques, including K-Means, Decision Trees (DT), and Support Vector Machines (SVM). Evaluation metrics such as precision, recall, true positive rate, false-positive rate, and ROC area are used to compare these methods. Utilizing the Weka data mining tool for training and classification, the DT and Simple K-Means methods demonstrate superior results, surpassing previous studies with precision (98.5%), recall (98.5%), true positive rate (98.5%), false-positive rate (0.2%), and ROC area (99.5%) [5].

1. **DATA PROCESSING**

The dataset, obtained from Kaggle, consists of 17 attributes and 2111 records, focusing on estimating obesity levels in individuals from Mexico, Peru, and Colombia. The class variable, NObesity, classifies records into seven categories: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. The dataset encompasses diverse features such as gender, age, height, weight, family history of overweight, frequency of high-caloric food consumption (FAVC), frequency of vegetable consumption (FCVC), number of main meals consumed daily (NCP), eating habits between meals (CAEC), smoking status (SMOKE), daily water intake (CH2O), monitoring calories intake daily (SCC), frequency of physical activity (FAF), time spent using technological devices (TUE), frequency of alcohol consumption (CALC), and mode of transportation used (MTRANS). Each feature plays a crucial role in understanding and classifying obesity levels, contributing to a comprehensive analysis of individual health profiles.

***Data Cleaning***

A table with numbers and text

Description automatically generatedTo ensure the integrity and accuracy of our dataset, several key steps in data cleaning were undertaken. The "Smoke" column contained "?" values (Fig 3.1), which were replaced with Nan (Not a Number) using the NumPy library.

Fig: 3.1

*A screenshot of a computer code

Description automatically generated*Additionally, various columns exhibited missing values (Fig 3.2), and their treatment depended on the nature of the data.

Fig: 3.2

For numerical columns such as "Age," "Weight," and "Height," missing values were imputed using statistical measures. The mean was employed for continuous columns like "Age," "Weight," and "Height" to maintain the overall central tendency of the dataset. Conversely, for categorical columns like "Smoke" and "Mtrans," mode imputation was applied to replace missing values.

Duplicate records were identified and removed using the *drop\_duplicates()* function, ensuring the dataset's cleanliness and avoiding redundancy.To enhance the consistency and correctness of our dataset, appropriate data types were assigned to each column. This step was crucial for subsequent operations, enabling accurate analysis and model training on the refined dataset. After performing all these steps, the size of the cleaned data frame was (2088,17).

***Exploratory Data Analysis***

In the course of Exploratory Data Analysis (EDA), a detailed examination of the dataset was carried out to derive meaningful insights.

*Univariate Analysis:*

A graph of weight loss

Description automatically generatedThe distribution of the target variable, "NObeyesdad," was explored (Fig 3.3) to understand the prevalence of different obesity levels. It can be observed that Obesity type\_I is the most common one followed by others.

Fig : 3.3

A graph of a person and person

Description automatically generated with medium confidenceThe means of transport ("MTRANS") distribution (Fig 3.4) shed light on the preferred modes of transportation among individuals. Public transportation is the most opted one.

Fig : 3.4

A diagram of a weight scale

Description automatically generated with medium confidenceA comparison of a box diagram

Description automatically generatedBox plots for "Height" and "Weight"(Fig 3.5) provided insights into their spread across various quartiles.

Fig: 3.5

Distribution plots for "Age," "Height," and "Weight" (Fig 3.6) were analyzed to understand their distribution types.

A graph of a number of age

Description automatically generated

A graph of a weight scale

Description automatically generated with medium confidenceFig : 3.6

The positive skewness (1.5178) in age distribution indicates a concentration of younger individuals, with a few older individuals contributing to a longer right tail.

A skewness value close to zero (0.2428) suggests a relatively symmetric distribution for weight, with data evenly spread on both sides of the mean. The skewness value near zero (-0.0252) implies an approximately symmetric height distribution, indicating an even spread around the mean.

*Bivariate Analysis:*

A graph of different colored dots

Description automatically generatedRelationships between "Age" and different obesity levels were explored (Fig 3.7) to understand age distribution patterns. Age group (20-40) are more prone to obesity.

Fig: 3.7

The impact of the number of meals ("NCP") on obesity levels was visualized through a violin plot(Fig 3.8). It is observed that people having 3 meals are more likely to get obese. Also, a horizontal line is observed for type\_III, this indicates a stable median value for the variable under consideration.

Fig: 3.8

*Multivariate Analysis:*

A screenshot of a graph

Description automatically generatedA graph of a bar graph

Description automatically generatedWeight distribution across different genders and obesity levels was compared through a bar plot(Fig 3.9).It can be observed that male are more prone to obesity than females, also the tendency of having obesity\_type\_III is more common in both genders. Individuals with weight more than 75kgs are also more prone to develop obesity.

Fig: 3.9

The interplay between smoking ("SMOKE"), alcohol consumption ("CALC"), water intake ("CH2O"), and physical activity ("FAF") on obesity levels was examined (Fig 3.10) using a FacetGrid. It can be inferred that individuals consuming less water and more alcohol has chances of getting obesity. Also, smoking has less effect on obesity levels compared to water intake and alcohol consumption.

A group of graphs showing different colored dots

Description automatically generatedFig: 3.10

These analyses provided a comprehensive understanding of individual features, their relationships, and their collective impact on obesity levels, laying the foundation for subsequent modeling and predictions.

***Correlation check***

In the process of exploring relationships among variables and assessing the strength and direction of their linear associations, correlation analysis is a valuable step. Two common visualizations for this purpose are the heatmap (Fig 3.11) and pair plot.

The pair-wise correlation heatmap visualizes linear relationships between features using color intensity to represent correlation strength. Dark blue indicates strong negative correlations (as one increases, the other decreases), yellow indicates strong positive correlations (both increase or decrease together), and white indicates no correlation. This can help identify redundant features (highly correlated) or features important for your target variable (highly correlated with the outcome you're interested in).

Fig: 3.11

For instance, in the heatmap, there seems to be a strong positive correlation between Weight and Height (0.46). This means that there is a tendency for people with higher weight values to also have higher Height values. On the other hand, there appears to be a strong negative correlation between Weight and TUE (-0.8). This suggests that there might be a tendency for people with lower Weight values to have higher TUE values, and vice versa.

1. **DATA MANAGEMENT**

In the realm of data management, handling categorical variables through label and one-hot encoding, scaling numerical features for uniformity, and strategic feature selection are critical steps. Categorical encoding ensures models interpret qualitative data accurately. Scaling contributes to the stability of algorithms like SVM and KNN, preventing bias towards features with larger magnitudes. Feature selection optimizes model efficiency and interpretability by retaining relevant information and discarding redundancy. These practices collectively enhance the performance and interpretability of machine learning models, laying the foundation for effective predictive analytics.

***Handling categorical variables***

In the process of preparing the dataset for machine learning models, categorical variables were appropriately handled using a combination of label encoding and one-hot encoding. For binary categorical columns, such as 'Gender,' 'family\_history\_with\_overweight,' 'FAVC,' 'SMOKE,' and 'SCC,' label encoding was applied. This technique assigns a unique numerical label to each category, converting them into integers. Label encoding is suitable for binary columns where there are only two distinct categories, effectively representing them as 0s and 1s.

A graph with blue and white bars

Description automatically generatedOn the other hand, non-binary categorical columns, including 'CAEC,' 'CALC,' 'MTRANS,' and the target variable 'NObeyesdad,' underwent one-hot encoding. This method creates binary columns for each category within these variables, indicating the presence or absence of each category using 0s and 1s. The 'get\_dummies' function from the pandas library was utilized for this purpose, facilitating the transformation of categorical information into a format suitable for machine learning algorithms.

***Scaling the data***

Scaling data is a crucial preprocessing step in machine learning, and for this project, the Standard Scaler from the sklearn preprocessing module was employed. Before delving into the importance of scaling, the dataset was divided into a feature set and the target column. Scaling data using the Standard Scaler is integral to ensuring a fair and effective model training process in machine learning. By placing numerical features on a standardized scale, this preprocessing step prevents certain features from unduly influencing the model based on their magnitudes. The significance of scaling is particularly pronounced for algorithms like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), enhancing their stability and convergence during training. This practice promotes a balanced comparison among different features, allowing each to contribute proportionately to the model's learning process.

***Feature selection***

Utilizing a Random Forest Classifier for feature selection (Fig 4.1) is a robust approach to discern the significance of each feature within a dataset. The method evaluates the contribution of features by assessing their impact on reducing impurity across multiple decision trees. The resulting feature importance scores provide a clear hierarchy of influential features. This information is instrumental for feature selection, allowing the identification and prioritization of key attributes crucial for predicting the target variable. By focusing on the most impactful features, this technique optimizes model interpretability, performance, and generalization to new data, contributing to an efficient and effective machine learning model.

Fig: 4.1

A screenshot of a computer code

Description automatically generatedIt can be observed from the plot, that SMOKE and SCC have less importance in the model’s prediction capability. Hence these are removed from the dataset before splitting the dataset into train and test split.

Whereas WEIGHT has the most impact on model’s prediction and is the most important feature followed by age, height and so on.

***Splitting the dataset***

The process of splitting the dataset into training and testing sets is a crucial step in machine learning model development. This division allows for the assessment of a model's performance on independent data, contributing to a robust evaluation. The target variable y undergoes a transformation using np.argmax() to convert a one-hot encoded matrix into a single-column array, ensuring compatibility with certain machine learning models.

The train\_test\_split function from sklearn.model\_selection is then employed to randomly partition the dataset into training (x\_train, y\_train) and testing (x\_test, y\_test) sets. The test\_size parameter determines the proportion of the data allocated to the test set (in this case, 20%), facilitating a reliable evaluation of the model's ability to generalize to new, unseen data. The random\_state parameter ensures reproducibility by fixing the random seed, maintaining consistency across multiple runs of the model development process. This separation is pivotal for effective model training, validation, and eventual performance assessment.

1. **MODELLING**

In the initial modeling phase, both Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) algorithms were applied to the entire dataset, producing accuracy values of 93% and 79%, respectively. These results were obtained using random hyperparameters without the benefit of feature selection or grid search optimization. Subsequently, the modeling approach was refined by implementing feature selection techniques and utilizing grid search to identify optimal hyperparameters for SVM and KNN. By carefully selecting relevant features and systematically tuning hyperparameters, thisenhanced modeling strategy aims to improve overall model accuracy.

***Support Vector Machine (SVM)***

For the SVM modeling phase, a comprehensive grid search technique was employed to fine-tune the hyperparameters for optimal model performance. Cross-validation (cross folding) was used during the grid search (GridSearchCV) with "cv=5". The grid search explored various combinations of hyperparameter values, such as 'C' (regularization parameter), 'kernel' (specifying the type of kernel), 'degree' (relevant for 'poly' kernel), and 'gamma' (kernel coefficient). The considered options were:

* 'C': [0.1, 1, 10]
* 'kernel': ['linear', 'poly', 'rbf']
* 'degree': [2, 3, 4] (if 'poly' kernel is used)
* 'gamma': ['scale', 'auto'] (if 'rbf' kernel is used)

After an exhaustive search, the optimal hyperparameters were determined to be {'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}. This configuration yielded an impressive accuracy of 96.17%. Subsequently, this optimal SVM model was employed to generate a comprehensive classification report, providing insights into precision, recall, and F1-score for each class. This report is instrumental in evaluating the model's performance across various metrics, offering a holistic view of its effectiveness in classifying instances into different obesity levels.

***K-Nearest Neighbors (KNN)***

In the KNN modeling phase, a grid search strategy was employed to identify optimal hyperparameters for enhanced model performance. Cross-validation (cross folding) was used during the grid search (GridSearchCV) with "cv=5”. The grid search explored a range of values for 'n\_neighbors' (number of neighbors), 'weights' (weighting strategy for neighbors), and 'p' (distance metric). The parameter grid considered options as follows:

* 'n\_neighbors': [3, 5, 7, 9]
* 'weights': ['uniform', 'distance']
* 'p': [1, 2] (1 for Manhattan distance, 2 for Euclidean distance)

Following the grid search, the best hyperparameters were determined to be {'n\_neighbors': 5, 'p': 1, 'weights': 'distance'}. This configuration resulted in a commendable accuracy of 83.73%. Subsequently, the optimal KNN model, characterized by these hyperparameters, was trained on the dataset. The model's performance was further assessed by generating a detailed classification report, encompassing precision, recall, and F1-score for each class. This report serves as a valuable tool for evaluating the model's efficacy in accurately classifying instances into different obesity levels.

1. **COMPARISON (SVM Vs KNN)**

In the comparative analysis of SVM and KNN models, performance evaluation was conducted using advanced metrics and visualizations. Confusion matrices were employed to provide a detailed breakdown of the models' classification accuracy across different obesity levels. These matrices visually depicted the true positive, true negative, false positive, and false

negative classifications, offering insights into the models' precision and recall.

A graph with numbers and squares

Description automatically generatedA graph with numbers and squares

Description automatically generatedWe can see from the SVM confusion matrix that it predicted 3 wrong for class I, 2 wrong for class II, 5 wrong for class V and 6 wrong for class VI. Whereas KNN performed poorly on class I. Both the classifiers performed well on class IV with zero wrong predictions. The predictions for class III are also nearly same.

A graph of multicolored lines

Description automatically generatedA graph of multi-class results

Description automatically generated with medium confidenceImplemented a function to plot multiclass ROC curves for both SVM and KNN models. The ROC curves visualize the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different threshold values for classifying instances. The Area Under the Curve (AUC) values are also calculated for each class, providing a quantitative measure of the models' discriminatory power.

The interpretation of AUC values is as follows:

AUC values close to 1 indicate excellent performance, where the model has a high true positive rate and a low false positive rate.

AUC values around 0.5 suggest random performance, indicating that the model's discriminative ability is no better than chance.

AUC values below 0.5 indicate poor performance, where the model is more likely to classify instances incorrectly. So, we can see that SVM performed well with classes 0, 3 and 4. Whereas, KNN performed well with classes 3 and 4. For the micro-average AUC, it combines the true positive and false positive rates across all classes, providing an aggregate measure of the overall model performance across different classes. A higher micro-average AUC suggests better overall performance.

1. **HYPOTHESIS TESTING**

Two hypotheses were tested to explore associations within the dataset. The first hypothesis investigated the relationship between smoking status ('SMOKE') and obesity levels, utilizing a Chi-squared test for independence. The second hypothesis examined potential differences in the median physical activity frequency ('FAF') across various obesity levels, employing an Analysis of Variance (ANOVA) test. In both cases, the null hypotheses were rejected, indicating significant associations and differences, respectively. These tests served as valuable tools in validating assumptions and uncovering insights within the dataset.

1. **CONCLUSION**

In conclusion, this project delved into the task of classifying obesity levels in individuals from Mexico, Peru, and Colombia. Through meticulous data cleaning, exploratory analysis, and strategic preprocessing steps, the dataset was refined for modeling. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) emerged as robust models, achieving accuracies of 96.17% and 83.73%, respectively, following hyperparameter optimization.

The application of hypothesis testing revealed significant associations between smoking status and obesity levels, as well as variations in physical activity frequency among different obesity categories. The comprehensive evaluation included confusion matrix visualizations and ROC curve analyses, providing a nuanced understanding of model performance.

This project contributes valuable insights into lifestyle factors influencing obesity and underscores the potential applications of machine learning in healthcare. Future research could explore additional features and advanced techniques to further enhance predictive capabilities. Overall, this studyadvances our understanding of obesity classification and its determinants.

1. **REFERENCES**

[1] Faria Ferdowsy, Kazi Samsul Alam Rahi, Md. Ismail Jabiullah, Md. Tarek Habib,A machine learning approach for obesity risk prediction,Current Research in Behavioral Sciences,Volume 2,2021,

[2] A. Ramya and K. Rohini, "Comparative evaluation of machine learning classifiers with Obesity dataset," 2021 International Conference on Computing Sciences (ICCS), Phagwara, India, 2021,

[3] M. Dirik, “Application of machine learning techniques for obesity prediction: a comparative study,” Journal of Complexity in Health Sciences, Vol. 6, No. 2, pp. 16–34, Oct. 2023,

[4] S. Garg and P. Pundir, "MOFit: A Framework to reduce Obesity using Machine learning and IoT," 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2021

[5] Rodolfo Cañas Cervantes, Ubaldo Martinez Palacio,Estimation of obesity levels based on computational intelligence,Informatics in Medicine Unlocked,Volume 21,2020,