HUMAN ACTIVITY RECOGNITION

**Logo

Description automatically generated**

**CLARK UNIVERSITY**

**School of Management**

Prof.Yue Gao

**STAT4650 Machine Learning**

**Project Proposal**

Class Section: 01

**Group Members:**

Bargavi Kongara

**The Research Problem:**

Human activity recognition is a field of study that aims to identify and understand human behavior through analysis of data generated by various sensors and sources, such as wearable devices, cameras, and smart home systems. Despite its potential for numerous applications, there are several challenges and research problems associated with human activity recognition:

* Data quality and consistency: Ensuring that the data collected from different sensors is accurate, consistent, and of high quality is a major challenge in human activity recognition.
* Activity recognition accuracy: Developing algorithms and models that can accurately recognize a wide range of human activities is a major research problem in the field.
* Scalability and generalizability: Developing systems that can recognize activities in real-world scenarios and adapt to different users and environments is a challenge.
* Privacy and security: Ensuring that personal data collected through human activity recognition systems is protected and secure is a critical research problem.
* Integration with other systems: Integrating human activity recognition systems with other systems, such as health monitoring systems, smart homes, and security systems, is a challenge.
* Ethical considerations: Ensuring that human activity recognition systems are used in an ethical and responsible manner, and that personal data is protected, is a critical research problem.

**Objective:**

The objective of our project is to develop a system that can automatically recognize and classify human activities based on data collected from sensors such as accelerometers, gyroscopes, or cameras. Our goal is to develop algorithms that can accurately identify different activities, such as walking, running, sitting, or laying and provide insights into human behavior and physical activity. Additionally, we want to develop a system that can provide personal health and wellness information, monitor employee performance, or inform home automation systems.

**Business Problem:**

Human activity recognition projects using machine learning are aimed at solving various business problems. In the field of health and wellness, the technology can be used to monitor and track physical activity levels, provide insights into personal health and wellness, and improve overall health outcomes. In the workplace, human activity recognition can be used to monitor employee performance in various settings, such as call centers or manufacturing plants, and improve productivity and efficiency. In the home, the technology can be used for home automation, by automating various tasks based on the presence and activity of the occupants.

Additionally, human activity recognition can be used for personal safety, by monitoring the movements of elderly or vulnerable individuals and alerting carers or family members in case of a problem. The technology can also be used for fraud detection, by detecting and preventing fraudulent activities in various settings such as banking, insurance, and retail. In sports, human activity recognition can be used to analyze and improve the performance of athletes by monitoring their movements and providing insights into their training and competition performance.

**Preliminary list of Dataset:**

The Human Activity Recognition database consists of data collected from 30 individuals performing daily tasks, each carrying a smartphone with motion sensors attached to their waist. The goal is to categorize the actions into six different categories.

Data source: <https://www.kaggle.com/code/essammohamed4320/human-activity-recognition-scientific-prespective/notebook>

The type of dataset that will be used in the project is csv format and we have numerical and categorical data in our dataset. There are no missing values in the dataset. Hence, we are considering the entire data set which has 7352 rows and 563 fields.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Description** | **Data Type** |
| tBodyAcc-XYZ | Body acceleration in X, Y, Z dimensions in the time domain | Int64 |
| tGravityAcc-XYZ | Gravity acceleration in X, Y, Z directions in the time domain | Int64 |
| tBodyAccJerk-XYZ | Body acceleration jerk in X, Y, and Z directions in the time domain | Int64 |
| tBodyGyro-XYZ | Body Angular speed in X, Y, and Z directions in the time domain | Int64 |
| tBodyGyroJerk-XYZ | Body angular acceleration in X, Y, Z directions in the time domain | Int64 |
| tBodyAccMag | Body acceleration Magnitude in the time domain | Int64 |
| tGravityAccMag | Gravity acceleration magnitude in the time domain | Int64 |
| tBodyAccJerkMag | Body acceleration jerk magnitude in the time domain | Int64 |
| tBodyGyroMag | Body angular speed magnitude in the time domain | Int64 |
| tBodyGyroJerkMag | Body Angular Acceleration Magnitude in the time domain | Int64 |
| fBodyAcc-XYZ | Body acceleration in X, Y, Z directions in the frequency domain | Int64 |
| fBodyAccJerk-XYZ | Body acceleration jerk in frequency domain | Int64 |
| fBodyGyro-XYZ | Body angular speed in the frequency domain | Int64 |
| fBodyAccMag | Body acceleration magnitude in the frequency domain | Int64 |
| fBodyAccJerkMag | Body acceleration jerk magnitude in the frequency domain | Int64 |
| fBodyGyroMag | Body angular speed magnitude in the frequency domain | Int64 |
| fBodyGyroJerkMag | Body Angular Acceleration Magnitude in the frequency domain | Int64 |
| Activity | Activities performed by humans such as walking, walking upstairs, walking downstairs, sitting, standing, laying | Object |
| Subject | data recorded directly on the subject by carrying custom hardware or smartphones with accelerometers and gyroscopes. | Object |

This dataset contains one dependent variable, activity which is the final observation of human activities such as walking, standing, walking upstairs, walking downstairs, etc. The remaining variables such as 3-axial linear acceleration and 3-axial angular velocity measurements are independent variables.

The set of variables that were estimated from these signals are:

* mean(): Mean value
* std(): Standard deviation
* mad(): Median absolute deviation
* max(): Largest value in the array
* min(): Smallest value in the array
* sma(): Signal magnitude area
* energy(): Energy measure. Sum of the squares divided by the number of values.
* iqr(): Interquartile range
* entropy(): Signal entropy
* arCoeff(): Auto regression coefficients with Burg order equal to 4
* correlation(): correlation coefficient between two signals
* maxInds(): index of the frequency component with the largest magnitude
* meanFreq(): Weighted average of the frequency components to obtain a mean frequency
* skewness(): skewness of the frequency domain signal
* kurtosis(): kurtosis of the frequency domain signal
* bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.
* angle(): Angle between two vectors.

**Applications of the Dataset:**

Objective of data set is to classify activities into one of the six activities performed.

A set of experiments were carried out to obtain the HAR dataset.

* A group of 30 volunteers with ages ranging from 19 to 48 years is selected for this task.
* Each person was instructed to follow a protocol of activities (WALKING, WALKINGUPSTAIRS, WALKINGDOWNSTAIRS, SITTING, STANDING, LAYING) while wearing a waist-mounted Samsung Galaxy S II smartphone.
* We will capture triaxial linear acceleration and angular velocity signals using the phone accelerometer and gyroscope at a sampling rate of 50Hz. The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed width sliding windows of 2.56 sec and 50% overlap.
* The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.
* The acceleration signal, which has gravitational and body motion components, was separated using another Butterworth low-pass filter into body acceleration and gravity.
* The gravitational force is assumed to have only low frequency components, therefore we found from the experiments that 0.3 Hz was an optimal corner frequency for a constant gravity signal.

**Softwares that we will use:**

Our team selected Python and Microsoft Excel as the software to analyze this dataset as they are widely used for data analysis and very simple to use. Python is open source, and well-supported. Python is more user-friendly for the approaches we'll employ in this project. There are many distributions of Python language and the most popular environment is Jupyter notebook which we will use for further part of this project. Excel lets you modify fields and functions that perform computations for you when dealing with more complicated data. Segmented data can be examined more thoroughly and visualized without the aid of additional tools, even for larger data sets.

**Analysis and Modeling techniques:**

Through the process of modeling, we will be training a machine learning algorithm to predict the labels from the features, tuning it for the business need and validating it on holdout data. The following 5 methods will be used to analyze this dataset:

* Logistic Regression: It will mainly be used by us for classification and predictive analysis. Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables.
* KNN: It is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. We aim to us KNN typically as a classification algorithm (On the assumption that similar points can be found near to one another).
* Naïve Bayes: It is a collection of classification algorithms based on Bayes’ theorem. It is not a single individual algorithm, but a family of algorithms where all of them share a common principle.
* SVM: SVM is a relatively simple Supervised Machine Learning Algorithm used for classification or regression. SVM is more preferred for classification but is sometimes very useful for regression.
* Decision Tree: Decision Tree is the most powerful and popular tool for classification and prediction. It has a flow chart structure where each internal node denotes an attribute.

**Error metrics that we plan to use & Algorithms for assessing them:**

1. ROC Curve: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classifier.
2. AUC: The Area Under the Curve (AUC) measures the performance of a classifier over all possible thresholds.
3. Accuracy: It measures the proportion of correct predictions made by the model.

To assess the above error metrics, we decided to use Logistic Regression.

**Descriptive Statistics:**

|  |  |
| --- | --- |
| *Activity* | |
|  |  |
| Mean | 3.315152 |
| Standard Error | 0.019717 |
| Median | 3 |
| Mode | 1 |
| Standard Deviation | 1.690628 |
| Sample Variance | 2.858222 |
| Kurtosis | -1.19717 |
| Skewness | 0.153416 |
| Range | 5 |
| Minimum | 1 |
| Maximum | 6 |
| Sum | 24373 |
| Count | 7352 |

From the above table we can see the descriptive statistics for the dataset that we have taken, the Mean value of Activity is 3.31, while the Median is 3 and the Mode resulted in 1. This is made through converting non-numeric column in numeric one. Further, Kurtosis is –1.91717, which means the distribution is flat, a negative kurtosis value suggests a relatively flat distribution, with fewer outliers or extreme values compared to a normal distribution. The Skewness is 0.153416 which indicates it is an almost substantially skewed distribution. This indicates that the distribution is slightly skewed to the right (positive skewness) or to the left (negative skewness). A positive skewness value means that the long tail of the distribution is on the positive side of the mean.

**Limitations and Concerns:**

Data Collection: Reliable and accurate data collection can be a challenge, as it requires proper equipment and techniques to capture human activity. Inaccurate or incomplete data can negatively impact the performance of the model.

Data Variability: Human activity can vary greatly from person to person and can be affected by factors such as posture, limb length, and body mass. This can result in variability in the data and make it difficult for the model to generalize to new examples.

**References:**

Khelalef, A.; Ababsa, F.; Benoudjit, N. An efficient human activity recognition technique based on deep learning. *Pattern Recognit. Image Anal.* **2019**, 29, 702–715. [[CrossRef](https://link.springer.com/article/10.1134/S1054661819040084)]

De-La-Hoz-Franco, E.; Ariza-Colpas, P.; Quero, J.M.; Espinilla, M. Sensor-based datasets for human activity recognition—A systematic review of literature. *IEEE Access* **2018**, 6, 59192–59210. [[CrossRef](https://ieeexplore.ieee.org/document/8478653)]

Ronao, C.A.; Cho, S.B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **2016**, 59, 235–244. [[CrossRef](https://www.sciencedirect.com/science/article/abs/pii/S0957417416302056?via%3Dihub)]

Wan, S.; Qi, L.; Xu, X.; Tong, C.; Gu, Z. Deep learning models for real-time human activity recognition with smartphones. *Mob. Netw. Appl.* **2020**, 25, 743–755. [[CrossRef](https://link.springer.com/article/10.1007/s11036-019-01445-x)]

Sousa Lima,W.; Souto, E.; El-Khatib, K.; Jalali, R.; Gama, J. Human activity recognition using inertial sensors in a smartphone: An overview. *Sensors* **2019**, 19, 3213. [[CrossRef](https://www.mdpi.com/1424-8220/19/14/3213)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31330919/)]