Empowering Smart Health Apps With Machine Learning

Sanjana Arora: sanjanaa@uvic.ca

Motivation and Overview

Nowadays people use their smartphones to keep a track of the number of calories consumed during the day, number of steps taken, kind of meals consumed, etc. Based on the collected data, a person's obesity level is predicted. Further, using all the collected data, a user's physical activity is classified into six categories being performed.

Datasets

The first dataset used is from UCI Machine Learning Repository which has Obesity levels in individuals from the countries of Mexico, Peru. This Dataset is based on their eating habits and physical condition of these individuals. Dataset contains 17 attributes and 2111 records.

Second dataset is the Human Activity Recognition database. It was built from the recordings of 30 study participants performing daily physical activities like standing, walking etc., while carrying a waist-mounted smartphone with embedded sensors. This dataset has 563 columns and 10299 records.

Data Preprocessing

Correlation

Correlation heat maps are used to remove features providing redundant information.

Data Normalization

Minmax scalar normalization technique is used for normalizing the range of features.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Dimensionality Reduction

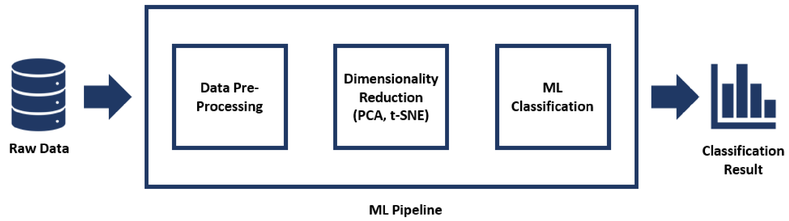
Two dimensionality reductions are applied including Principal Component Analysis and t- Distributed Stochastic Neighbor Embedding

One-Hot Encoding

One Hot Encoding is done on categorical features.

Experiments

- **Metrics:** Training Accuracy, Testing Accuracy and Validation Accuracy
- **Models Implemented:** XGBoost, Random Forest, SVM
  - XGBoost achieves highest training and testing accuracies.
- Regularization improved the XGBoost classifier performance.
- RandomForest classifier using label powerset performed better in comparison with classifier chains of Naïve Bayes algorithm for classifying the data into different age groups.



- **Dimensionality Reduction:** Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).
- **Models Implemented**: AdaBoost, Random Forest, XGBoost
- **GridSearch – 5-Fold Cross Validation** is implemented for hyper-parameter tuning.
- XGBoost Outperform all the implemented algorithms.
- T-SNE is used to visualize clusters in 2-D

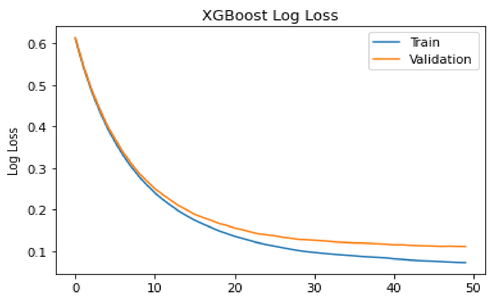


Figure 1. XGBoost Log Loss on Obesity Dataset.

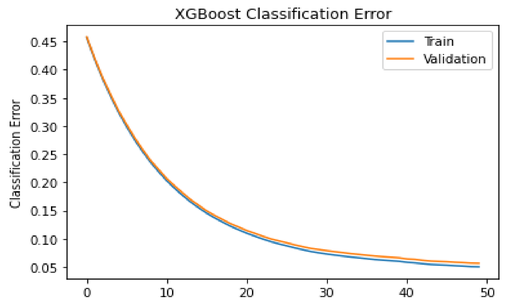


Figure 2. XGBoost Classification Error on Obesity Dataset.

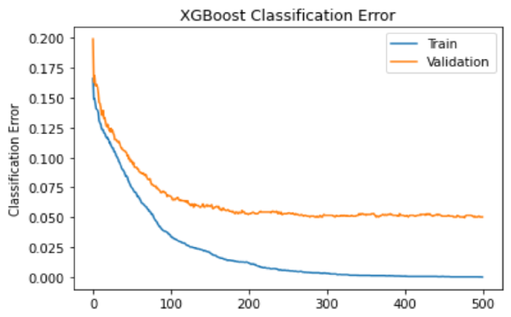


Figure 3. XGBoost Classification Error on Human Activity Recognition.

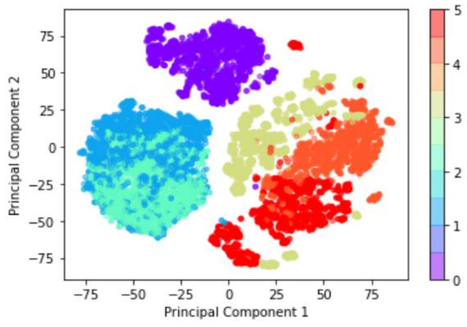


Figure 4. t-SNE Clusters on Human Activity Recognition .

Model Summary

Model Name	Train Accuracy	Validation Accuracy	Test Accuracy
SVM	95.41%	95.97%	95.03%
Random Forest	100%	97.39%	96.92%
XGBoost	100%	97.63%	95.98%
RF (Label Powerset)	100%	98.81%	98.34%
KMeans	26.85%	25.11%	26.71%

Approach	Model Name	Train Accuracy	Validation Accuracy	Test Accuracy
PCA	**AdaBoost on PCA Data**	55.34%	53.26%	51.91%
	Random Forest on PCA Data	100%	94.22%	89.44%
	XGBoost on PCA Data	100%	94.96%	91.14%
t-SNE	**Random Forest on t-SNE Data**	100%	94.83%	84.83%
	XGBoost on t-SNE Data	95.06%	91.29%	82.89%

Results

The dataset is divided into 60% training, 20% testing and 20% validation data.

For the consumer health analysis, both the ensemble machine learning techniques random forest and XGBoost classifier achieve 100% training accuracy.
Unsupervised clustering machine learning algorithm, KMeans, performs very poorly on the dataset.

For Human Activity Recognition dataset, XGBoost achieves ~91% accuracy on test data. Training is done on PCA reduced dataset. While, t-SNE results are visualized, and clusters can be seen.

Future Works

In the future we wish integrate smart bands and smart shoes sensor data for further improving the accuracy of our project.

References

[1] https://unsplash.com/s/photos/smartband
[2] https://www.reddit.com/r/Honor/comments/jo3ydr/what_is_your_favorite_home_workout_wear/
[3] https://miro.medium.com/max/1482/0*Srg7htj4TOMP5ldX.png