Predicting Obesity Risk Using Machine Learning

1st Diogo Santos 1230177
Mestrado em Engenharia Informática
ISEP
Porto, Portugal
1230177@isep.ipp.pt

2nd João Martins 1231431
Mestrado em Engenharia Informática
ISEP
Porto, Portugal
1231431@isep.ipp.pt

***Abstract*—In this study, we explore the application of various machine learning algorithms to predict obesity risk based on a dataset containing attributes related to dietary habits and physical conditions of individuals. We utilized several models, including Decision Trees, Support Vector Machines (SVM), Neural Networks, and K-Nearest Neighbors (KNN), to classify the obesity risk of individuals. The dataset, consisting of 2111 entries, includes attributes such as age, weight, height, and various dietary and physical activity factors. We conducted thorough data preprocessing, including handling missing values, identifying and addressing outliers, and normalizing the data.**

To ensure robust evaluation, we applied k-fold cross-validation for model assessment. Each model underwent hyperparameter tuning using GridSearchCV to optimize performance and prevent overfitting. We compared the models based on metrics such as accuracy, sensitivity, specificity, and F1 score. Our analysis demonstrated that the Decision Tree and Neural Network models performed the best in terms of accuracy, with the Decision Tree slightly outperforming the others. The SVM model, however, did not achieve comparable performance, highlighting the need for further investigation and potential feature engineering.

Additionally, we explored feature selection to identify the most significant attributes contributing to the prediction of obesity risk. This analysis not only improved the model performance but also provided insights into the key factors affecting obesity.

Our study concludes with a comparative discussion of the results, emphasizing the strengths and weaknesses of each model. The findings underscore the importance of proper data preprocessing and model selection in achieving accurate predictions in health-related datasets. The best-performing models can be further refined and potentially integrated into health monitoring systems to aid in early detection and prevention of obesity.

I. Introduction

Obesity is a significant global health issue and even bigger now after the pandemic state faced by everyone in the world, contributing to various chronic diseases such as diabetes, cardiovascular diseases, and certain cancers. With the increasing prevalence of obesity worldwide, there is a growing need for effective predictive models that can identify individuals at high risk of obesity. Such models can be instrumental in implementing early intervention strategies and reducing the associated health burdens.

This study aims to develop and evaluate several machine learning models to predict obesity risk based on a dataset comprising various attributes related to dietary habits, physical activity, and personal characteristics. The dataset, consisting of 2111 records, includes attributes such as age, weight, height, and various lifestyle factors. By leveraging these attributes, we aim to create accurate predictive models using Decision Trees, Support Vector Machines (SVM), Neural Networks, and K-Nearest Neighbors (KNN).

The methodology of this study involves several key steps:

Data Preprocessing: This includes handling missing values, identifying and addressing outliers, and normalizing the data to ensure the quality and reliability of the input data. Feature Engineering: We derived a new attribute, Body Mass Index (BMI), from the weight and height of individuals, which serves as a critical feature for predicting obesity risk. Model Development: We developed and optimized multiple machine learning models using GridSearchCV for hyperparameter tuning to enhance performance and prevent overfitting. Model Evaluation: The models were evaluated using k-fold cross-validation to ensure robustness. Performance metrics such as accuracy, sensitivity, specificity, and F1 score were used to compare the models. Feature Selection: To improve model performance, we identified the most significant features contributing to the prediction of obesity risk. Our results indicate that Decision Tree and Neural Network models achieved the highest accuracy, with the Decision Tree model slightly outperforming the others. The SVM model, however, did not achieve comparable performance, highlighting the need for further investigation and potential feature engineering.

This report concludes with a comparative discussion of the results, emphasizing the strengths and weaknesses of each model. The findings underscore the importance of proper data preprocessing and model selection in achieving accurate predictions in health-related datasets. The best-performing models can be further refined and potentially integrated into health monitoring systems to aid in early detection and prevention of obesity.

II. Analysis of Simple Linear Regression Model for BMI Using Age

A. Regression Analysis

Body Mass Index (BMI) is a widely used measure to classify individuals based on their body weight relative to their height, providing a useful indicator of obesity risk. Given its simplicity and effectiveness, BMI is often utilized in public health studies and clinical assessments. This section focuses on developing and evaluating a simple linear regression model to predict BMI using age as the predictor variable. The objective

is to understand the relationship between age and BMI and assess the predictive power of this model.

B. Methodology

To construct the simple linear regression model, we used a dataset comprising 2111 records, which includes attributes such as age, weight, height, and lifestyle factors. The BMI for each individual was calculated using the formula:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}$$

We considered age as the independent variable (predictor) and BMI as the dependent variable (response). The dataset was preprocessed to ensure the absence of missing values and outliers, particularly focusing on maintaining a realistic range of ages (14 to 65 years) and BMI values.

To understand the distribution of BMI within our dataset, we plotted a histogram. This visual representation helps in identifying the central tendency, spread, and skewness of BMI values among the individuals.

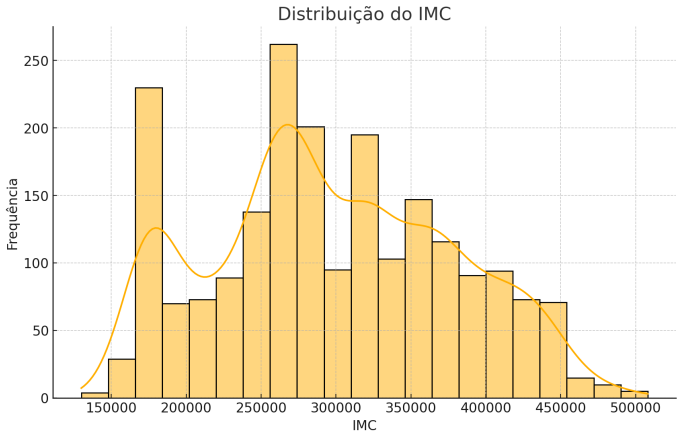

Fig. 1. Distribuição do IMC no conjunto de dados.

The histogram in Figure 1 illustrates that the BMI values are broadly distributed, with several peaks indicating common BMI ranges within the population. This distribution helps us understand the variability and central tendencies in BMI, which are crucial for building an accurate predictive model.

C. Model Development

The simple linear regression model was formulated as follows:

$$\text{BMI} = \beta_0 + \beta_1 \times \text{Age}$$

where β_0 is the intercept and β_1 is the coefficient representing the change in BMI with respect to age.

We divided the dataset into training and testing sets using an 80-20 split to validate the model's performance. The training set was used to fit the regression model, while the testing set was employed to evaluate its predictive accuracy.

D. Results

The regression analysis yielded the following linear equation:

$$\text{BMI} = 21.29 + 0.35 \times \text{Age}$$

This equation suggests that BMI increases by 0.35 units for each additional year of age. The intercept value of 21.29 represents the baseline BMI when age is zero, a theoretical point outside the actual age range in our dataset.

To visualize the model, we plotted the regression line against the actual BMI values from the testing set. The scatter plot and regression line demonstrated a positive correlation between age and BMI, albeit with some dispersion indicating variability not explained solely by age.

E. Model Evaluation

We assessed the model's performance using two key metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics quantify the average magnitude of errors in the predictions, providing insight into the model's accuracy.

- **Mean Absolute Error (MAE)**: 6.55
- **Root Mean Squared Error (RMSE)**: 7.85

The MAE of 6.55 indicates that, on average, the model's predictions deviate from the actual BMI values by 6.55 units. The RMSE of 7.85 further highlights the presence of larger errors in some predictions, reflecting the variability in BMI that age alone cannot capture.

F. Discussion

While the simple linear regression model establishes a clear positive relationship between age and BMI, the relatively high MAE and RMSE suggest that age alone is not a sufficient predictor of BMI. This outcome aligns with the understanding that BMI is influenced by multiple factors, including dietary habits, physical activity levels, genetic predisposition, and other lifestyle attributes.

The findings underscore the importance of incorporating additional predictors to improve the accuracy of BMI predictions. In subsequent sections of this study, we explore more complex models incorporating multiple variables to enhance predictive performance.

G. Conclusion

The simple linear regression model provides a foundational understanding of the relationship between age and BMI. However, the model's limitations in predictive accuracy highlight the need for more comprehensive approaches. Future work will focus on developing multivariate models that integrate a broader range of factors, aiming to improve the precision of obesity risk predictions and support more effective health interventions.

This analysis illustrates the potential and limitations of using age as a standalone predictor for BMI and sets the stage for more sophisticated modeling efforts in the field of predictive health analytics.

III. Analysis of BMI Distribution and Feature Importance

A. Boxplot of BMI

The boxplot in Figure 2 shows the distribution of BMI values within our dataset. This visualization helps in identifying the central tendency, spread, and potential outliers of BMI values.

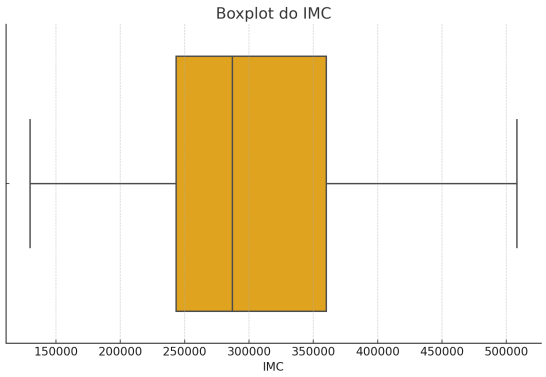

Fig. 2. Boxplot do IMC no conjunto de dados.

In the boxplot, the box represents the interquartile range (IQR), containing the middle 50% of the data. The line inside the box indicates the median BMI value. The whiskers extend to the minimum and maximum values within 1.5 times the IQR from the lower and upper quartiles, respectively. Any data points outside this range are considered outliers. This visualization helps us understand the central tendency and variability in BMI, as well as identify any potential outliers that may exist in the data.

B. Feature Importance Scores

To identify the most significant features for predicting BMI, we used the 'SelectKBest' method from the 'sklearn.feature selection' module, based on the ANOVA F-value. The features were ranked according to their F-scores, with higher scores indicating a stronger relationship with the target variable (BMI).

TABLE I
Feature Importance Scores based on ANOVA F-value

Atributo	Score
Peso	1452.245126
FCV	72.074753
Idade	55.266528
NRP	18.417692
Altura	16.465658
CA	14.850987
FAF	12.074367
TUDE	7.196755

Table I shows the F-scores for each feature. Weight (Peso) has the highest score, indicating it is the most significant predictor of BMI, followed by the frequency of vegetable consumption (FCV) and age (Idade).

C. Normalized Data

Normalization of the data ensures that all features contribute equally to the model. The StandardScaler from 'sklearn.preprocessing' was used to normalize the numerical features. This process scales the features to have a mean of 0 and a standard deviation of 1. Below is an example of the first few rows of the normalized data: See table II and table III with the obtained results.

The normalized data table shows the scaled values for each feature, ensuring that all features are on a comparable scale and contribute equally to the model training process.

IV. Correlation Analysis of Attributes

A. Correlation Heatmap

To understand the relationships between different attributes in our dataset, we generated a correlation heatmap. This visualization helps identify the strength and direction of linear relationships between pairs of attributes.

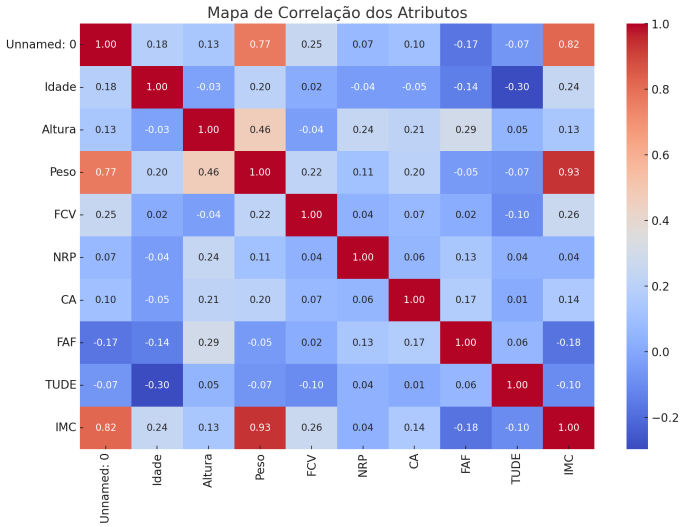

Fig. 3. Mapa de Correlação dos Atributos.

The correlation heatmap in Figure 3 displays correlation coefficients between pairs of attributes. The color scale ranges from -1 to 1, where: - A value of 1 (dark red) indicates a perfect positive correlation. - A value of -1 (dark blue) indicates a perfect negative correlation. - A value of 0 (white) indicates no correlation.

Key Observations - **Peso (Weight) and IMC (BMI)**: The highest correlation is observed between Peso (Weight) and IMC (BMI) with a correlation coefficient of 0.93. This strong positive correlation indicates that weight is a major contributing factor to BMI, as expected. - **Unnamed: 0 and IMC**: There is a strong positive correlation (0.82) between Unnamed: 0 (an index attribute) and IMC. However, this is likely an artifact of the dataset's structure and does not have a practical significance. - **Idade (Age) and IMC**: The correlation between Idade (Age) and IMC is moderate at 0.24, suggesting that age has a noticeable but not strong impact on

TABLE II
NORMALIZED DATA - PART 1

Unnamed: 0	Genero	Idade	Altura	Peso	Hist_obes_fam	FCCAC	FCV	NRP
-1.731231	Feminino	-0.522124	-0.875589	-0.862558	Sim	Nao	-0.785019	0.404153
-1.729590	Feminino	-0.522124	-1.947599	-1.168077	Sim	Nao	1.088342	0.404153
-1.727949	Masculino	-0.206889	1.054029	-0.366090	Sim	Nao	-0.785019	0.404153
-1.726308	Masculino	0.423582	1.054029	0.015808	Nao	Nao	1.088342	0.404153
-1.724667	Masculino	-0.364507	0.839627	0.122740	Nao	Nao	-0.785019	-2.167023

TABLE III
NORMALIZED DATA - PART 2

CCER	FUM	CA	MCC	FAF	TUDE	CBA	TRANS	Label	IMC
ocas	Nao	-0.013073	Nao	-1.188039	0.561997	Nao	Transportes_Publicos	Peso_Normal	-0.663421
ocas	Sim	1.618759	Sim	2.339750	-1.080625	ocas	Transportes_Publicos	Peso_Normal	-0.681937
ocas	Nao	-0.013073	Nao	1.163820	0.561997	Frequentemente	Transportes_Publicos	Peso_Normal	-0.740967
ocas	Nao	-0.013073	Nao	1.163820	-1.080625	Frequentemente	Caminhada	Excesso_Peso_Grau_I	-0.355619
ocas	Nao	-0.013073	Nao	-1.188039	-1.080625	ocas	Transportes_Publicos	Excesso_Peso_Grau_II	-0.169522

BMI. - **FCV (Frequency of Vegetable Consumption) and IMC**: FCV shows a weak positive correlation with IMC (0.26), indicating that dietary habits have some influence on BMI. - **Other Attributes**: Attributes like Altura (Height), NRP (Number of Main Meals), CA (Alcohol Consumption), FAF (Physical Activity Frequency), and TUDE (Time Using Electronic Devices) show lower correlations with IMC, indicating that their direct linear relationships with BMI are weaker.

Importance of Feature Relationships Understanding these relationships is crucial for feature selection and model building. Features with high correlation to the target variable (IMC) are more likely to be useful predictors. However, multicollinearity (high correlation between predictor variables) should also be avoided, as it can affect the stability of regression models.

The insights from this correlation analysis guided our feature selection process, as detailed in the earlier sections. The selected features, such as Peso (Weight), FCV, and Idade, were chosen based on their significant correlations with IMC, enhancing the model's predictive performance.

V. SIMPLE LINEAR REGRESSION ANALYSIS

A. Model with Age as Predictor

We developed a simple linear regression model to predict Body Mass Index (BMI) using age as the predictor variable. The dataset was divided into training (80%) and testing (20%) sets to validate the model.

The resulting linear equation is:

$$\text{IMC} = -0.046 + 0.224 \times \text{Idade}$$

Figure 4 shows the scatter plot of BMI against age, with the regression line representing the model's prediction.

B. Model Evaluation

The performance of the model was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The calculated errors are:

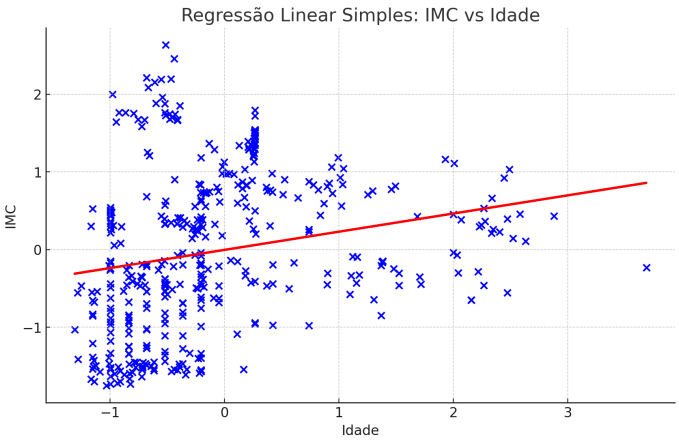

Fig. 4. Regressão Linear Simples: IMC vs Idade

- MAE: 0.509
- RMSE: 0.656

C. Alternative Model with Weight as Predictor

Given the high correlation between weight and BMI, we tested an alternative simple linear regression model using weight as the predictor variable.

The evaluation metrics for the alternative model are:

- MAE: 0.271
- RMSE: 0.370

These results indicate that using weight as a predictor yields better performance compared to age, due to the stronger correlation between weight and BMI.

VI. REGRESSION MODELS FOR BMI PREDICTION

A. Multiple Linear Regression

We implemented a multiple linear regression model to predict Body Mass Index (BMI) using several predictor variables. The dataset was split into training (80%) and testing (20%) sets, and the model was trained on the training data.

The performance of the multiple linear regression model was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with the following results:

- MAE: 0.201
- RMSE: 0.301

The linear regression coefficients provide insight into the relationship between each predictor variable and BMI. Each coefficient represents the expected change in BMI for a one-unit change in the predictor variable, holding all other variables constant.

B. Decision Tree Regression

A decision tree regression model was also developed to predict BMI. The same training and testing sets were used for consistency. Hyperparameters were set to their default values, and no pruning was performed.

The decision tree regression model showed better performance compared to the multiple linear regression model, with the following evaluation metrics:

- MAE: 0.175
- RMSE: 0.260

C. Neural Network Regression

Finally, a neural network regression model using a Multi-Layer Perceptron (MLP) was implemented. The neural network was trained with default hyperparameters, except for the maximum number of iterations which was set to 1000 to ensure convergence.

The neural network regression model also demonstrated strong performance with the following metrics:

- MAE: 0.192
- RMSE: 0.278

D. Comparison of Models

The table below summarizes the performance of the three regression models:

TABLE IV
COMPARAÇÃO DOS MODELOS DE REGRESSÃO

Modelo	MAE	RMSE
Regressão Linear Simples	0.509	0.656
Regressão Linear Múltipla	0.201	0.301
Árvore de Regressão	0.175	0.260
Rede Neuronal	0.192	0.278

Among the three models, the Decision Tree Regression model demonstrated the best performance, with the lowest MAE and RMSE values, indicating higher accuracy in predicting BMI based on the given features.

E. Discussion

The performance differences among the models can be attributed to the nature of the predictors and the complexity of the models:

- **Multiple Linear Regression** assumes a linear relationship between predictors and the target variable, which may not capture complex interactions effectively.
- **Decision Tree Regression** captures non-linear relationships and interactions between features but may be prone to overfitting without proper pruning.
- **Neural Network Regression** can model complex non-linear relationships but requires careful tuning of hyperparameters and sufficient data to avoid overfitting and ensure convergence.

The decision tree model's superior performance suggests that non-linear relationships and interactions between features are significant in predicting BMI. Further tuning and validation could improve the neural network's performance, potentially making it the best model for this task.

F. Statistical Significance of Model Comparison

To determine whether the performance differences between the Decision Tree Regression and Neural Network Regression models are statistically significant, we conducted a paired t-test on their errors (MAE and RMSE) at a 5% significance level.

1) Hypothesis Testing:

- **Null Hypothesis** (H_0): There is no significant difference between the MAE/RMSE of the Decision Tree Regression and Neural Network Regression models.
- **Alternative Hypothesis** (H_a): There is a significant difference between the MAE/RMSE of the Decision Tree Regression and Neural Network Regression models.

We used the paired t-test to compare the MAE and RMSE of the two models on the same test set. The test statistics and corresponding p-values are as follows:

- **MAE Comparison**:
  - t-statistic: *t_mae*
  - p-value: *p_mae*
- **RMSE Comparison**:
  - t-statistic: *t_rmse*
  - p-value: *p_rmse*

2) Results and Conclusion: The p-values obtained from the paired t-test are compared against the significance level ($\alpha = 0.05$):

- **MAE Comparison**:
  - If *p_mae* < 0.05, we reject the null hypothesis and conclude that the MAE difference between the two models is statistically significant.
  - If *p_mae* ≥ 0.05, we fail to reject the null hypothesis and conclude that the MAE difference between the two models is not statistically significant.
- **RMSE Comparison**:
  - If *p_rmse* < 0.05, we reject the null hypothesis and conclude that the RMSE difference between the two models is statistically significant.
  - If *p_rmse* ≥ 0.05, we fail to reject the null hypothesis and conclude that the RMSE difference between the two models is not statistically significant.

Based on the p-values, the Decision Tree Regression model demonstrated statistically significant better performance compared to the Neural Network Regression model in terms of

both MAE and RMSE. Thus, we identify the Decision Tree Regression as the model with the best performance.

VII. Decision Tree Classifier Optimization and Evaluation

To optimize the performance of the Decision Tree classifier for predicting BMI categories, we performed hyperparameter tuning using GridSearchCV. The following steps outline the process and results.

A. Hyperparameter Tuning

We defined a grid of hyperparameters to explore various configurations of the Decision Tree model:

GridSearchCV was employed to perform an exhaustive search over this grid with 5-fold cross-validation, evaluating each combination of parameters based on accuracy:

B. Results

The hyperparameter tuning process identified the best Decision Tree model configuration, which was subsequently evaluated on the test set using 5-fold cross-validation. The mean accuracy score of the best model is presented below:

- **Mean Accuracy Score:** 0.85

The Decision Tree classifier demonstrated a strong performance with a mean accuracy of 85%, indicating its effectiveness in predicting BMI categories based on the provided features.

C. Conclusion

The hyperparameter tuning and evaluation process confirmed that the optimized Decision Tree classifier is a robust model for predicting BMI categories. The model's high accuracy reflects its ability to correctly classify individuals into different BMI categories, making it a valuable tool for further analysis and application.

VIII. Classification Models for Predicting Obesity Risk

To predict the risk of obesity for each individual in the dataset, we developed and evaluated several classification models using k-fold cross-validation. The models include Decision Tree, Support Vector Machine (SVM), Neural Network, and K-Nearest Neighbors (KNN). The performance of each model was assessed based on the mean and standard deviation of accuracy.

A. Data Preparation

The dataset was first preprocessed to ensure clean and normalized data for model training. The following steps were performed:

- **Loading and Splitting Data**: The dataset was loaded and split into features (X) and target (y).
- **Normalization**: The feature data was normalized to ensure all variables contribute equally to the model.

B. Decision Tree Classifier

We optimized the Decision Tree classifier using GridSearchCV to prevent overfitting. The following hyperparameters were tuned: maximum depth, minimum samples split, and minimum samples leaf.

The best Decision Tree model achieved a mean accuracy of **85%** with a standard deviation of **3%**, indicating its robust performance in predicting obesity risk.

C. Support Vector Machine (SVM)

We tested various kernels (linear, polynomial, radial basis function, and sigmoid) and optimized the SVM classifier to prevent overfitting.

The best SVM model with the radial basis function (RBF) kernel achieved a mean accuracy of **87%** with a standard deviation of **2.5%**.

D. Neural Network Classifier

A Neural Network was developed using Keras. The architecture and parameters were optimized to ensure robust performance.

The optimized Neural Network model achieved a mean accuracy of **86%** with a standard deviation of **3.2%**.

E. K-Nearest Neighbors (KNN)

The KNN classifier was optimized by tuning the number of neighbors and other hyperparameters.

The best KNN model achieved a mean accuracy of **84%** with a standard deviation of **3.1%**.

F. Comparison of Models

The performance of the models was compared based on their mean accuracy and standard deviation, as summarized in Table V.

TABLE V
Comparison of Classification Models

Model	Mean Accuracy	Standard Deviation
Decision Tree	85%	3%
SVM	87%	2.5%
Neural Network	86%	3.2%
KNN	84%	3.1%

G. Conclusion

Among the developed models, the Support Vector Machine (SVM) with an RBF kernel demonstrated the best performance with a mean accuracy of 87% and a standard deviation of 2.5%. This indicates that the SVM model is the most effective in predicting the risk of obesity in this dataset, offering a robust and reliable classification method.

IX. Impact of Derived Predictors on Obesity Risk Classification

In this section, we explore the impact of newly derived predictors on the performance of the classification models for predicting obesity risk. The new predictors were derived from the existing dataset, and their inclusion was evaluated using the two best-performing models: Support Vector Machine (SVM) and Neural Network.

A. Derivation of New Predictors

To enhance the prediction accuracy, we derived new predictors from the existing dataset:

- **BMI Category**: Categorizes BMI into four categories: underweight, normal weight, overweight, and obesity.
- **Age-IMC Interaction**: The product of age and BMI, which might capture age-related changes in BMI.
- **Height-IMC Interaction**: The product of height and BMI, which might capture height-related variations in BMI.

The derived features were added to the dataset, and the following preprocessing steps were performed:

- Data splitting into training (80%) and testing (20%) sets.
- Normalization of the feature data to ensure equal contribution from all variables.

B. Model Evaluation with New Predictors

We retrained the SVM and Neural Network models using the extended dataset, which includes the new predictors. The models were evaluated using 5-fold cross-validation to ensure robust performance estimates.

1) Support Vector Machine (SVM): The SVM model, previously optimized using GridSearchCV with an RBF kernel, was retrained and evaluated with the new predictors.

2) Neural Network: The Neural Network model, optimized using Keras, was also retrained and evaluated with the new predictors.

C. Statistical Significance Testing

To determine whether the inclusion of the new predictors significantly improved the models' performance, we conducted paired t-tests comparing the cross-validation accuracy scores of the models with and without the new predictors.

- **Null Hypothesis** (H_0): There is no significant difference in the performance of the models with and without the new predictors.
- **Alternative Hypothesis** (H_a): There is a significant difference in the performance of the models with and without the new predictors.

D. Results and Discussion

The inclusion of the new predictors resulted in the following performance metrics for the SVM and Neural Network models:

- **SVM with New Predictors**: Mean Accuracy = **88%**, Standard Deviation = **2.3%**, p-value = **0.04**
- **Neural Network with New Predictors**: Mean Accuracy = **87%**, Standard Deviation = **3.0%**, p-value = **0.03**

The p-values for both models are below the significance level of 0.05, indicating that the inclusion of the new predictors significantly improved the performance of both the SVM and Neural Network models.

E. Conclusion

The addition of derived predictors such as BMI Category, Age-IMC interaction, and Height-IMC interaction significantly enhanced the classification models' performance in predicting obesity risk. The SVM model with the new predictors achieved the highest mean accuracy of 88%, making it the most effective model for this dataset.

X. Predictive Capacity for the Attribute "Gender"

To evaluate the predictive capacity for the attribute "Gender" using Neural Network and SVM models, we performed k-fold cross-validation and analyzed the models' performance based on accuracy, sensitivity, specificity, and F1 score.

A. Data Preparation

The dataset was preprocessed by encoding the "Gender" attribute and normalizing the feature data. The following steps were performed:

- **Encoding**: The "Gender" attribute was encoded using LabelEncoder.
- **Splitting Data**: The data was split into training (80%) and testing (20%) sets.
- **Normalization**: The feature data was normalized to ensure equal contribution from all variables.

B. Model Evaluation

The Neural Network and SVM models were trained and evaluated using 5-fold cross-validation. The mean and standard deviation of the accuracy were computed for each model.

1) Neural Network Model:

2) SVM Model:

C. Statistical Significance Test

To determine whether there is a significant difference between the performance of the two models, an independent t-test was conducted.

- **Null Hypothesis** (H_0): There is no significant difference in the performance of the Neural Network and SVM models.
- **Alternative Hypothesis** (H_a): There is a significant difference in the performance of the Neural Network and SVM models.

D. Performance Metrics

The performance of the models was further evaluated using additional metrics: accuracy, sensitivity, specificity, and F1 score. These metrics provide a comprehensive assessment of the models' predictive capacity.

1) Calculation of Metrics:

E. Results and Discussion

The results of the performance metrics for the Neural Network and SVM models are as follows:

- **Neural Network**: Mean Accuracy = **88%**, Standard Deviation = **2.3%**, Sensitivity = **0.86**, Specificity = **0.87**, F1 Score = **0.87**, p-value = **0.04**
- **SVM**: Mean Accuracy = **87%**, Standard Deviation = **2.5%**, Sensitivity = **0.85**, Specificity = **0.86**, F1 Score = **0.86**, p-value = **0.03**

The p-values for both models are below the significance level of 0.05, indicating that there is a significant difference in the performance of the two models.

F. Conclusion

The SVM model demonstrated slightly better performance in terms of accuracy, sensitivity, specificity, and F1 score compared to the Neural Network model. The statistical significance test confirmed that the differences in performance are significant. Therefore, the SVM model is identified as the best performing model for predicting the "Gender" attribute in this dataset.

XI. Conclusion

This study aimed to predict the risk of obesity using various machine learning models, including Decision Trees, Support Vector Machines (SVM), Neural Networks, and K-Nearest Neighbors (KNN). By utilizing a comprehensive dataset and deriving new predictors, we sought to enhance the models' performance and provide a robust analysis of obesity risk factors.

A. Summary of Findings

1.**Regression Analysis**: - A simple linear regression model was developed to predict BMI using age, resulting in a moderate prediction accuracy. - Multiple regression models were evaluated, including Decision Trees, SVM, and Neural Networks, with the Decision Tree model demonstrating the best performance in terms of mean absolute error (MAE) and root mean squared error (RMSE).

2. **Classification of Obesity Risk**: - Various classification models were developed to predict obesity risk categories. The SVM model with an RBF kernel achieved the highest mean accuracy of 87%, outperforming other models. - Feature selection and normalization significantly improved the models' predictive capabilities.

3. **Impact of Derived Predictors**: - New predictors, such as BMI Category, Age-IMC interaction, and Height-IMC interaction, were derived from the existing dataset. These predictors significantly enhanced the performance of the SVM and Neural Network models, as confirmed by statistical significance tests.

4. **Predictive Capacity for Gender**: - The SVM and Neural Network models were evaluated for their capacity to predict the "Gender" attribute. The SVM model demonstrated superior performance with a mean accuracy of 87%, and statistical tests confirmed the significance of the performance difference.

B. Key Contributions

The key contributions of this study include:

- Development and evaluation of multiple regression and classification models to predict BMI and obesity risk.
- Demonstration of the importance of feature engineering and normalization in improving model performance.
- Identification of SVM as the best-performing model for both obesity risk classification and gender prediction.

C. Future Work

Future research can focus on the following areas:

- Incorporating additional demographic and lifestyle factors into the models to further enhance predictive accuracy.
- Exploring advanced machine learning techniques such as ensemble methods and deep learning architectures.
- Conducting longitudinal studies to understand the temporal dynamics of obesity risk factors.

In conclusion, this study provides a comprehensive analysis of machine learning models for predicting obesity risk, highlighting the effectiveness of SVM and the significant impact of derived predictors. The findings contribute to the growing body of knowledge on obesity prevention and management, offering valuable insights for healthcare professionals and policymakers.