A study on Heterogeneity Activity Recognition Data Set

Chandra S Kundu

Content

- Introduction
- Goal
- Heterogeneity Activity Recognition Dataset
- Preprocessing Data
- Models
- Result Analysis
- Scope of Improvement

Introduction

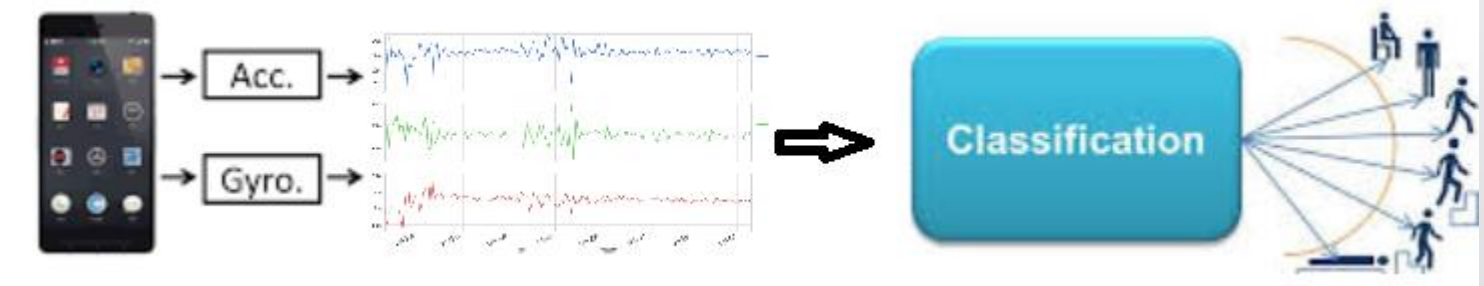
- Smart Devices becoming sophisticated day by day
- These devices incorporating powerful sensors

Introduction

- Data from the sensors can be used to recognize human activity
- Solve health-care issues such as automatic and intelligent daily activity monitoring for elderly people

Goal

Apply and compare machine learning algorithms on the data that is collected from two frequent used accelerometer and gyroscope sensors, to predict human activities.



Heterogeneity Activity Recognition Dataset

- Data Collected from UCI machine learning data repository
- Dataset size: 43930257 instances and 10 attributes. 3.3 GB.
- Data gathered from accelerometer and gyroscope sensor by conducting an extensive analysis on 9 users using 31 smartphones, 4 smartwatches and 1 tablet
- Sample Data:

Index	Arrival_Tir	Creation_T	x	y	z	User	Model	Device	gt
0	1.42E+12	1.42E+18	-5.95819	0.688065	8.135345	a	nexus4	nexus4_1	stand
1	1.42E+12	1.42E+18	-5.95224	0.670212	8.136536	a	nexus4	nexus4_1	stand
2	1.42E+12	1.42E+18	-5.99509	0.653549	8.204376	a	nexus4	nexus4_1	stand
3	1.42E+12	1.42E+18	-5.94272	0.676163	8.128204	a	nexus4	nexus4_1	stand
4	1.42E+12	1.42E+18	-5.99152	0.641647	8.135345	a	nexus4	nexus4_1	stand
5	1.42E+12	1.42E+18	-5.96533	0.629746	8.128204	a	nexus4	nexus4_1	stand
6	1.42E+12	1.42E+18	-5.99152	0.635696	8.16272	a	nexus4	nexus4_1	stand
7	1.42E+12	1.42E+18	-5.91534	0.630936	8.105591	a	nexus4	nexus4_1	stand
8	1.42E+12	1.42E+18	-5.98438	0.694016	8.067505	a	nexus4	nexus4_1	stand

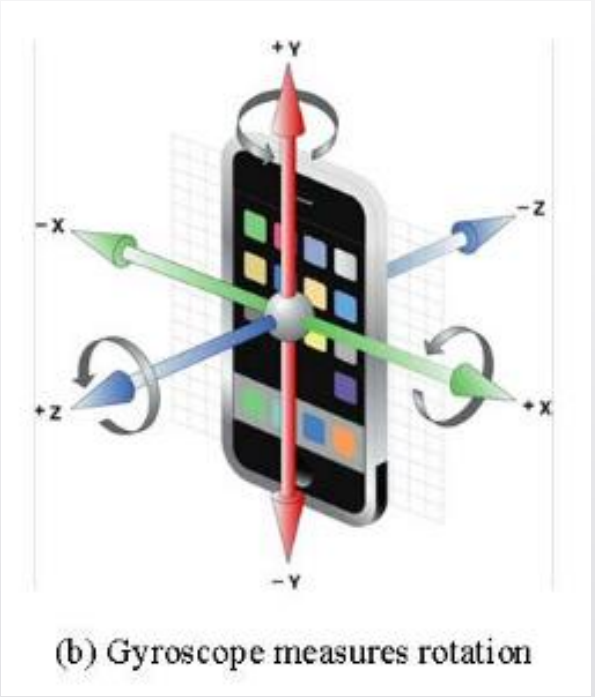
Accelerometer Sensor

(a) Accelerometer measures acceleration

- Computes the linear acceleration of the device on the X- axis (lateral), Y-axis (longitudinal), and Z-axis (vertical)
- For example, if activity changes from walking to biking, abrupt change of acceleration pattern along the vertical axis

Gyroscope Sensor

- Measure the device's rotation rate by detecting the roll, pitch, and yaw motions along the x, y, and z axis respectively



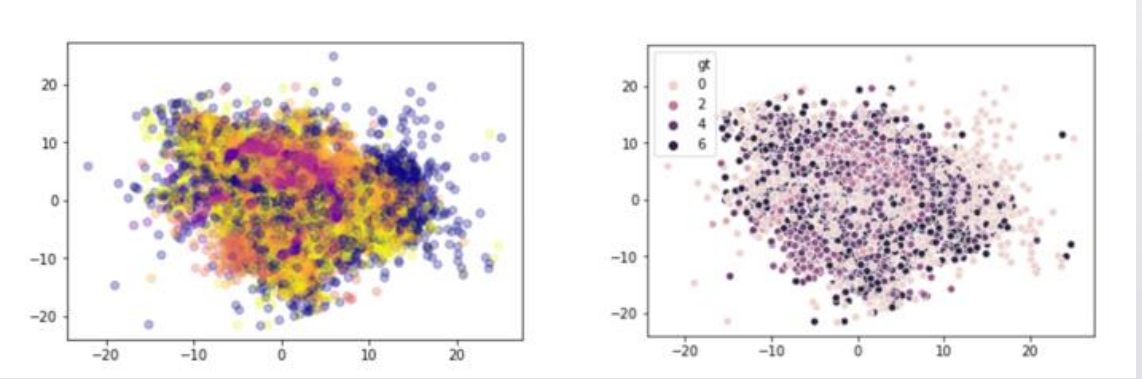
(b) Gyroscope measures rotation

Preprocessing Data

- Dataset almost preprocessed
- Drop the row which has missing value in the ground-truth or activity class.
- Remove extra features such as 'Index', 'Arrival_Time', 'Creation_Time', 'User', 'Model', 'Device', for our project, which are not necessary
- Do cluster analysis to see the pattern of the data

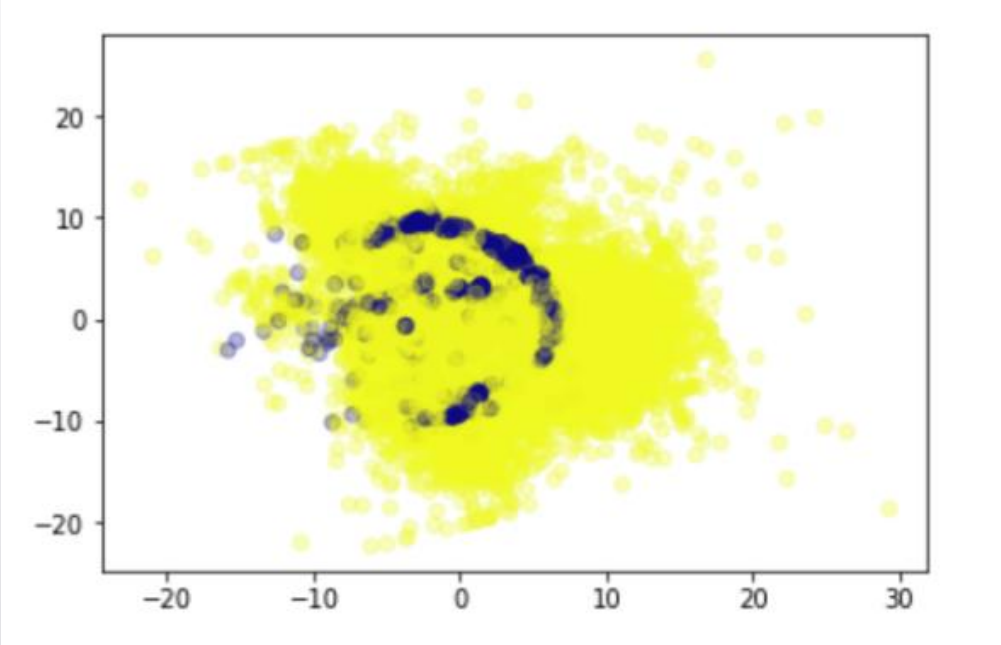
Cluster Analysis

- Apply dimensionality reduction technique(PCA) to draw every datapoints by giving same color to every datapoints in the same activity
- Difficult to detect activity using any classification technique
- Poorly calibrated sensor or any other factors can cause this problem



Cluster Analysis

- Convert multiclass activity into binary class
  - Assign standing and sitting as 0
  - Assign walking, stair up and down, and biking as 1
- Apply dimensionality reduction technique(PCA) to draw every datapoints by giving same color to every datapoints in the same activity
- Data clearly divided into two part.

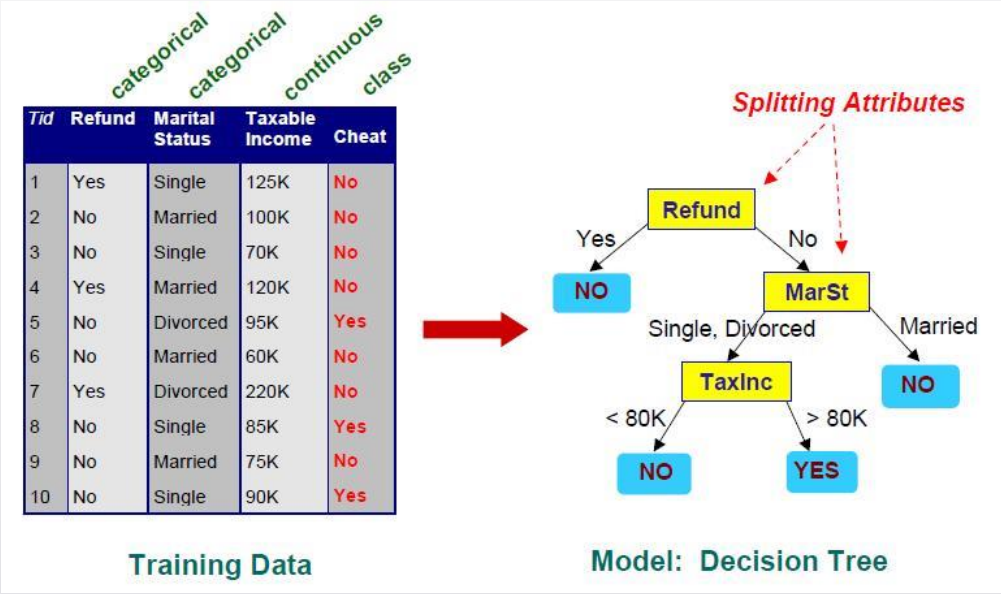


Models

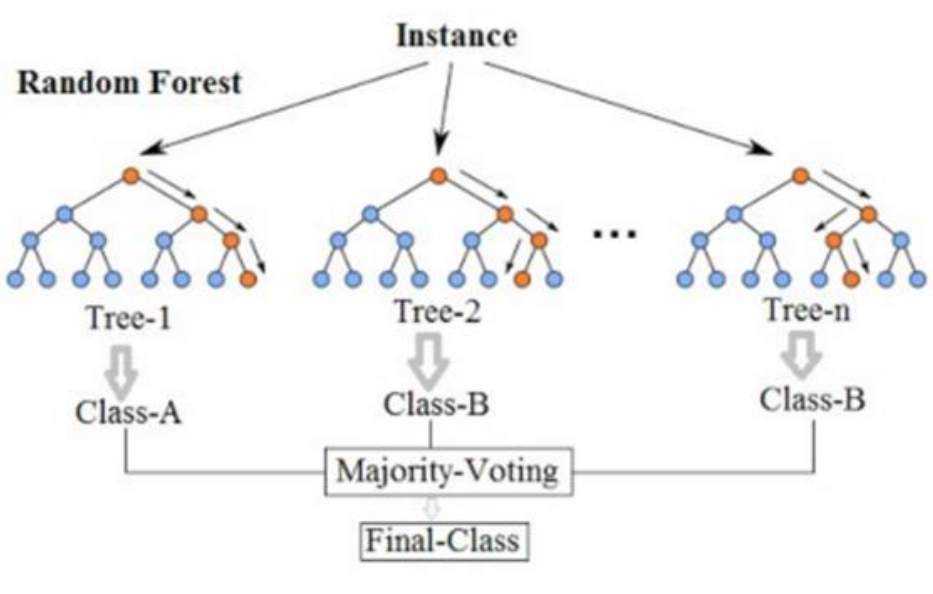
- Decision Tree
- K-Nearest Neighbors
- Random Forest
- Support Vector Machine

Decision Tree

- Classifies data in a tree structure
- Requires little effort from users for data preparation and computationally efficient



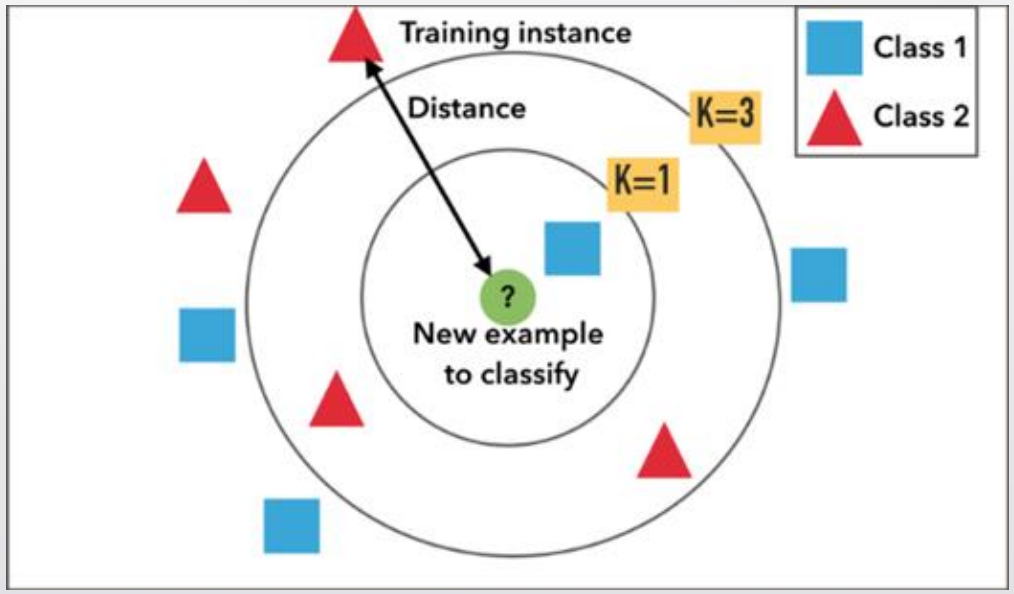
Random Forest



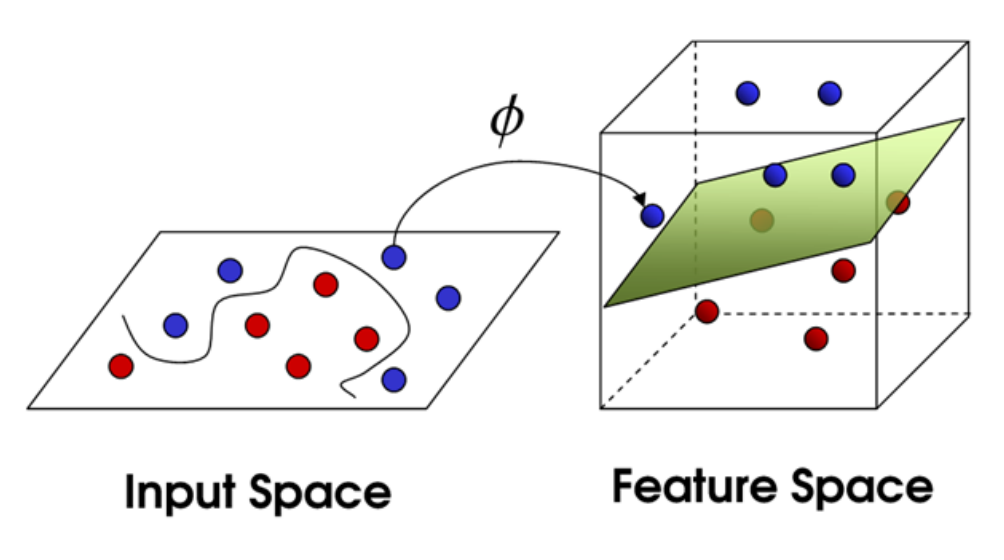
- Ensemble learning method for classification that operates by constructing a multitude of decision trees
- Requires little effort from users for data preparation and computationally efficient

K-Nearest Neighbors

- Classifies data a majority vote of its neighbors,
- A memory-based approach
- Allows the algorithm to respond quickly to changes in the input during real-time use

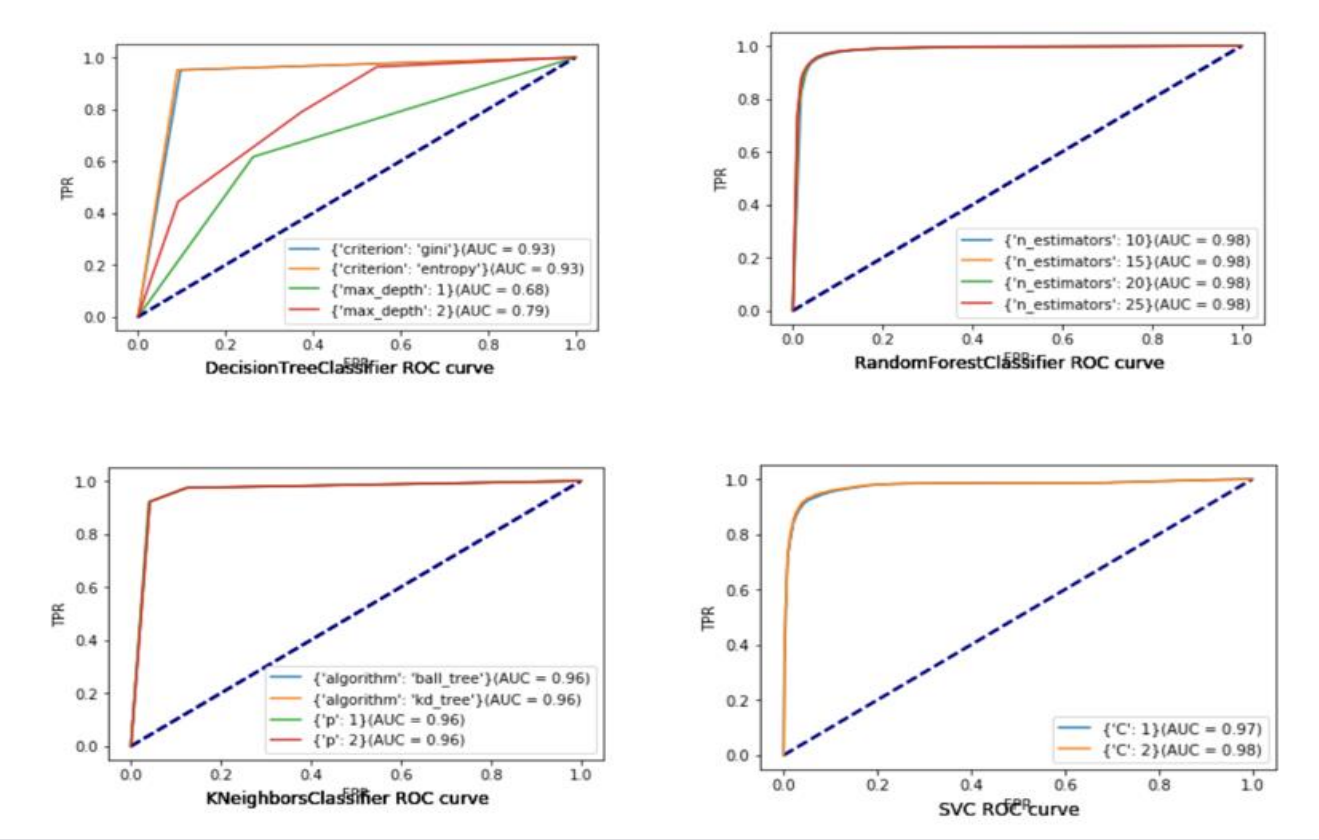


K-Nearest Neighbors

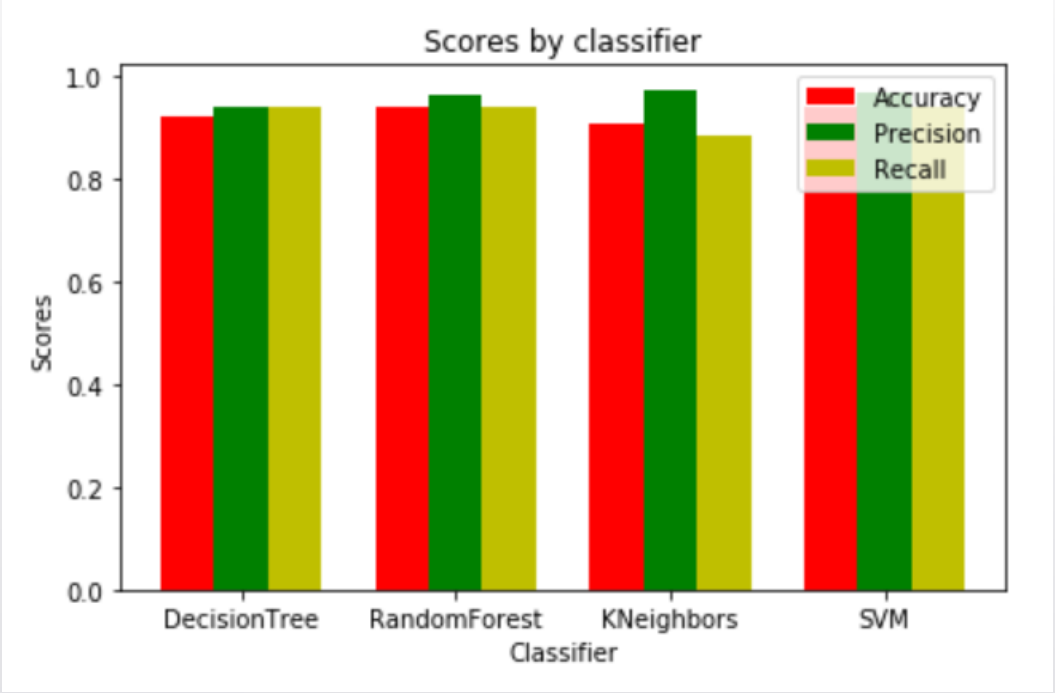


- Find hyperplane to separate classes by maximizing margin between class.
- Better performance since classifies data only on a subset of points

Result Analysis

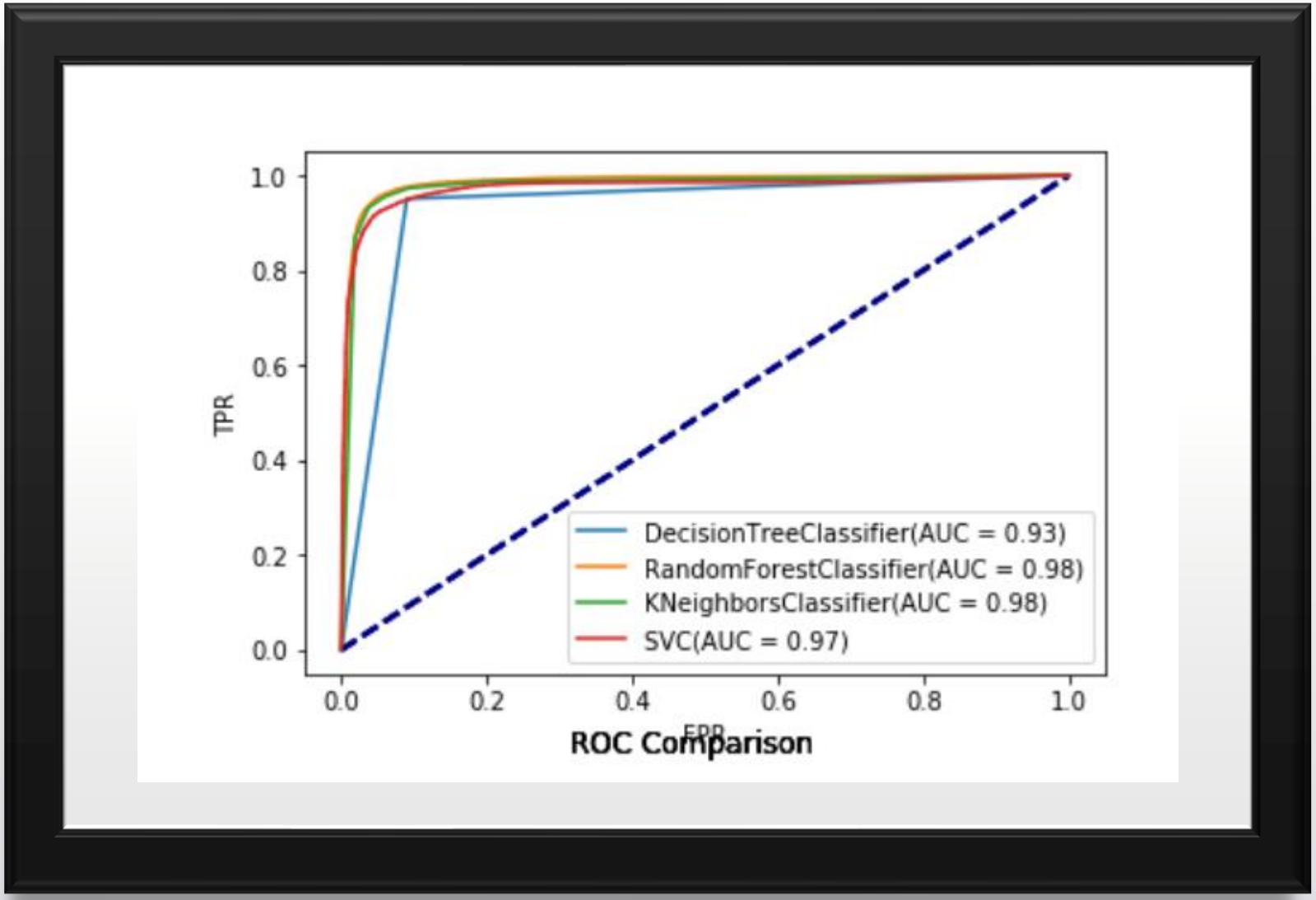


Result Analysis



	DecisionTree	RandomForest	KNeighbors	SVM
Accuracy	0.922000	0.939500	0.928000	0.884500
Precision	0.944700	0.959199	0.965764	0.964807
Recall	0.936073	0.948250	0.923135	0.855403

Result Analysis



Result Analysis

- In terms of AUC, Random forest and KNN outperform other classifier
- Random forest has the most accuracy and recall score
- KNN gives the most precision
- Best Classifier: Random Forest

Scope of Improvement

- Activation time and Creation time is other two features in the dataset that can be used.
  - Data processing would be more complex
  - Time series technique can be used
- More complex algorithm such us neural network can be used

Reference

- Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen "Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition" In Proc. 13th ACM Conference on Embedded Networked Sensor Systems (SenSys 2015), Seoul, Korea, 2015. [Web Link]

Thank you