ELECTRICAL AND COMPUTER ENGINEERING
NORTH SOUTH UNIVERSITY

CSE445 PROJECT REPORT

Human Activity Classification Using Machine Learning Algorithms.

Group Members

Name	ID	Email Address
Tahmid Hasan Pranto	1711697042	tahmid.pranto@northsouth.edu
Azizul Hakim Tareq	1711080042	azizul.tareq@northsouth.edu
Golam Kibria Anik	1712345642	golam.anik@northsouth.edu
Mohammed Ariful Karim	1711153042	mohammad.karim01@northsouth.edu

SUBMITTED TO : SYED ATHAR BIN AMIR – SAA3

CONTENTS

Abstract—

Human activity recognition can be proven beneficial to several sectors including human-machine interaction, building smart devices with exclusive features and understanding human nature. Smartphones are one of the smart gadgets that stays close to a person for the better part of a day. Smartphones have several sensors which are capable of reading angular motion, acceleration, geographic location and tri-axial (3-d) motion data very accurately. Machine learning techniques have already proven their capability to work with data in large scale and having large number of feature. In this project we used a sensor read value dataset to classify human activity based on the data read by their cellphone sensors. The activity has been classified for 6 classes and a total of 3 machine learning models have been tested on this dataset. SVC performes with a highest testing accuracy of 95% followed by logistic regression with testing accuracy of 82% followed by Ridge classifier CV with testing accuracy of 81%. For further investigation about the models performance, we showed the precision, recall, f1-score along with corresponding confusion matrix.

Index Terms—**Human Activity Classification, Linear Regression, Support Vector Classification, Ridge Classifier Cross-Validation.**

I. Introduction

Smartphones have become an inseparable companion of human being. From entertainment to very sensitive and sophisticated tasks are often done by smartphones now-a-days. According to famous statistical data analysis company Statista, [1] the number of smartphone user in 2020 is 3.5 billion which is expected to reach 3.8 billion towards the end of 2021. As the phone user base is growing exponentially, marketing strategies are being designed to make the users use their phone as long as possible with exciting features and applications. Research done by RescueTime shows that smartphone users spends 3 hours and 15 minutes a day on average while the top 20% among this user-base uses their phone for as long as 4 and a half hour per day. [2] Another facts is the picking count as we on average pick our cellphone for 58 times a day. [2] So, without any doubt, we spend most of our time with our cellphone among all the electronic devices we use.

Human activity identification is a task that can aid several other possibilities like human robot collaboration [3], human smart-home interaction [4] and so on. As we stay close to our smartphone for a large portion of our day, data read by smartphone can be used to determine the current activity of a human. In this project we intend to classify or correctly identify human activity based on vast data read by sensors in regular smartphones. The rest of this report will start by describing our data then going through the process and methods we used finally conclude by analyzing our results. The sections are organized as follows: bla, bla and bla.

II. Related Works

Machine learning techniques and algorithms are ever evolving from the inception point where Arthur Samuel first came up with the Artificially Intelligent solution of the checker game in 1959. [5] Since then, hardware became cheaper and faster while so many new and advanced algorithm has been invented by brilliant minds of so many enthusiasts. Among them, classification and regression can be stated as the most important and most used algorithms. [6] Regression outputs the probability percentage of some event happening while classification outputs a single indicator of a unit belonging to a category. [6] Although, traditional classification and regression seems like a very basic approach towards solving complex machine learning problem, advanced algorithms are basically problem specific fine tuned versions of these techniques.

Several classification algorithms have already proven their capability in terms of multi-class classification. Algorithms like Logistic Regression, Support Vector Machines, Tree-based classifiers, Stochastic Gradient Descent, ADABoos, XGBoost, Neighbour-based classifiers have already proven their capability in terms of accuracy as well as performance. Classification algorithms have been used in verity of fields and sectors. These algorithms have been used in the medical sector for classifying breast cancer [7], brain histology [8] and facies [9]. Remotely sensed geospatial data is another sector where machine learning algorithms are frequently applied for tasks like land use classification [10], land cover mapping. [11] and other use cases. [12] ML has also been used in entertainment sector for music genre classification [13], movie rating prediction [14] and as well as movie success prediction [15].

III. Dataset

The dataset we're using has been taken from Kaggle [16] which serves a vast amount of open-source data. The name of the dataset is, "Human Activity Recognition with Smartphones", which has been collected by a company named UCI Machine Learning. They collected 3-axial linear acceleration and angular velocity from the built-in accelerometer and gyroscope sensors at a constant 45Hz rate from the smartphones of 30 persons who participated in this study. These study people are aged between 19-48 years. Based on these input features, this dataset have 6 different target labels named as activities (WALKING, WALKINGUPSTAIRS, WALKINGDOWNSTAIRS, SITTING, STANDING, LAYING). 70% of the study population was used for train set generation while the rest 30% was used to generate test dataset.

The training set contains a total of 7352 number of instances and the test set contains 2947 number of instances. Each instances in the dataset contains 562 features.

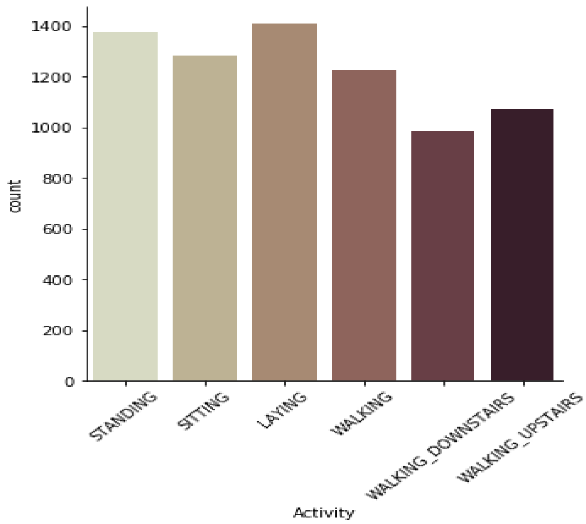

Figure 1. Number of instances per class.

From figure 1, we can see that the dataset is well balanced, it's not skewed towards some particular class.

IV. Methodology

A. Pre-Processing

This dataset have been pre-processed for noise reduction. These noises would appear due to the effect of gravity on the sensor or due to overlapping of signals. 0.3Hz frequency was cut to eliminate the expected noise due to gravity and several noise filters have been applied to reduce the noise due to signal overlapping. There was no null value or missing data. But, the target or label column which is "Activity" wasn't numeric. As machine learning model doesn't understand anything but numbers, we had to encode it to numeric value. A greater part of the data is consisting of sensor read negative numbers. Feature normalization has already been proved to aid the performance of machine learning classification models. [17] There's also range difference among the data which might also decrease the performance. So, we applied feature scaling and normalization with the help of Pipeline.

B. Pipeline

As we intended to compare the performance of different machine learning algorithms for classifying human activity, to deal with the requirements of different algorithms becomes difficult. For better performance, some model wants the features to be normalized where the term "feature scaling" means to compress the data within a fixed range for all the features.To implement this variety of requirements and data transformation, a better approach is to use Pipeline. Pipeline can handle these varieties within itself while not transforming the original data over and over again. Machine learning pipeline follows the flow-diagram below. (figure 2)

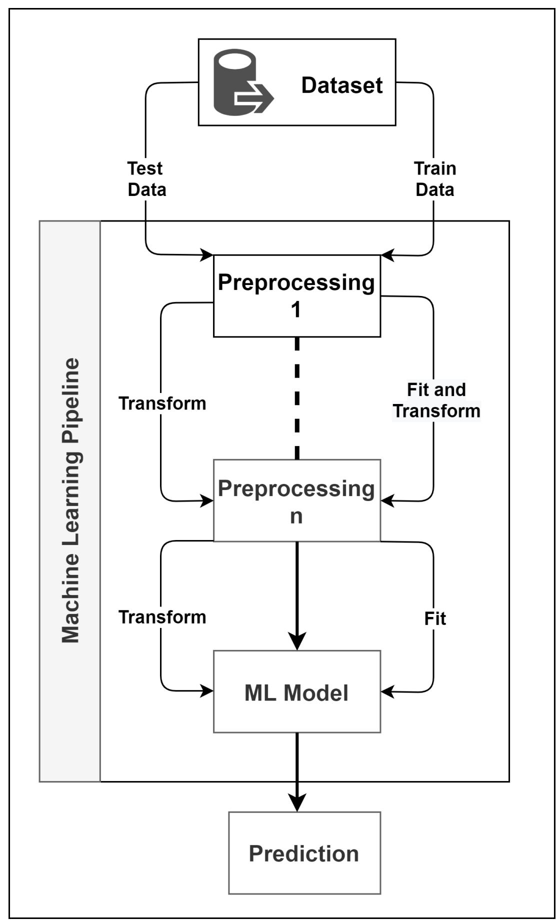

Figure 2. Flow diagram of a machine learning pipeline.

C. Model Selection

D. Logistic Regression

In our case, it's a multi class classification which is also known as softmax regression. In this process first we compute the following vector S which holds the predicted score of each class.

$$S_k(x) = X^T\theta_k \tag{1}$$

we use this above formula in the softmax function to normalize the predicted output so that for each class, the results are normalized to a particular range. The sofmax function is given below.

$$\sigma(S_k(x)) = \frac{e^{(S_k(x))}}{\sum_{j=1}^{k} e^{(S_j(x))}} \tag{2}$$

E. Support Vector Classification

By default, support vector classification is linear. But, with kernel trick, it can perform multi-class classification. As we have many features and we can't linearly decide the decision boundary, we used kernel trick. For soft margin classification that we're using, the algorithm tries to minimize the following equation.

$$W^T x + b \tag{3}$$

while maintaining a constraint as follows.

$$t^i(W^T x + b) \tag{4}$$

By drawing hyper-plane to separate the data, SVC performs the classification while with kernel trick, data is projected to higher dimensional space for a more accurate decision boundary.

F. Ridge Classification CV

Ridge Classification works in the similar fasion of Logistic regression except for the fact that it uses "l2 norm" for regularization. But, the advantage of sckikit-learn ridge classification is it comes with built in k-fold cross validation advantage.

V. Result & Analysis

A machine learning classifier can be made in several ways. We can aim for classifiers which are very strict about mistakenly classifying positive or negative classes and we can also give the model independence about these positive or negative error rate and focus on the models ability and accuracy. Precision and recall are the machine learning measurements that deals with this positive and negative error rates. For specific use cases we can trade between precision and recall. But, in classifier in general, we use F1 score to evaluate a model. All these classification scores can be decided from confusion matrix's which is a representation of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) values.

A. Analysis methods

1) Precision: Precision score defines what portion of positive classes are classified correctly. High precision means low False positive classifications by the model. Equation 5 is used to calculate precision.

$$\frac{TP}{TP + FP} \tag{5}$$

2) Recall: Recall score point out to the portion of actual positive classes were classified correctly. Recall is the measurement to look for while making a very strict classifier which doesn't mistakenly misclassifies a positive class. Equation 6 is used to calculate recall.

$$\frac{TP}{TP + FN} \tag{6}$$

3) F1 Score: For a general classifier, F1 score is important because it is a combination of both precision and recall which is a measurement of overall accuracy of the model. Equation 7 is used to measure F1-Score of the classification model.

$$2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

4) Confusion Matrix: Only train test accuracy doesn't tell us how good a classification model is. Also if we want to see if some class is being misclassified into some other classes, and we want to visually recongnize those conflicting class, confusion matrix enables us to do that. Confusion matrix is like a pair plot, where rightly or wrongly classified labels are shown as per their count in a matrix representation. From this type of matrix, we get more insight that enlightens us with deeper understanding of our models performance.

B. Result

The table below (table I) shows the training and testing accuracy of the individual models.

Model	Training Accuracy	Testing Accuracy
SVC	1.0	0.95
Logistic Regression	0.86	0.82
Ridge CV	0.84	0.81

Table I
TRAIN AND TEST ACCURACY SCORES OF THE MODELS.

The confusion matrix's generated based on the output of each model is attached below.

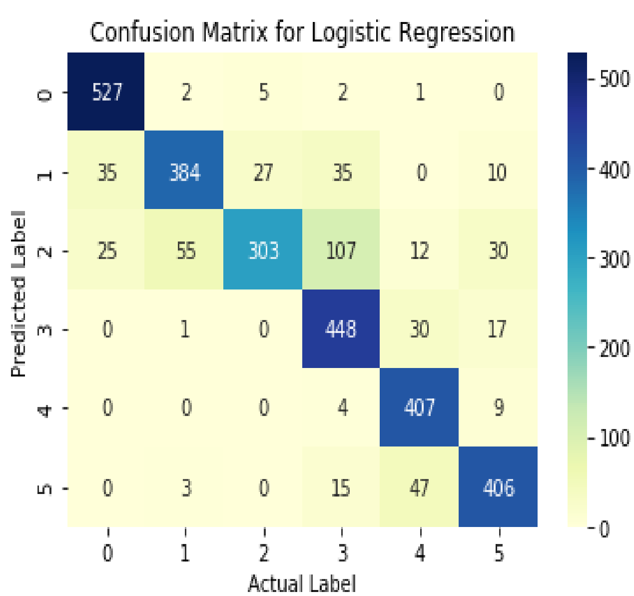

Figure 3. Logistic Regression Confusion Matrix.

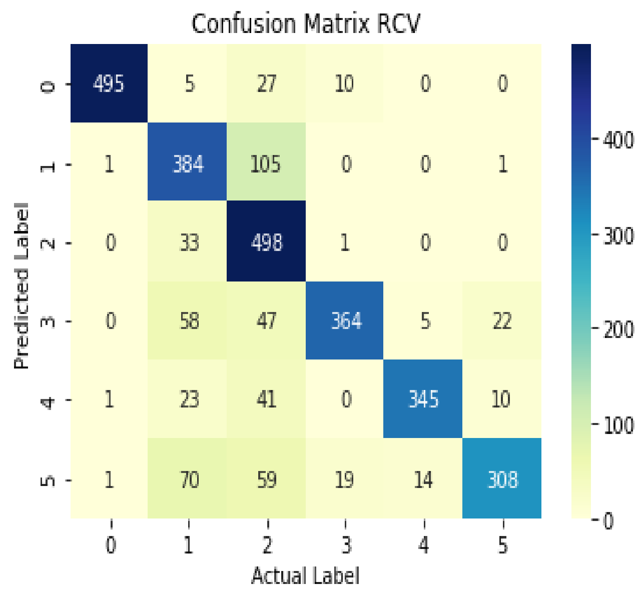

Figure 4. RCV Confusion Matrix.

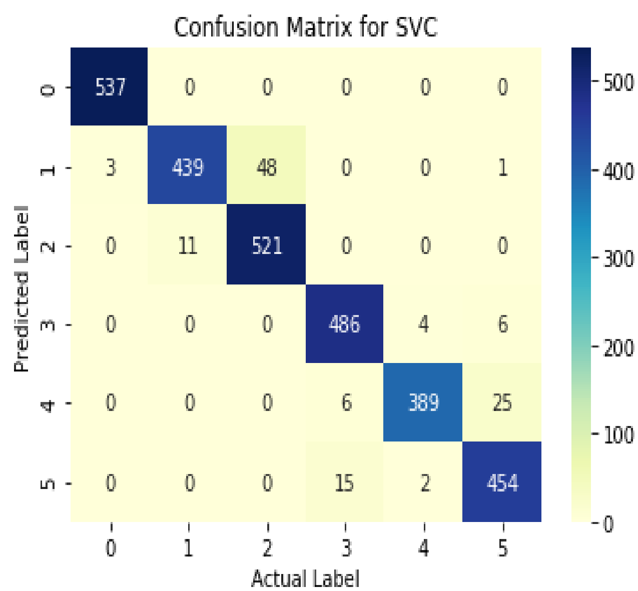

Figure 5. SVC Confusion Matrix.

From table I, we can see that SVC, Logistic Regression, Ridge Classification CV has performs well on this dataset. SVC performed the best with 100% training and 95% testing accuracy. Logistic regression and Ridge classifier also scored more than 80% which is also a notable performance on this dataset having many features. The confusion matrix's for the models which are SVC, Logistic Regression and Ridge CV is shown below on figure 5, 3 and 4.

From confusion matrix, using the formula for precision, recall and f1-score we can get more insight about the performances of these models. The table below shows these scores for each indevidual models.

Model	Precision	Recall	F1-Score
SVC	0.96	0.96	0.96
Logistic Regression	0.85	0.84	0.84
Ridge CV	0.85	0.81	0.82

Table II
PRECISION, RECALL AND F1-SCORES OF THE MODELS.

The precision, recall and f1-scores again shows that SVC performing best while the other two models performing well for classifying human activity.

VI. CONCLUSION

Human and machinery has a very close bonding now-a-day as we are being dependent on the machines now more than ever before. The fundamental aspect of machines is, they make our life easier and comfortable. As time goes, machines are becoming smart enough to decide things without human supervision. But, for the machines to be smart enough to understand human behavior, activity classification based on sensor read data can be considered as a preliminary task. In this project we tried to focus on the performance of different machine learning models to classify human activity into 6 different classes. Codes of this project can be found on GitHub on this URL : https://github.com/Tahmid1406/Human-Activity-Classification-Using-ML-Techniques. The implementation is generic enough to use it in similar scenario.

REFERENCES

[1] Statista, "Number of smartphone users worldwide from 2016 to 2021," https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/, accessed: 2020-12-25.

[2] A. Matei, "Shock! horror! do you know how much time you spend on your phone?" https://www.theguardian.com/lifeandstyle/2019/aug/21/cellphone-screen-time-average-habits, accessed: 2020-12-25.

[3] A. M. Zanchettin, A. Casalino, L. Piroddi, and P. Rocco, "Prediction of human activity patterns for human–robot collaborative assembly tasks," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3934–3942, 2018.

[4] Y. Du, Y. Lim, and Y. Tan, "A novel human activity recognition and prediction in smart home based on interaction," *Sensors*, vol. 19, no. 20, p. 4474, 2019.

[5] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[6] I. Goodfellow, Y. Bengio, and A. Courville, "Machine learning basics," *Deep learning*, vol. 1, pp. 98–164, 2016.

[7] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensarİ, "Breast cancer classification using machine learning," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*. IEEE, 2018, pp. 1–4.

[8] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC medical informatics and decision making*, vol. 17, no. 1, pp. 1–13, 2017.

[9] B. Hall, "Facies classification using machine learning," *The Leading Edge*, vol. 35, no. 10, pp. 906–909, 2016.

[10] Y. Qian, W. Zhou, J. Yan, W. Li, and L. Han, "Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery," *Remote Sensing*, vol. 7, no. 1, pp. 153–168, 2015.

[11] N. N. Kussul, N. S. Lavreniuk, A. Y. Shelestov, B. Y. Yailymov, and I. N. Butko, "Land cover changes analysis based on deep machine learning technique," *Journal of Automation and Information Sciences*, vol. 48, no. 5, 2016.

[12] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2784–2817, 2018.

[13] H. Bahuleyan, "Music genre classification using machine learning techniques," *arXiv preprint arXiv:1804.01149*, 2018.

[14] M. Marović, M. Mihoković, M. Mikša, S. Pribil, and A. Tus, "Automatic movie ratings prediction using machine learning," in *2011 Proceedings of the 34th International Convention MIPRO*. IEEE, 2011, pp. 1640–1645.

[15] K. Lee, J. Park, I. Kim, and Y. Choi, "Predicting movie success with machine learning techniques: Ways to improve accuracy," *Information Systems Frontiers*, vol. 20, no. 3, pp. 577–588, 2018.

[16] U. M. Learning, "Human activity recognition with smartphones," https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones, accessed: 2020-12-24.

[17] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020.

Table of Contribution

Task	Contributor
Coding	All - Via multiple online meetings and screen-sharing
Introduction	Azizul Hakim Tarek
Related Works	All via paper distribution among the group
Dataset Section	Golam Kibria Anik
Preprocessing	Mohammed Ariful Karim
Model Selection	Tahmid Hasan Pranto
Result & Analysis	Tahmid Hasan Pranto
Formatting	Tahmid Hasan Pranto

Table III
CONTRIBUTION TABLE.