**Obesity Classification Application with Machine Learning on Central and**

**South American Peoples**

Timothy Kurniawan, John Pham, Timothy Smith II

**I. Introduction**

Obesity is a global health concern, with 43% of adults (18 years and older) overweight and 16% living as obese.1 While obesity causes are well researched in the U.S., research is limited in South and Central America where physiology and nutrition vary significantly. The following statistical research explores improvements to obesity level classification using machine learning (ML) models based on eating habits and physical condition. By testing different approaches and the combination of multiple methods in ensemble models, it is identified which techniques perform best under various conditions, focusing on both global and regional health patterns. Our model, trained on data from Mexico, Peru, and Colombia, is further tested on U.S. obesity datasets to identify potential cross-regional differences—findings that could significantly impact nutrition guidelines and health protocols. In addition, by testing American obesity data through the model which is trained on South American data, the results uniquely indicate whether statistical obesity health research between North and South America are cross applicable, a field which has not yet been explored before our research.

Several studies have explored the application of different ML techniques to predict obesity levels based on physical, lifestyle, and demographic data. Notably, ensemble learning methods such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) have demonstrated their ability to enhance predictive accuracy by aggregating multiple models to minimize overfitting and increase robustness.

One noteworthy study on obesity prediction used ensemble models combining RF and XGBoost on variables which were determined significant predictors through the Recursive Feature Elimination (RFE) method of feature selection.2 The researchers used the same dataset that our analysis used. This study used RF and XGBoost classifiers for their reliability and scaling capacity. Hyperparameter were tuned with Bayesian optimization to improve generalization. The ensemble learning model achieved 93.79% accuracy on predicting obesity.

A similar analysis utilized neural networks to predict obesity levels and identify key influencing factors. By comparing chi-square and F-classify feature selection methods, the authors demonstrated that F-classify yielded superior results in optimizing model performance. Bayesian optimization was employed to fine-tune hyperparameters, further enhancing predictive accuracy.3

**II. Data Set**

The dataset used in this study contains information for estimating obesity levels in individuals from Mexico, Peru, and Colombia, based on their eating habits, physical activity, and general health indicators.4 It includes 2111 instances and 16 input features, with one target variable (NObeyesdad) representing obesity classification. The dataset is multivariate and supports classification, regression, and clustering tasks, with all features stored as integers or encoded categories. Approximately 23% of the data was collected directly from users via a web platform, while the remaining 77% was synthetically generated using the SMOTE algorithm in Weka to balance class distributions. Importantly, the dataset contains no missing values. Table 1 provides a detailed description of each feature and target.

**Table 1: Dataset Features**

| **Variable** | **Type** | **Description** |
| --- | --- | --- |
| Gender | Categorical |  |
| Age | Continuous |  |
| Height | Continuous |  |
| Weight | Continuous |  |
| family\_history\_with\_overweight | Binary | Has a family member suffered with obesity |
| FAVC | Binary | Eats high caloric food frequently |
| FCVC | Integer | Usually eats vegetables in meals |
| NCP | Continuous | Main meals eaten daily |
| CAEC | Categorical | Eats food between meals |
| SMOKE | Binary | Smoker |
| CH2O | Continuous | Amount of water consumed daily |
| SCC | Binary | Monitors daily calories |
| FAF | Continuous | Physical activity frequency |
| TUE | Integer | Time spent using technological devices such as cell phones, video games, television, computer |
| CALC | Categorical | Frequency of drinking alcohol |
| MTRANS | Categorical | Primary transportation method |
| NObeyesdad | Categorical | Obesity level |

**III. Methodology**

This study examines Immersive obesity classification given health and genetic factors. The methodology that we adopt are listed below.

**1.** Handling Missing Values:

Rows containing missing values were dropped to ensure model training was based on complete and accurate data. After removing these rows, the target labels were updated to match the cleaned feature set indices to maintain correct alignment between inputs and outputs.

**2.** Encoding Categorical Variables:

Categorical features were converted to numerical representations to facilitate machine learning model compatibility. The following variables were encoded: family\_history\_with\_overweight, Gender, FAVC, CAEC, SMOKE, SCC, CALC, and MTRANS. Encoding was performed using simple label encoding, where each unique categorical value was assigned a unique integer.

**3.** Feature Scaling:

All numerical features were standardized using StandardScaler from scikit-learn. This scaling transformed features to have a mean of 0 and a standard deviation of 1, helping models like SVM and neural networks converge more efficiently and perform optimally.

**4.** Data Splitting:

The cleaned and preprocessed data was split into training and testing subsets using an 80/20 split. The training set was used for model fitting and hyperparameter tuning, while the testing set was reserved for final model evaluation to assess generalization performance.

**5.** Feedforward Neural Network (FFNN):

A feedforward neural network was constructed to classify obesity levels based on the input features. The network consisted of the following architecture: An input layer corresponding to the number of features, 2 hidden layers, each using the Rectified Linear Unit (ReLU) activation function to introduce non-linearity, and a final output layer using a softmax activation function to handle multi-class classification.

The model was compiled using a categorical cross entropy loss function, optimized with the Adam optimizer. Training was performed over a defined number of epochs, with monitoring of the validation loss to prevent overfitting. Model performance was evaluated using accuracy metrics, a confusion matrix, and a full classification report showing precision, recall, and F1-score for each class.

**6.** Feedforward Neural Network (FFNN):

A feedforward neural network was constructed to classify obesity levels based on the input features. The network consisted of the following architecture: An input layer corresponding to the number of features, 2 hidden layers, each using the Rectified Linear Unit (ReLU) activation function to introduce non-linearity, and a final output layer using a softmax activation function to handle multi-class classification.

**7.** Support Vector Machine (SVM):

An SVM classifier was trained to classify obesity levels using hyperparameter optimization with GridSearchCV. The grid search included tuning the following parameters: Kernel type (e.g., radial basis function (RBF)), Regularization parameter C, and Kernel coefficient gamma. Five-fold cross-validation was employed during grid search to ensure robustness in hyperparameter selection. The best model was selected based on cross-validation accuracy and was subsequently evaluated on the test set. Evaluation metrics included overall test accuracy and a classification report.

**8.** SVM with Reduced Feature Set:

Following initial SVM training, a reduced feature set is constructed by removing features deemed insignificant. A new SVM model is instantiated using the same optimized hyperparameters identified in the original grid search. This retraining aims to test whether model performance could be maintained or improved with fewer features. To further enhance robustness, the reduced model is embedded in a pipeline combining feature standardization with a BaggingClassifier ensemble. The base classifier was a linear SVM (C=10, gamma=1), wrapped in a bagging ensemble using 20 bootstrapped subsets, each trained on 80% of the data. The final pipeline is evaluated on both training and test sets using accuracy and classification metrics.

**9.** K-Nearest Neighbors (KNN):

A K-Nearest Neighbors classifier was implemented to provide a distance-based alternative approach to obesity classification. Hyperparameter tuning was performed via GridSearchCV, where the following parameters were explored: Number of neighbors (1 to 20), Weight function (uniform vs. distance), and Distance metric (euclidean vs. manhattan). Five-fold cross-validation was used during grid search to select the best combination of parameters. The best KNN model was then evaluated on the test set using accuracy metrics.

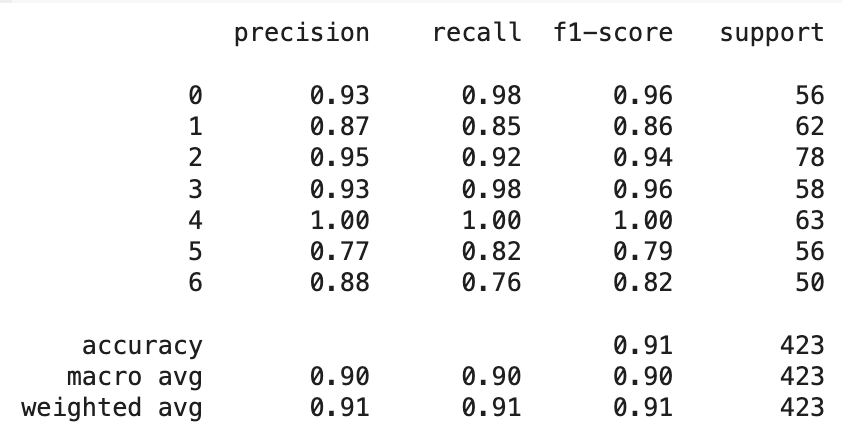
**10.** Testing the Peru Model on USA Dataset:

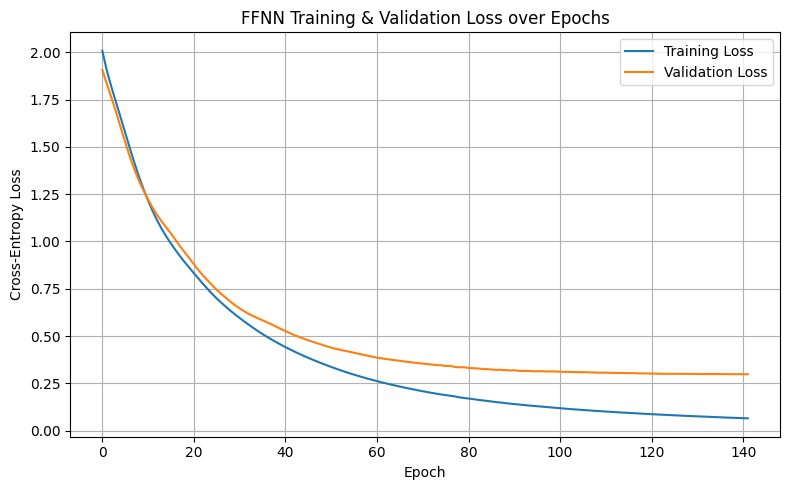
The SVC model, trained on the Peru dataset, was used to predict obesity levels on the USA dataset. The USA dataset was processed in the same way as the Peru dataset (i.e., missing value removal, label encoding, scaling), ensuring that the input features matched the format expected by the model. The performance of the Peru-trained model on the USA dataset was assessed by calculating the accuracy and comparing it with the results from the Peru test set. This transfer learning approach demonstrated the generalizability of the model across different geographical datasets.

**IV. Experiment Results**

**1.** Feedforward Neural Network (FFNN):

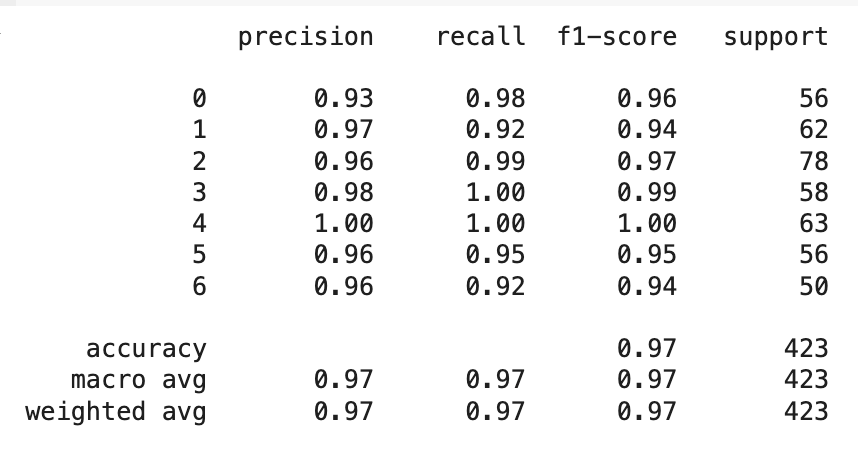
After setting the training loop to 1000 epochs, with early stopping if the loss doesn’t improve in 10 epochs, the following was achieved after stopping at epoch 142.





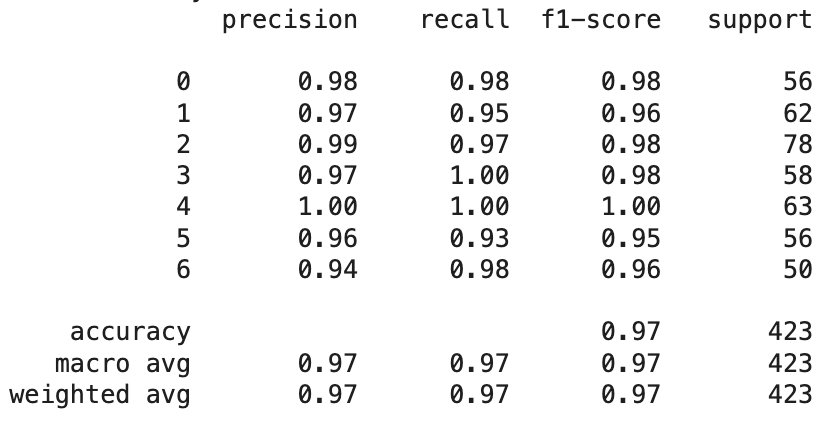
**2.** Support Vector Machine (SVM):

Hyperparameters: C:10, Gamma: 1, Kernel:Linear



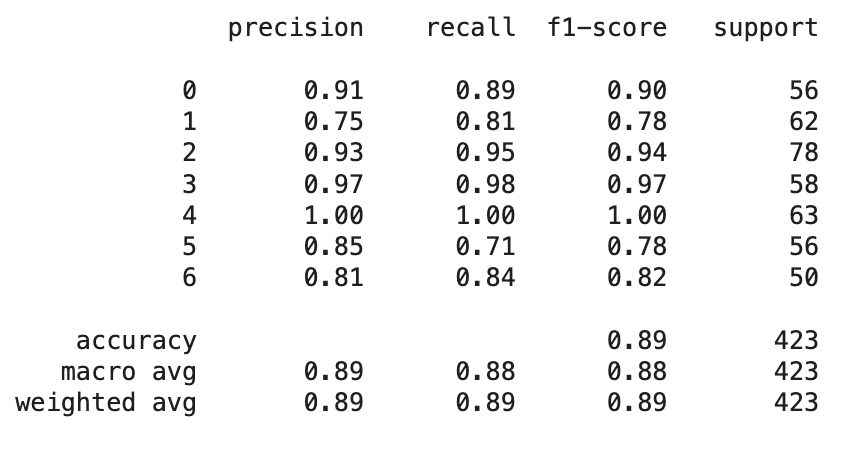
**3.** SVM with reduced feature set:

Hyperparameters: C:10, Gamma: 1, Kernel:Linear



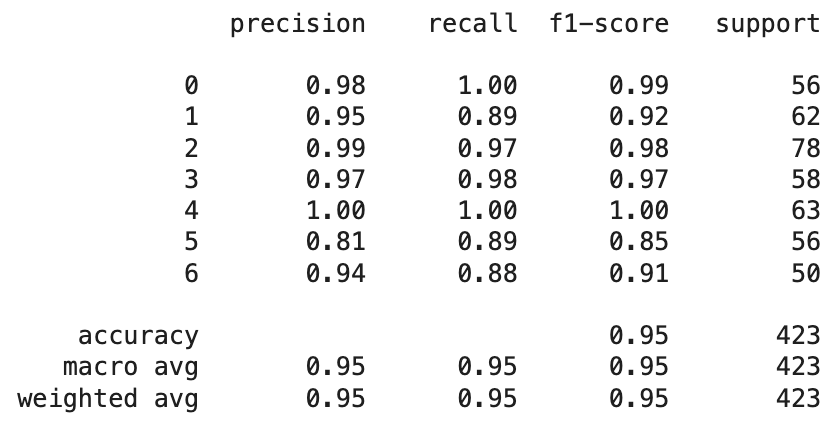
**4.** KNN:

Hyperparameters:N:3,Weights:Uniform,Metric: Manhattan

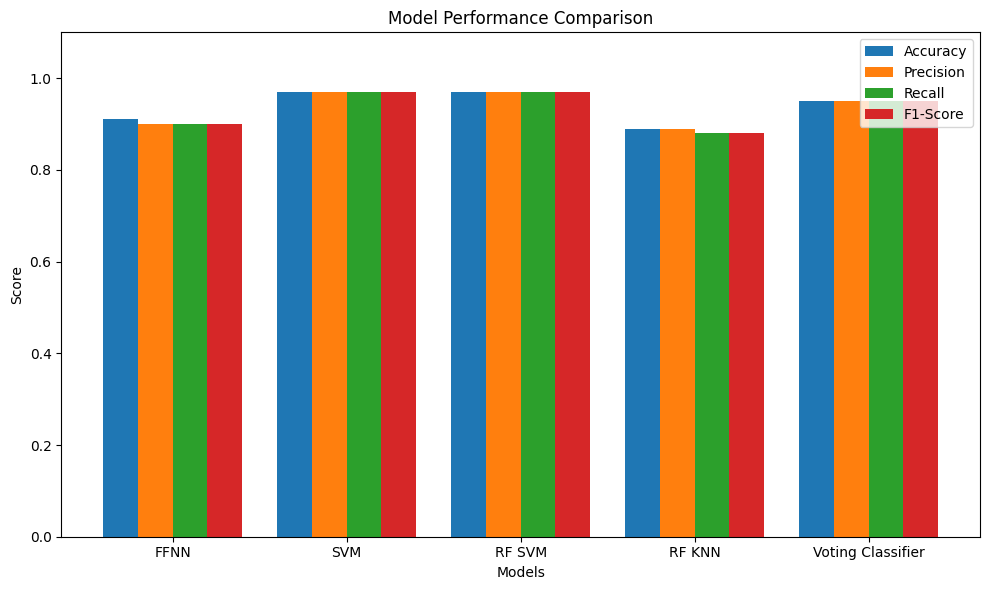


**5.** Ensemble voting model:

Soft voting, Weighted 3:1:1, SVM:KNN:FFNN



**6.** Comparing between models



**7.** Test on USA Dataset:

The SVM model achieved 96.7% test accuracy on the Peru dataset and 95.3% on the USA dataset, demonstrating strong cross-geographical generalization. While key features remained consistent, FCVC (vegetable consumption) showed significantly higher importance in the Peru data, suggesting regional differences in dietary impact on obesity classification.

**V. Conclusion**

This project explored the use of machine learning techniques to classify obesity levels in individuals from Central and South America, with a focus on identifying the most effective models and evaluating their generalizability across regions. Among the models tested, the Support Vector Machine (SVM) demonstrated the highest classification accuracy, achieving 96.7% on the Peru dataset and an impressive 95.3% accuracy when applied to the U.S. data. This strong performance suggests that statistical models trained on regional health and lifestyle data can generalize effectively on other populations, a finding that holds important implications for health research. Furthermore, the analysis revealed that certain features, such as vegetable consumption (FCVC), held varying levels of importance between regions, indicating that while general patterns exist, regional dietary and lifestyle behaviors must be considered when designing health interventions. Despite the high accuracy and strengths of the models, limitations such as reliance on partially synthetic data and limited geographic scope suggests the need for further training and validation. Future research should focus on expanding the dataset to include more countries, features, etc. Testing with other advanced ensemble and deep learning models would also enhance classification robustness. Ultimately, this study demonstrates the promising role of machine learning in advancing public health diagnostics and guiding data-driven health policies.

**References**

| [1] | “Obesity and overweight,” World Health Organization, Mar 2024,  https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight |
| --- | --- |
| [2] | G. N. Nikhita, P. S. Varma, S. P. Teja and G. S. Rao, "Obesity level prediction using ML based on eating habits and physical condition," 2024 9th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2024, pp. 1811-1816, doi: 10.1109/ICCES63552.2024.10860098. https://ieeexplore.ieee.org/document/10860098 |
| [3] | F. H. Yagin, M. G·· ul“ u, Y. Gormez, A. Casta- neda-Babarro, C. Colak, G. Greco, et al., "Estimation of obesity levels with a trained neural Network Approach optimized by the bayesian technique", Applied Sciences, vol. 13, no. 6, pp. 3875, 2023. |
| [4] | “Estimation of Obesity Levels Based On Eating Habits and Physical Condition,” UC Irvine Machine Learning Repository, 2019, https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition |