Empowering Smart Health Apps with Machine Learning

Tavanpreet Singh Oberoi
Electrical and Computer Engineering
University of Victoria
Victoria, Canada
tavanpreetoberoi@uvic.ca

Siddharth Chadda
Electrical and Computer Engineering
University of Victoria
Victoria, Canada
sidchadda@uvic.ca

Sanjana Arora
Electrical and Computer Engineering
University of Victoria
Victoria, Canada
sanjanaa@uvic.ca

Hridya Divakaran
Electrical and Computer Engineering
University of Victoria
Victoria, Canada
hridyadivakaran@uvic.ca

Abstract

Nowadays, people use their smartphones to collect data relating to the number of steps taken, the number of calories consumed, the kind of meals taken, the amount of water taken, etc. This paper and project involve deriving insights from the population survey and using them to detect human activity that is being performed. For example, a person's walking, running, sitting, etc. is recognized by using the collected data through a smartphone. This can help as an assistive technology in the healthcare sector and eldercare when combined with other technologies like the Internet of Things (IoT). In this paper, we discuss various machine learning algorithms to derive different kinds of information from a given dataset and compare the performances of each of the algorithms.

1 Introduction

Obesity and overweight are defined as excessive fat buildup in specific body areas that can be damaging to health. The number of people suffering from obesity has more than doubled since 1980, and in 2014, more than 1900 million adults, 18 years old or older, are suffering from weight change. Obesity interventions can be spatially tailored if high-risk groups are recognized. Some of the reasons for obesity include an increase in the consumption of energy-dense, high-fat foods, as well as a decrease in physical activity due to the nature of the sedentary job, new transportation modes, and increased urbanization. This project is split into two halves. The first phase calculates the level of obesity based on a population survey, and the second phase uses this calculation to identify human activity such as walking, standing, sitting etc.

The first phase of the project focuses on identifying and classifying people based on the type of obesity they have, the method of transportation they use, and their age group, utilizing various machine learning classification approaches such as support vector machine and random forest classification. In the second stage, we used Principal Component Analysis and t-SNE to lower the dimension of the dataset in order to forecast human activity, utilizing a variety of machine learning and data mining strategies based on ensemble learning techniques.

1.1 Problem Definition and Previous Work

A number of e-health applications are available in the market to collect user data in terms of their eating habits and other personal data. However, there are comparatively few solutions available in the industry that can recognize human activity using the collected data. Our project achieves the dual objectives of analyzing and understanding the survey data of the general population of an area and then using the analysis to recognize the human activity. This solution can assist the eldercare and healthcare sectors to keep track of their patients' activities and also serve as an interesting feature for the consumer to keep track of their activities.

Obesity has been a study topic of interest, with many studies focusing on the variables that cause the disorder. Here we have done some review of works proposed by different authors that implement data mining techniques on datasets related to health issues. A logistic regression model was developed by Davila-Payan et al. (2015) to evaluate the probability of body mass index on children aged 2 to 17 years old in small geographic areas. Their findings revealed that estimations in small geographic areas are necessary for developing successful interventions and assisting in the design of potential solutions to the problem. A paper was formed by Girija Chetty and Mathew white to compare traditional naive Bayes classifier and unsupervised k-means clustering approaches for processing smartphone sensor signals for activity recognition with several new machine learning and data mining approaches based on decision trees and ensemble learning techniques such as random forests and random committee. The proposed methods' experimental evaluation with a publicly accessible smartphone activity recognition database considered improvement in recognition performance of proposed machine learning algorithms.

1.2 Input Data

The data comes from UCI Machine Learning Repository and includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records. The records are labelled with the class variable NObesity (Obesity Level), which allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III [1].

The Human Activity Recognition dataset 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 inertial sensors. This dataset has 563 columns and 10299 records [2].

For each record in the dataset the following is provided:

- Triaxial acceleration sensor data and the estimated body acceleration.
- Triaxial Angular velocity from the gyroscope.
- A 561-feature vector with time and frequency domain variables.
- Its activity label.

2 Approach

2.1 Consumer Health data analysis

2.1.1 Data Pre-processing

As a very first step of the data pre-processing, we checked for the presence of missing values in the dataset. Further, we observed that the age column of the dataset included some floating values, so we changed the floating values of the age column into integer values for simplification and formed bins of the age groups. To check for the presence of outliers in the dataset, we used the boxplot visualization. Age column included some outliers; however, age between 40-60 is not necessarily outliers, therefore, the outliers in age weren't treated. Further, we used a correlation heat map to identify the features that are highly correlated and remove those features. The columns Age and TUE have high correlations and therefore, we dropped the TUE feature. Finally, we performed the data standardization using the MinMax scaler technique to normalize the ranges of the features so that each of the features contributes equally to the classification.

2.1.2 Support Vector Machines

A support vector machine (SVM) is a type of supervised machine learning classification algorithm. When compared to other machine learning algorithms, SVMs are implemented in a unique approach. Support Vector Machines are not new but are still a powerful tool for classification due to their tendency not to over fit thus we used SVM to classify individuals based on their obesity level .There are 7 type of obesity level in our dataset. Once the model is trained we used the model to predict the accuracy.

2.1.3 Random Forest

Random Forest is a classification task-focused supervised machine learning method. Random Forest is a member of the ensemble algorithm family. Using an ensemble of Decision Trees trained using the "bagging" approach, the Random Forest algorithm creates a "Forest." The primary principle behind "bagging" is that merging different machine learning models improves the overall accuracy of the outcome. To put it another way, the random forest method combines the results of numerous decision trees to provide a high-accuracy classification. Thus for the second part we used random forest for classifying individuals based on the transportation mode .We used this to understand about the physical activity. The graphs were plotted for better visualization.

2.1.4 XGBoost Classifier

XGBoost stands for eXtreme Gradient Boosting, it is also a decision tree-based supervised ensemble machine learning algorithm which uses gradient descent boosting for optimization. This algorithm has been used to classify whether or not people monitor their calories consumption. The XGBoost algorithm was developed at the University of Washington, the algorithm is extremely optimized for solving classification problems. The algorithm achieves this by using a combination of parallel-processing-based tree pruning, regularization, and the gradient descent algorithm to minimize the classification error.

The figures below capture the logloss and classification error evaluation metrics of the XGBoost classifier with the regularization parameter.

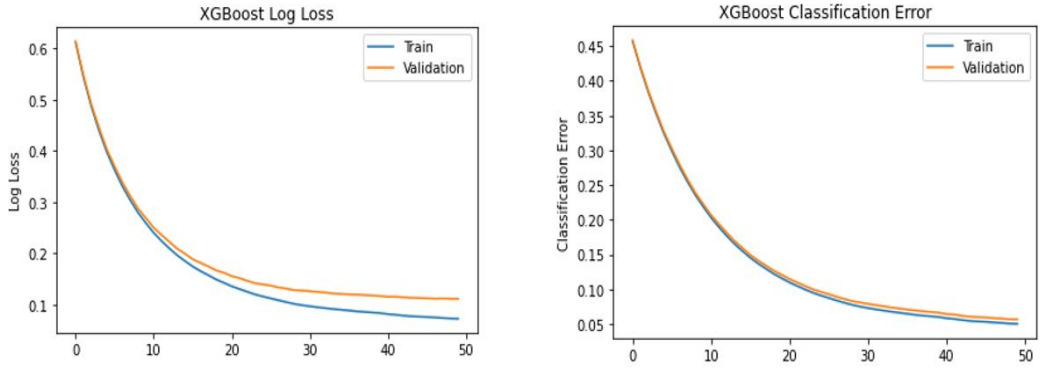

Figure 1: XGBoost Results on Obesity Dataset.

2.1.5 K-means Clustering

For classifying the dataset into whether or not people monitor their calories consumptions we have also used an unsupervised clustering algorithm called Kmeans. This algorithm assumes that there are k clusters and assigns the data to the clusters with the nearest mean by minimizing the mean squared error.

2.1.6 Classifier Chains of Naive Bayes Algorithm

To classify the population into 4 different age groups, we have used classifier chains of the Naïve Bayes Algorithm. Since the classification of one-hot encoded age groups is a multi-label classification

problem; we have used chains of Naïve Bayes classification algorithm to combine the computational efficiency of the binary relevance method [3]. However, the classifier chains of the Naïve Bayes Algorithm did not yield very good results as captured in the results section. Therefore, we tried another method for solving the multi-label classification of the population into four different age groups.

2.1.7 Random Forest using Label PowerSet

To solve the multi-label classification problem, we have used another approach called the label powerset method with the Random Forest classifier [4]. Label PowerSet is a problem transformation approach to multi-label classification that transforms a multi-label class problem with a multi-class classifier trained on unique label combinations found in the training sets. Therefore, by combining the label powerset with random forest classifier we have combined two powerful techniques to perform multi-label classification and hence, we have achieved good accuracy results as discussed in the results section.

2.2 Machine Learning Algorithms to recognizing Human Activity

To tackle the problem of Activity Recognition using machine learning we have implemented two different approaches.

2.2.1 Principal Component Analysis (PCA)

The multi-dimensional Human Activity Recognition dataset is used having 563 features, thus in order to reduce the complexity of the dataset, Principal Component Analysis for dimensionality reduction is implemented. The objective of PCA-Principal Component Analysis is to reduce the dimensionality of large datasets by transforming large number of features into smaller one that retains most of the information the large number of features wants to convey. Through PCA the dimensionality of the dataset is reduced to just 34 features which explains 90 percent variance of the dataset. Further, Classification algorithm including AdaBoost, Random Forest and XGBoost are applied, and the results are discussed. The hyper-parameters are tuned using GridSearchCV [5].

GridSearch aids in fitting your estimator (model) to your training data by looping through predefined hyperparameters. As a result, the best hyperparameters from the list are chosen in the end. CV stands for 'Cross Validation'. Cross-validation is a resampling technique for evaluating machine learning models on a small sample of data. The process includes only one parameter, k, which specifies the number of groups into which a given data sample should be divided. As a result, the process is frequently referred to as k-fold cross-validation. Here 5 fold cross-validation is used to evaluate the results.

The figures below capture the logloss and classification error evaluation metrics of the XGBoost classifier with the regularization parameter.

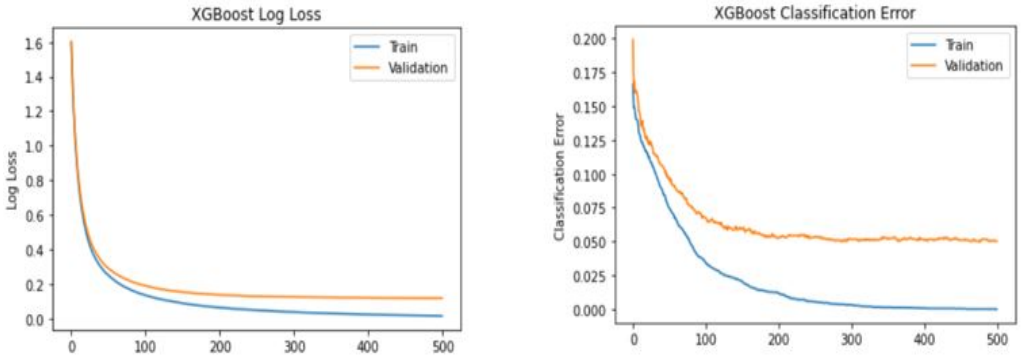

Figure 2: XGBoost Results on Human Activity Dataset.

2.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Further, t-SNE is applied on the PCA reduced dataset to visualize the activity clusters by converting 34 dimensional dataset to 2 dimensional dataset [6]. It is the non-linear dimensionality reduction algorithm, used to explore high-dimensional data. It converts multidimensional data into two or more dimensions that humans can see. Further, Classification algorithms Random Forest and XGBoost are applied, and the results are discussed. AdaBoost is discarded since the results on the PCA reduced dataset are bad, and the 2-Dimensional dataset is projected to lose more accuracy.

The figure below visualizes the human activity clusters in 2-dimensions.

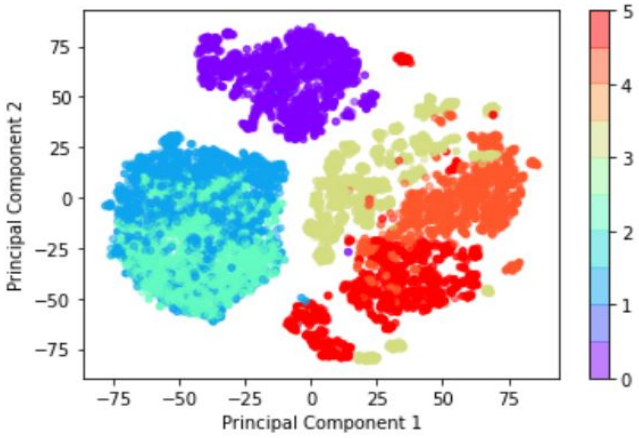

Figure 3: Visualization on t-SNE reduced Human Activity Dataset.

3 Results

3.1 Classifying area population based on personal data

The machine learning algorithms are applied to the dataset that is split into 60% training data, 20% testing data and 20% validation data. We used various machine learning algorithms for classifying different target variables such as classifying the dataset based on obesity levels, whether or not a person keeps track of calories, and identifying the age group of the population. Ensemble machine learning algorithms such as Random Forest and XGBoost achieve an accuracy of 100% for training dataset, approximately 96.9% accuracy on testing dataset and $\sim 97.3\%$ validation accuracy. Unsupervised Kmeans clustering algorithm achieved an accuracy of approximately 25% and therefore, this algorithm has not been able to perform well on the given dataset.

The figure below describes the results observed from different classification algorithms.

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%

Figure 4: Results for consumer health data analysis.

3.2 Recognizing Human Activity using data recorded in Smart Phones

The three classification techniques are implemented on PCA reduced dataset, while best two are implemented on t-SNE reduced dataset. The dataset was already split into training and testing data. Further training dataset is split into training and validation sets with 80% as train data and 20% as validation data. The split is done using sklearn library train_test_split function. Validation dataset is used for initial testing, so that if the model performs well on the initial tests, then only testing set is exposed to the model for prediction.

AdaBoost is implemented on the PCA reduced dataset giving training error of $\sim 45\%$. It could be because of inclusion of irrelevant features in the dataset. To improve the accuracy, hyper-parameters are tuned, but the results remain the same. Hence, AdaBoost with such low accuracy is discarded for further experiments.

Next, Random Forest Classifier is applied on the PCA reduced dataset and all the accuracies have been improved. To further improve the accuracy, grid search is applied, and the best tuned parameters are selected. But Random Forest Classifier takes minutes to train on the dataset and to overcome this challenge, XGBoost is applied which yields the most optimized results with testing accuracy of $\sim 91\%$.

To better visualize the dataset, t-SNE is implemented on the PCA reduced dataset. It can be observed from the plots of t-SNE model with 2 components that segregated classes can be identified. Though reducing the dimension also impact the accuracy of the models and testing accuracy has been decreased to $\sim 82\%$ from $\sim 91\%$ for XGBoost as best classifier. It can be said that t-SNE can be applied to visualize the dataset, though models must be trained on PCA reduced dataset.

The figure below describes the results observed from different classification algorithms.

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%

Figure 5: Results for recognizing human activity.

4 Contributions

Siddharth Chadda was responsible for identifying the problem and finding a suitable dataset. Tavanpreet Oberoi compiled the report in Overleaf. All the four authors contributed equally to writing the project report sections, identifying suitable machine learning algorithms and poster making. Finally, Siddharth Chadda presented the poster.

References

[1] https://www.kaggle.com/ankurbajaj9/obesity-levels?select=ObesityDataSet_raw_and_data_sinthetic.arff

[2] https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones

[3] http://scikit.ml/api/skmultilearn.problem_transform.lp.html

[4] https://scikit-learn.org/stable/auto_examples/multioutput/plot_classifier_chain_yeast.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[6] https://medium.com/analytics-vidhya/pca-vs-t-sne-17bcd882bf3d