4 - HUMAN ACTIVITY RECOGNITION

A Project Report

Submitted by (Team 30)

KESAVARAJ - 2022701008

POORVAJA - 2020101096

VIVEK - 2022701001

ROHAN NAIN - 2020102023

In the partial fulfillment for the completion of the project

Of

Statistical Methods in Artificial Intelligence

Table of Contents:

1. Introduction:

By continuously monitoring and analyzing user activity it is possible to provide automated recommendations. Common consumer devices such as smart phones and smart watches generally ship with IMUs (Inertial Measurement Unit), which are packaged accelerometer and gyroscope sensors. Through the information provided by these IMUs, machine learning techniques can then be used to train activity classifiers, giving users, doctors, and app developers access to an individual's lifestyle and activity choices. Various machine learning techniques can be used to train models for classifying human activities. Several models are tested in the paper like decision trees, random forest, SVM, Ridge and Linear regression. Also, we did the K fold validation and considered the best accuracy of each model. Among these techniques we examine what is the "best" classification technique with a complete set of features and a subset of features. We used the PAMAP2 Dataset from the UCI repository of machine learning datasets.

First we pre-processes the data in which we combined the data of all the 9 subjects into a single data file. Data consists of 54 attributes, like Timestamp, Activity Id, Heart rate and 17 recordings from IMU sensors worn on hand, chest and ankle. There are 24 activities and some of the activities are ironing, walking, lying, standing, sitting, Nordic walking, vacuum cleaning. Recordings from IMU sensors include temperature, 3-axis acceleration, 3-axis angular velocity, and 3-axis magnetometer, etc. We considered only 12 activities and later we compared the subset of 18 attributes, Heart rate and IMU recordings from one location. Heart rate attribute has missing values and those values were filled using linear interpolation. We considered a different number of tuples to train each classifier using random sampling considering huge data. We used 4.25 lakh samples to train SVM, Multilayer perceptron and Adaboost. Whereas for other models we considered complete data. As the number of training samples are not the same, we introduced a new measure speed to compare the time taken to train the model. Speed can be defined as the number of training tuples divided by time taken to train and units is lakh samples per minute. Machine learning techniques were compared based on accuracy and speed.

2. Machine Learning

2.1 Decision Tree(Max depth=15):

Decision Tree is a Supervised learning technique that can be used for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Here, we used Gini for the measurement of uncertainty. An attribute with lowest Gini impurity is used as the split node for making decisions. The max depth is 15, which defines the maximum depth of the tree.We have run a 5-fold cross validation and for each validation the training score, validation score, and test score, time taken is displayed in the below table.

Fold No	**Training Score**	**Validation Score**	**Test Score**	**Time Taken(in mins)**
1	0.9856	0.9850	0.9847	4.14
2	0.9849	0.9838	0.9839	2.73
3	0.9848	0.9838	0.9837	2.54
4	0.9857	0.9847	0.9846	2.95
5	0.9866	0.9856	0.9854	2.45

Table: 2.1.1

Of the 5-fold validations, the best test score accuracy is for validation-5 and the value is 98.54%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28760	2	10	0	0	0	1	0	1	27	11	0
sitting	5	27606	154	0	0	0	0	0	0	22	2	0
standing	0	16	28245	7	0	0	0	0	1	71	125	0
walking	0	0	1	35422	0	12	12	141	11	15	1	0
running	0	0	0	40	14680	4	1	55	0	3	0	56
cycling	0	2	0	100	11	24498	96	17	0	86	2	7
Nordic walking	0	0	0	260	2	57	27936	29	0	5	1	1
ascending stairs	0	0	5	334	27	16	3	16627	424	69	8	0
descending stairs	0	0	0	336	81	0	0	318	14911	53	13	0
vacuum cleaning	0	8	74	205	0	2	0	18	29	25701	297	0
ironing	1	5	253	0	0	10	0	3	3	92	35387	0
rope jumping	0	0	0	0	68	10	0	0	0	0	0	7410

Fig: 2.1.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ascending stairs is misclassified as descending stairs (424)
2. descending stairs is misclassified as walking (336)
3. ascending stairs is misclassified as walking (334)
4. nordic walking is misclassified as walking (260)
5. ironing is misclassified as standing (253).

On executing the same code on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9802	0.9785	0.9784	1.09
2	0.9796	0.9774	0.9775	1.18
3	0.9779	0.9761	0.9760	1.13
4	0.9797	0.9771	0.9778	1.2
5	0.9797	0.9775	0.9776	1.06

Table: 2.1.2

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28583	13	43	4	0	0	0	110	12	26	21	0
sitting	120	27138	443	0	0	1	1	0	1	63	22	0
standing	19	137	27892	29	0	5	0	23	4	129	227	0
walking	1	9	47	35103	2	47	123	62	25	189	7	0
running	0	2	0	39	14544	19	6	14	26	0	2	187
cycling	3	15	0	159	15	24315	196	62	46	7	0	1
Nordic walking	10	3	0	140	1	98	28017	14	5	2	0	1
ascending stairs	0	0	73	238	3	12	6	16598	266	313	3	1
descending stairs	1	17	44	38	10	50	5	254	15031	188	74	0
vacuum cleaning	74	106	202	155	0	7	3	68	13	25475	231	0
ironing	23	73	352	2	0	0	0	16	9	185	35094	0
rope jumping	0	0	0	0	100	26	8	0	0	0	0	7354

Fig: 2.1.2

Best Accuracy on split data is 97.8% From the confusion matrix it can be inferred that the more number of misclassified labels is seen for below attributes

1. ironing is misclassified as standing (352)
2. Ascending stairs is misclassified as vacuum cleaning (313)
3. Ascending stairs is misclassified as descending stairs (266)

. There is a slight decrease in the accuracy of the decision tree which is trained on the subset of the data and the misclassified activities ironing vs standing and ascending stairs vs descending stairs are common. Time taken to train the model with subset data is less compared to complete data.

2.2 Adaboost:

Adaboost is one of the ensemble methods where we train multiple models and assign class labels to a sample based on the vote from each classifier. AdaBoost algorithm is a boosting technique where weights are updated with higher weights to wrongly classified instances. While building this classifier, the base classifier is a decision tree. Here the max depth of the base decision tree is 9, and the number of estimators used are 100. Gini impurity criteria is used in the base classifier. We have run a 5-fold cross validation and for each validation the training score, validation score, and test score, time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9983	0.9978	0.9979	94.98
2	0.9987	0.9979	0.9981	63.15
3	0.9979	0.9974	0.9975	79.84
4	0.9988	0.9984	0.9985	126.33
5	0.9980	0.9974	0.9971	111.55

Table: 2.2.1

Of the 5-fold validations, the best test score accuracy is for validation-4 and the value is 99.85%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28809	0	0	0	0	0	0	0	0	3	0	0
sitting	1	27769	19	0	0	0	0	0	0	0	0	0
standing	0	10	28350	0	0	0	0	0	0	0	105	0
walking	0	0	0	35615	0	0	0	0	0	0	0	0
running	0	0	0	0	14839	0	0	0	0	0	0	0
cycling	0	0	0	2	0	24817	0	0	0	0	0	0
Nordic walking	0	0	0	1	0	0	28290	0	0	0	0	0
ascending stairs	0	0	0	4	0	0	0	17385	124	0	0	0
descending stairs	0	0	0	0	0	0	0	153	15559	0	0	0
vacuum cleaning	0	0	0	0	0	0	0	0	0	26334	0	0
ironing	0	0	15	0	0	0	0	0	0	0	35739	0
rope jumping	0	0	0	0	1	0	0	0	0	0	0	7487

Fig: 2.2.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ascending stairs is misclassified as descending stairs (124)
2. standing is misclassified as ironing (105)

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9848	0.9812	0.9812	29.52
2	0.9828	0.9789	0.9784	42.32
3	0.9825	0.9781	0.9777	42.37
4	0.9840	0.9793	0.9793	30.90
5	0.9841	0.9807	0.9795	15.67

Table: 2.2.2

Of the 5-fold validations, the best test score accuracy is for validation-5 and the value is 98.12%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28775	8	1	4	0	0	0	1	0	9	14	0
sitting	3	27687	78	0	0	0	0	0	5	11	5	0
standing	0	13	28037	0	0	0	0	0	0	138	277	0
walking	0	0	0	35061	0	2	129	206	7	210	0	0
running	0	0	0	18	14724	1	3	12	36	2	0	43
cycling	0	0	0	12	1	24718	31	26	22	9	0	0
Nordic walking	0	0	0	109	1	52	28122	0	0	7	0	0
ascending stairs	0	0	0	130	0	0	0	16193	884	306	0	0
descending stairs	0	0	1	21	0	16	0	1219	14212	187	56	0
vacuum cleaning	0	1	62	28	0	2	0	279	83	25477	402	0
ironing	0	2	46	0	0	0	0	6	21	416	35263	0
rope jumping	0	0	0	1	33	1	0	0	1	0	0	7452

Fig: 2.2.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ascending stairs is misclassified as descending stairs (884)
2. vacuum cleaning is misclassified as ironing (402)
3. ascending stairs is misclassified as vacuum cleaning (306)

There is a drop in accuracy for ensemble trained on the subset of attributes. Both the ensembles misclassified standing as ironing. There is a significant drop in training time with subset data.

2.3 Support Vector Machine

Support vector machine or SVM, creates a hyperplane to split the classes. With the help of support vectors we try to maximize the margin, the gap between classifier and support vectors. Here to separate the data we have used rbf kernels which transform the input data into higher dimensions to convert non linear separable classes to linearly separable classes. Here we have taken C=10, C is the regularization parameter. The regularization parameter is a degree of importance that is given to misclassifications.

We have run a 5-fold cross validation and for each validation the training score, validation score, and test score, time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9985	0.9979	0.9977	137.7
2	0.9985	0.9977	0.9976	110.9
3	0.9986	0.9979	0.9978	123.3
4	0.9984	0.9978	0.9974	123.7
5	0.9985	0.9978	0.9978	119.5

Table: 2.3.1

Of the 5-fold validations, the best test score accuracy is for validation-5 and the value is 99.78%. The confusion matrix for the best classifier is displayed below

Confusion Matrix

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28811	1	0	0	0	0	0	0	0	0	0	0
sitting	0	27758	30	1	0	0	0	0	0	0	0	0
standing	0	11	28449	0	0	0	0	0	0	0	5	0
walking	0	0	0	35598	0	0	9	2	0	0	0	6
running	0	0	0	0	14834	0	0	0	3	0	0	2
cycling	0	0	0	0	7	24798	0	1	3	0	0	10
Nordic walking	0	0	0	5	10	0	28275	0	0	0	0	1
ascending stairs	0	0	0	4	21	0	2	17382	98	0	0	6
descending stairs	0	0	0	7	6	0	4	37	15649	0	0	9
vacuum cleaning	0	0	0	0	3	0	0	8	3	26319	0	1
ironing	0	0	300	0	4	0	1	0	0	1	35448	0
rope jumping	0	0	0	0	5	0	0	0	0	0	0	7483

Fig: 2.3.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ironing is misclassified as standing (300).
2. ascending stairs is misclassified as descending stairs (98)
3. descending stairs is misclassified as ascending stairs (37)
4. sitting is misclassified as standing (30)
5. ascending stairs is misclassified as running (21)

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9992	0.8571	0.8566	77.07
2	0.9987	0.8576	0.9981	47.17
3	0.9994	0.8575	0.9988	32.39
4	0.9985	0.9979	0.9978	25.48
5	0.9990	0.9985	0.9985	45.85

Table: 2.3.2

Of the 5-fold validations, the best test score accuracy is for validation-5 and the value is 99.85%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28807	5	0	0	0	0	0	0	0	0	0	0
sitting	0	27762	24	0	0	0	0	0	2	1	0	0
standing	0	26	28410	0	0	0	0	0	0	0	29	0
walking	0	0	0	35604	1	0	5	3	2	0	0	0
running	0	0	0	0	14836	0	0	1	1	0	0	1
cycling	0	0	0	0	2	24812	1	3	0	1	0	0
Nordic walking	0	0	0	6	1	0	28284	0	0	0	0	0
ascending stairs	0	0	0	2	0	0	0	17431	72	8	0	0
descending stairs	0	0	0	1	2	1	1	56	15648	3	0	0
vacuum cleaning	3	2	2	0	0	0	0	3	7	26315	2	0
ironing	0	0	20	0	0	0	0	0	0	0	35734	0
rope jumping	0	0	0	0	2	0	0	1	0	0	0	7485

Fig: 2.3.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ascending stairs is misclassified as descending stairs (72)
2. Descending stairs is misclassified as ascending stairs (56)
3. standing is misclassified as ironing (29)

The accuracy of model trained with the subset of data has increased compared to model trained with complete data. Time taken to train is also comparatively less.

2.4 Ridge Regression

Linear regression is a supervised classification technique used to classify the sample based on its features. To make the Linear regression model more robust against overfitting, L2 regularization was incorporated with a penalty to alter the strength of regularization in order to overcome overfitting. Alpha(Regularization strength) is taken as 0.01, regularization improves the conditioning of the problem and reduces the variance of the estimates. The Stochastic Average Gradient (SAG) was used as it generally provides fast convergence for large feature-set. Stochastic gradient descent (often abbreviated SGD) is an iterative method for optimizing an objective function with suitable smoothness properties.We have run a 5-fold cross validation and for each validation the training score, validation score, and test score, time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.7171	0.7165	0.7156	8.13
2	0.7168	0.7172	0.7156	7.65
3	0.7173	0.7176	0.7160	7.88
4	0.7174	0.7170	0.7161	8.03
5	0.7173	0.7171	0.7159	7.67

Table: 2.4.1

Of the 5-fold validations, the best test score accuracy is for validation-4 and the value is 71.61%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	26941	350	817	110	0	37	8	12	0	220	317	0
sitting	333	22728	619	41	3	96	0	0	2	389	3578	0
standing	43	1721	21026	117	5	0	0	1	0	253	5299	0
walking	13	52	394	32956	1	736	732	70	16	273	372	0
running	49	302	303	607	9665	518	1531	472	514	196	479	203
cycling	2	68	3	234	25	23013	185	151	71	143	924	0
Nordic walking	175	816	540	10134	1220	2376	10513	576	184	390	1292	75
ascending stairs	69	313	2018	3460	770	386	568	5783	652	2669	824	1
descending stairs	117	527	882	4852	533	667	839	995	3432	1032	1836	0
vacuum cleaning	41	1579	1239	1070	65	659	26	177	73	17992	3413	0
ironing	0	404	1385	13	0	50	0	0	0	216	33686	0
rope jumping	7	98	114	124	4029	664	1085	248	68	99	2	950

Fig: 2.4.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. standing is misclassified as ironing (5299)
2. descending stairs is misclassified as walking (4852)
3. Rope jumping is misclassified as running (4029)
4. Ascending stair is misclassified as walking (3460)

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.5117	0.5113	0.5106	2.72
2	0.5120	0.5111	0.5107	2.50
3	0.5121	0.5124	0.5112	2.48
4	0.5118	0.5127	0.5110	2.46
5	0.5108	0.5108	0.5095	2.46

Table: 2.4.2

Of the 5-fold validations, the best test score accuracy is for validation-2 and the value is 51.112%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	26714	332	499	515	0	198	14	0	0	313	227	0
sitting	13935	8480	1395	226	0	2	0	0	0	124	3627	0
standing	2013	1261	17932	1719	0	30	1	0	0	770	4739	0
walking	451	368	4082	23645	58	468	600	17	0	2579	3347	0
running	408	89	80	1600	5626	1909	2855	642	659	77	893	1
cycling	162	47	0	1015	143	19695	126	93	37	463	3038	0
Nordic walking	3403	598	177	5046	1838	4875	5952	18	21	2338	4025	0
ascending stairs	179	242	2434	4059	1241	719	590	3112	209	3028	1700	0
descending stairs	612	224	1257	2590	880	3413	461	1453	429	1903	2490	0
vacuum cleaning	1220	746	4371	3707	774	342	5	575	142	9026	5426	0
ironing	3250	906	2690	173	0	87	10	0	0	266	28372	0
rope jumping	195	7	8	456	2486	1402	2612	104	40	8	170	0

Fig: 2.4.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. sitting is misclassified as lying (13935)
2. Vacuum cleaning is misclassified as ironing (5426)
3. Nordic walking is misclassified as walking (5046)
4. Nordic walking is misclassified as cycling (4875)

Although the time taken to run each validation has decreased, the accuracy of the model when trained with split datasets has dropped drastically, it can also be seen that the number of misclassifications have increased by a lot. Misclassified samples are different in both the trained models.

2.5 Neural Network:

Neural network is modeled loosely after the human brain, which is designed to recognize patterns. Neural networks help us cluster and classify. The network consists of an input layer, hidden layer and output layer. The layers are made of nodes. A node is just a unit where computation happens. We considered the number of nodes in each hidden layer as 512. A node combines input from a neuron connected with a set of coefficients, or weights, that either amplify or dampen that input, thereby assigning significance to inputs with regard to the task the algorithm is trying to learn. These input-weight products are summed and then the sum is passed through a node's so-called activation function. Among the activation functions available, Relu (rectified linear unit) was used, which outputs the sum directly if the input is positive, otherwise it outputs zero. The number of epochs used is 150, which indicates the number of iterations on the entire training dataset. The training, validation, test scores and time taken for each validation is displayed below.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9992	0.8571	0.8566	77.07
2	0.9987	0.8576	0.9981	47.17
3	0.9994	0.8575	0.9988	32.39
4	0.9985	0.9979	0.9978	25.48
5	0.9990	0.9985	0.9985	45.85

Table: 2.5.1

Of the 5-fold validations, the best test score accuracy is for validation-5 and the value is 99.85%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28807	5	0	0	0	0	0	0	0	0	0	0
sitting	0	27762	24	0	0	0	0	0	2	1	0	0
standing	0	26	28410	0	0	0	0	0	0	0	29	0
walking	0	0	0	35604	1	0	5	3	2	0	0	0
running	0	0	0	0	14836	0	0	1	1	0	0	1
cycling	0	0	0	0	2	24812	1	3	0	1	0	0
Nordic walking	0	0	0	6	1	0	28284	0	0	0	0	0
ascending stairs	0	0	0	2	0	0	0	17431	72	8	0	0
descending stairs	0	0	0	1	2	1	1	56	15648	3	0	0
vacuum cleaning	3	2	2	0	0	0	0	3	7	26315	2	0
ironing	0	0	20	0	0	0	0	0	0	0	35734	0
rope jumping	0	0	0	0	2	0	0	1	0	0	0	7485

Fig: 2.5.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ascending stairs is misclassified as descending stairs (72)
2. Descending stairs is misclassified as ascending stairs (56)
3. standing is misclassified as ironing (29)

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9972	0.9946	0.9949	129.47
2	0.9985	0.9959	0.9959	89.03
3	0.9984	0.9962	0.9960	83.92
4	0.9987	0.9959	0.9963	79.73
5	0.9973	0.9950	0.9949	146.21

Table: 2.5.2

Of the 5-fold validations, the best test score accuracy is for validation-4 and the value is 99.63%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28790	9	3	0	0	0	0	1	1	2	6	0
sitting	0	27758	17	0	0	0	0	1	2	10	1	0
standing	0	9	28317	0	0	0	0	4	0	15	120	0
walking	0	0	0	35555	0	9	23	16	2	8	2	0
running	0	0	0	1	14819	0	1	4	8	3	0	3
cycling	0	0	0	3	1	24787	15	3	5	4	0	1
Nordic walking	0	0	1	17	10	6	28236	5	4	1	11	0
ascending stairs	0	1	0	14	7	1	0	17346	113	25	3	3
descending stairs	0	1	1	9	8	2	4	184	15460	28	14	1
vacuum cleaning	1	4	24	16	0	0	2	39	41	26150	57	0
ironing	0	1	32	5	1	0	0	4	3	46	35662	0
rope jumping	0	0	0	0	94	4	2	2	1	0	0	7385

Fig: 2.5.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. Descending stairs is misclassified as ascending stairs (184)
2. standing is misclassified as ironing (120)
3. ascending stairs is misclassified as descending stairs (113)
4. Rope jumping is misclassified as running (94)

It can be seen that even by taking a subset of the data the prediction accuracy and the amount of misclassifications is almost similar. Time taken to train the neural network with the subset is slightly lesser than training it with complete data. Both the classifiers misclassified Descending stairs is misclassified as ascending stairs, standing is misclassified as ironing, ascending stairs is misclassified as descending stairs

2.6 Random Forest:

Random forest is one of the ensemble learning methods for classification and regression. During training it constructs multiple trees and for classification, output of the random forest is the class selected by most trees and for regression the mean prediction of individual trees is returned. In our code we have taken 50 trees in the forest with a maximum depth of 20.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9995	0.9994	0.9993	17.41
2	0.9998	0.9996	0.9997	15.66
3	0.9994	0.9991	0.9992	16.54
4	0.9997	0.9995	0.9995	16.89
5	0.9996	0.9994	0.9994	17.86

Table: 2.6.1

Of the 5-fold validations, the best test score accuracy is for validation-1 and the value is 99.97%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28810	2	0	0	0	0	0	0	0	0	0	0
sitting	0	27783	5	0	0	0	0	0	0	1	0	0
standing	0	0	28455	0	0	0	0	0	0	0	10	0
walking	0	0	0	35612	0	0	3	0	0	0	0	0
running	0	0	0	0	14839	0	0	0	0	0	0	0
cycling	0	0	0	0	0	24819	0	0	0	0	0	0
Nordic walking	0	0	0	3	0	0	28288	0	0	0	0	0
ascending stairs	0	0	0	3	0	0	0	17504	4	2	0	0
descending stairs	0	0	0	2	0	0	0	7	15695	8	0	0
vacuum cleaning	0	0	0	0	0	0	0	1	0	26324	9	0
ironing	0	0	23	0	0	0	0	0	0	3	35728	0
rope jumping	0	0	0	0	0	0	0	0	0	0	0	7488

Fig: 2.6.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ironing is misclassified as standing (23)
2. standing is misclassified as ironing (10)
3. Vacuum cleaning is misclassified as ironing (9)

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9986	0.9974	0.9972	10.48
2	0.9986	0.9971	0.9971	10.44
3	0.9981	0.9968	0.9967	8.61
4	0.9983	0.9966	0.9967	8.35
5	0.9987	0.9974	0.9973	7.27

Table: 2.6.2

Of the 5-fold validations, the best test score accuracy is for validation-4 and the value is 99.73%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28797	1	0	7	0	0	0	0	0	5	2	0
sitting	0	27673	112	1	0	0	0	0	0	2	1	0
standing	0	0	28415	0	0	0	0	0	0	13	37	0
walking	0	0	2	35575	0	4	32	0	0	2	0	0
running	0	0	0	0	14828	0	1	5	3	0	0	2
cycling	0	0	0	15	0	24758	25	19	2	0	0	0
Nordic walking	0	0	0	14	1	3	28268	3	0	2	0	0
ascending stairs	0	0	0	98	0	0	0	17353	23	37	2	0
descending stairs	0	0	2	22	0	0	0	30	15601	53	4	0
vacuum cleaning	0	0	35	23	0	0	0	3	0	26230	43	0
ironing	0	0	60	0	0	1	0	0	0	7	35686	0
rope jumping	0	0	0	0	0	7	0	0	0	0	0	7481

Fig: 2.6.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. sitting is misclassified as standing (112)
2. ascending stairs is misclassified as walking (98)
3. ironing is misclassified as standing (60)
4. descending stairs is misclassified as vacuum cleaning (53).

Accuracy of both the models is same but time taken to train model with the subset is half the time taken to train model with complete data.

2.7 Miscellaneous:

2.7.1 Decision tree with entropy:

This model is a decision tree with different criteria. The base parameters used here are max depth with a value of 15. Entropy criteria is used in the classifier. In the decision tree, data is split based on values of the feature vector associated with each data point. The training, validation, test scores and time taken for each validation is shown in table below.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9954	0.9950	0.9948	10.17
2	0.9953	0.9948	0.9947	10.05
3	0.9952	0.9945	0.9946	9.99
4	0.9951	0.9945	0.9944	9.97
5	0.9954	0.9947	0.9947	8.65

Table: 2.7.1.1

Of the 5-fold validations, the best test score accuracy is for validation-0 and the value is 99.48%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28793	1	10	0	0	0	0	0	0	6	2	0
sitting	0	27576	165	0	0	0	0	0	6	32	10	0
standing	0	69	28333	0	0	0	0	1	0	6	56	0
walking	0	0	5	35582	0	15	2	3	0	7	1	0
running	0	0	1	0	14822	6	0	0	1	4	4	1
cycling	0	0	1	0	0	24798	9	11	0	0	0	0
Nordic walking	0	0	0	1	0	11	28277	0	0	1	1	0
ascending stairs	1	0	0	1	8	7	0	17247	217	25	7	0
descending stairs	0	1	22	8	1	1	0	41	15624	6	8	0
vacuum cleaning	5	32	168	1	1	0	0	0	3	26076	48	0
ironing	0	30	368	0	0	0	0	0	2	20	35334	0
rope jumping	0	0	0	0	4	4	3	0	0	0	0	7477

Fig: 2.7.1.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. ironing is misclassified as descending stairs (368)
2. ascending stairs is misclassified as descending stairs (217)

3. vacuum cleaning is misclassified as standing (168)
4. sitting is misclassified as standing (165).

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.9907	0.9890	0.9882	2
2	0.9910	0.9889	0.9888	1.85
3	0.9906	0.9887	0.9883	2
4	0.9900	0.9881	0.9876	2
5	0.9907	0.9885	0.9883	1.97

Table: 2.7.1.2

Of the 5-fold validations, the best test score accuracy is for validation-2 and the value is 98.88%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28743	5	11	2	0	0	0	33	2	0	16	0
sitting	8	27674	81	0	0	0	0	0	0	4	22	0
standing	3	126	28140	5	0	0	2	23	6	59	101	0
walking	2	0	4	35297	2	80	112	35	10	73	0	0
running	1	0	0	3	14753	7	4	52	6	0	1	12
cycling	0	4	0	62	7	24614	94	29	0	1	3	5
Nordic walking	2	0	0	116	0	83	28059	10	0	19	1	1
ascending stairs	0	3	2	88	0	18	1	17062	210	123	6	0
descending stairs	1	16	2	29	9	1	2	332	15212	72	36	0
vacuum cleaning	2	46	127	133	6	19	0	118	23	25684	176	0
ironing	7	8	65	0	0	8	0	27	18	134	35487	0
rope jumping	0	0	0	0	10	27	0	0	0	0	0	7451

Fig: 2.7.1.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. vacuum cleaning is misclassified as walking (133)
2. vacuum cleaning is misclassified as standing (127)
3. Nordic walking is misclassified as walking (116)

Model trained on the subset of data has slightly less accuracy but when compared with respect to time taken for training, it outperformed the model trained on complete data. Both the models misclassified nordic walking as walking.

2.7.2 Linear regression:

Linear regression shows the relationship between dependent and one or more independent variables, since it is linear the change in value of dependent variable shows corresponding change in value of independent variable. Using this we can predict the value of unknown dependent variable.

Fold No	Training Score	Validation Score	Test Score	Time Taken
1	0.8577	0.9950	0.9948	10.17
2	0.8577	0.9948	0.9947	10.05
3	0.8578	0.8575	0.9946	9.99
4	0.8576	0.8578	0.9944	9.97
5	0.8574	0.8577	0.9947	8.65

Table: 2.7.2.1

Of the 5-fold validations, the best test score accuracy is for validation-0 and the value is 99.48%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	28055	52	316	0	82	0	0	2	0	148	113	44
sitting	187	26116	921	77	0	0	0	2	0	318	168	0
standing	17	523	25395	1	0	0	0	1	5	817	1702	4
walking	3	0	42	31917	1	102	1855	667	1010	10	8	0
running	21	2	9	216	11622	149	785	570	322	128	20	995
cycling	0	0	0	70	99	23473	508	277	242	147	1	2
Nordic walking	20	13	20	2365	80	367	23140	906	1083	39	111	147
ascending stairs	161	17	339	1575	728	184	1081	10469	1516	1272	86	85
descending stairs	86	53	30	2344	314	429	1292	2091	8139	742	87	105
vacuum cleaning	98	461	690	510	61	104	73	517	453	22045	1322	0
ironing	0	202	1289	0	0	50	1	32	28	432	33720	0
rope jumping	7	3	1	5	1070	54	484	209	10	2	3	5640

Fig: 2.7.2.1

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. Standing is misclassified as vacuum cleaning (817)
2. ironing is misclassified as descending stairs (368)
3. ascending stairs is misclassified as descending stairs (217)

4. vacuum cleaning is misclassified as standing (168)
5. sitting is misclassified as standing (165).

On running 5-fold cross validation on the split data the train, validation and test score along with time taken is displayed in the below table.

Fold No	Training Score	Validation Score	Test Score	Time Taken (in mins)
1	0.6678	0.6676	0.6676	0.76
2	0.6680	0.6670	0.6670	0.83
3	0.6678	0.6683	0.6683	0.75
4	0.6679	0.6683	0.6683	0.75
5	0.6677	0.6677	0.6677	0.76

Table: 2.7.2.2

Of the 5-fold validations, the best test score accuracy is for validation-2 and the value is 66.83%. The confusion matrix for the best classifier is displayed below

	lying	sitting	standing	walking	running	cycling	Nordic walking	ascending stairs	descending stairs	vacuum cleaning	ironing	rope jumping
lying	26745	537	385	297	0	241	45	0	1	357	204	0
sitting	551	21670	2205	152	0	8	208	0	18	138	2839	0
standing	821	1914	17760	2513	0	38	151	0	2	1648	3618	0
walking	47	2	1078	25907	1	180	2644	454	198	4661	434	9
running	47	50	90	602	8500	425	1717	796	980	97	134	1401
cycling	10	130	4	574	47	20234	1492	134	1337	645	178	34
Nordic walking	66	365	20	2969	695	2028	17959	1365	812	1687	294	31
ascending stairs	99	69	657	3105	1136	165	1283	6956	1358	2422	203	60
descending stairs	115	39	632	2047	1153	2519	1412	2546	2570	1934	663	82
vacuum cleaning	441	96	2307	2744	0	114	923	1658	1264	14105	2682	0
ironing	601	1178	2483	328	0	604	176	10	149	1712	28513	0
rope jumping	0	0	0	113	2370	329	658	254	65	2	0	3697

Fig: 2.7.2.2

From the confusion matrix it can be inferred that the more number of misclassified are there for below attributes

1. walking is misclassified as vacuum cleaning (4661)
2. standing is misclassified as ironing (3618)
3. ascending stairs is misclassified as walking (3105)

4. Nordic walking is misclassified as walking (2969)

Here, the accuracy has decreased and number of misclassifications increased eventhough the time taken to train has reduced.

3. Conclusion:

From the above report we can conclude that,

- It can be noted that the most misclassified activities from all model are:
    - Ironing ↔standing
    - Ascending stairs ↔ descending stairs
    - Nordic walking ↔walking

- Based on the accuracy results we can tell that Random forest and SVM performed the best, and Ridge regression has given the least accuracy. On a general note ensemble methods performed better than single classifiers.

- The training time of linear regression was highest, while the least training time was for Adaboost .

- Basen on speed (training samples/time) Linear regression performed the best but the accuracy was least, comparatively the decision tree with Gini index is the optimum solution.

- Training the models on a subset of data gives a slight decrease in accuracy, but with respect to time these models were faster.

- Accuracy of decision tree with entropy is greater than the accuracy of decision tree with Gini impurity.

- For this dataset Linear regression has outperformed the Ridge regression.

4. Contribution:

Vivek (2022701001) : implementing Adaboost and Decision tree(with Gini) algorithm and making presentation.

Kesavaraj (2022701008) : implementing Random forest and Ridge regression algorithm and making report.

Poorvaja (2020101096): implementing SVM and Decision tree (with entropy) regression algorithm and making report.

Rohan Nain(2020102023): implementing Neural Network and Linear regression algorithm and making presentation.