Wireless Sensor Network Project Report on

Classification of ADL data collected from WSN

Anusha Prakash (12IT10)

Bhuvan M S (12IT16)

Siddharth Jain (12IT78)

Under the Guidance of,

MS. Deepthi

Department of Information Technology, NITK Surathkal

Date of Submission: 21 April 2016

in partial fulfillment for the award of the degree

of

Bachelor of Technology In **Information Technology** At

Department of Information Technology

National Institute of Technology Karnataka, Surathkal

April 2016

Department of Information Technology, NITK Surathkal
Wireless Sensor Network Project
End Semester Evaluation Report (April 2016)

Course Code : IT TODO

Course Title: Wireless Sensor Network Project

Project Title: *Classification of ADL data collected from WSN*

Project Group:

Name of the Student	Register No.	Signature with Date
Anusha Prakash	12IT10	
Bhuvan M S	12IT16	
Siddharth Jain	12IT78	

Place:NITK, Surathkal

Date:18 April 2016 *(Name and Signature of Wireless Sensor Network Project Guide)*

Abstract

Keywords: *Wireless Sensor Network; Machine Learning CLassifier; Naive Bayes; Multilayer Perceptron; Bayes Network; Random Forest; Adaboost; Logistic Regression*

Contents

1 **Introduction** 1

2 **Literature Survey** 2

2.1 Problem Statement 2

2.2 Objectives 2

3 **Methodology** 3

3.1 ADL Classification Model 3

3.1.1 Data description 3

3.1.2 Work Flow 3

3.1.3 Data Preprocessing 3

3.1.4 Machine learning algorithms 3

3.1.5 Evaluation metrics 5

3.2 HEED Protocol 6

4 **Results and Analysis** 7

4.1 ADL Classification Model 7

4.1.1 User A 7

4.1.2 User B 7

4.2 HEED Protocol 8

5 **Conclusion & Future Work** 9

List of Figures

3.1 Flowchart representing the Methodology work flow. 11
4.1 ROC curves for User A. 12
4.2 PR curves for User A. 13
4.3 ROC curves for User B. 14
4.4 PR curves for User B. 15

List of Tables

1 Features details 3
2 Performance comparison of different classifiers for User A. 7
3 Performance comparison of different classifiers for User B. 8

1 Introduction

2 Literature Survey

2.1 Problem Statement

The aim of our project is to classify the ADL of the user using the data sensed by sensor nodes in the wireless sensor networks by modeling separate user dependent machine learning model and analyse the performance of various classifiers.

2.2 Objectives

1. To develop ADL prediction model on the data collected from sesor nodes in a Wireless Sensor Network.

2. Evaluate each prediction model using Receiver Operating Characteristics (ROC), Precision-Recall Curves, F1 Measure.

3. To implement simple version of HEED protocol to illustrate how the ADL data could be collected from the Wireless Sensor Network.

Table 1: Features details

Feature	**Type**
Start Timestamp	Numeric
End Timestamp	Numeric
Duration	Numeric
Location	Nominal
Type	Nominal
Place	Nominal
Class: Activity	10 Nominal Classes

3 Methodology

3.1 ADL Classification Model

3.1.1 Data description

3.1.2 Work Flow

The workflow of the methodology has been described in the Figure 3.1

3.1.3 Data Preprocessing

The cleaned, preprocessed and normalized data contains 408 number of instances for User-A and 2334 number of instances for User-B. The features are tabulated in the Table 1. The classes (ADL) are: $\text{Spare}_Time/TV, Grooming, Toileting, Sleeping, Breakfast, Showering, Snack,$

3.1.4 Machine learning algorithms

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

We have chosen following machine learning algorithms for the project.

3.1.4.1 Naive Bayes In machine learning, naive Bayes classifiers [5] are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive)

independence assumptions between the features.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, Naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method; Russell and Norvig note that "[naive Bayes] is sometimes called a Bayesian classifier, a somewhat careless usage that has prompted some Bayesians to call it the idiot Bayes model."

3.1.4.2 Multilayer Perceptron - 2 nodes in one hidden layer

3.1.4.3 Bayesian Network A Bayesian network [7] is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph. Formally, Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes that are not connected (there is no path from one of the variables to the other in the bayesian network) represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. For example, if m parent nodes represent m Boolean variables then the probability function could be represented by a table of $2\hat{m}$ entries, one entry for each of the $2\hat{m}$ possible combinations of its parents being true or false. Similar ideas may be applied to undirected, and possibly cyclic, graphs; such are called Markov networks.

Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences) are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence

diagrams.

3.1.4.4 Random Forest

3.1.4.5 Adaboost

3.1.4.6 Logistic Regression

3.1.5 Evaluation metrics

The following evaluation metrics are calculated for each prediction model.

3.1.5.1 Accuracy Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

3.1.5.2 F1-Measure The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

$$F1score = (2 * Precision * Recall)/(Precision + Recall) \tag{2}$$

3.1.5.3 Receiver Operating Characteristic Curve: ROC In statistics, a receiver operating characteristic (ROC) [2], or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

The Area Under the Curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. This measure is used to evaluate different binary classifiers. Higher the ROC AUC the better is the classifier.

3.1.5.4 Precision and Recall Curve In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

1. Precision (P): is defined as the fraction of ground truth positives among all the examples that are predicted to be positive.

$$Precision = TP/(TP + FP) \tag{3}$$

   Where, TP, FP stand for True Positives and False Positives respectively

2. Recall (R): is defined as the fraction of ground truth positives that are predicted to be positives.

$$Recall = TP/(TP + FN) \tag{4}$$

   Where, FN denotes False Negatives.

3. Precision-Recall Curve [3]: The tradeoff between the precision and recall can be studied by looking at the plot of how precision changes as a function of recall. This plot is known as the Precision-Recall (PR) curve. The precision is plotted on the y-axis and recall on x-axis. The area under precision recall curve is a cost-effective metric for evaluation of classifiers. High value of area under the PR curve indicates better performance. Area under the PR curve is preferred way of studying the performance of classifiers in skewed (i.e. imbalance between positives and negatives) datasets such as the one considered in this work [3].

3.2 HEED Protocol

Table 2: Performance comparison of different classifiers for User A.

Algorithm	**F1_Score**	**Accuracy**
MLP_CLASSIFIER_2	0.328498863313	0.441176470588
LOGISTIC_REGRESSION	0.752633628582	0.764705882353
BAYES_NETWORK	0.720978373888	0.735294117647
NAIVE_BAYES	0.764874797931	0.764705882353
RANDOM_FOREST	0.64279135996	0.656862745098
ADABOOST	0.110876854162	0.254901960784

4 Results and Analysis

The classifier algorithm learns on the training data and saves a trained model, which is hence used to predict the probabilities of the occurrence of each class for each test instance. These scores are used to calculate the weighted area under the ROC curve and the weighted area under the PR curve, F1 score and plot the ROC and PR curves.

4.1 ADL Classification Model

THe results of the ADL prediction models for User-A and User-B have been discussed in following subsections.

4.1.1 User A

4.1.1.1 ROC and PR plots The ROC curves and PR curves for all algorithms for User-A have been presented in Figure 4.1 and Figure 4.3 respectively.

TODO: DISCUSS

4.1.1.2 Comparison of Results Summary of the results containing the scores from various evaluation metrics as explained in the Methodology section is shown in Table 2.

TODO:DISCUSS

4.1.2 User B

4.1.2.1 ROC and PR plots The ROC curves and PR curves for all algorithms for User-B have been presented in Figure 4.3 and Figure 4.3 respectively.

Table 3: Performance comparison of different classifiers for User B.

Algorithm	**F1_Score**	**Accuracy**
MLP_CLASSIFIER_2	0.278021577872	0.349914236707
LOGISTIC_REGRESSION	0.287792488475	0.360205831904
BAYES_NETWORK	0.317570449807	0.391080617496
NAIVE_BAYES	0.289539715817	0.36192109777
RANDOM_FOREST	0.309216828748	0.368782161235
ADABOOST	0.0966266437965	0.245283018868

TODO: DISCUSS

4.1.2.2 Comparison of Results

Summary of the results containing the scores from various evaluation metrics as explained in the Methodology section is shown in Table 2.

TODO:DISCUSS

4.2 HEED Protocol

TODO: HEED Protocol results

5 Conclusion & Future Work

References

heed TODO: HEED citationTODO: Dataset citation Receiver Operating Characteristics. Wikipedia, The Free Encyclopedia.Web.2016. Precision Recall Curves. Wikipedia, The Free Encyclopedia.Web.2016. F1 Score. Wikipedia, The Free Encyclopedia.Web.2016. Naive Bayes. Wikipedia, The Free Encyclopedia.Web.2016. Bayesian Networks. Wikipedia, The Free Encyclopedia.Web.2016. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. TODO: LIT REVIEW CITATIONS

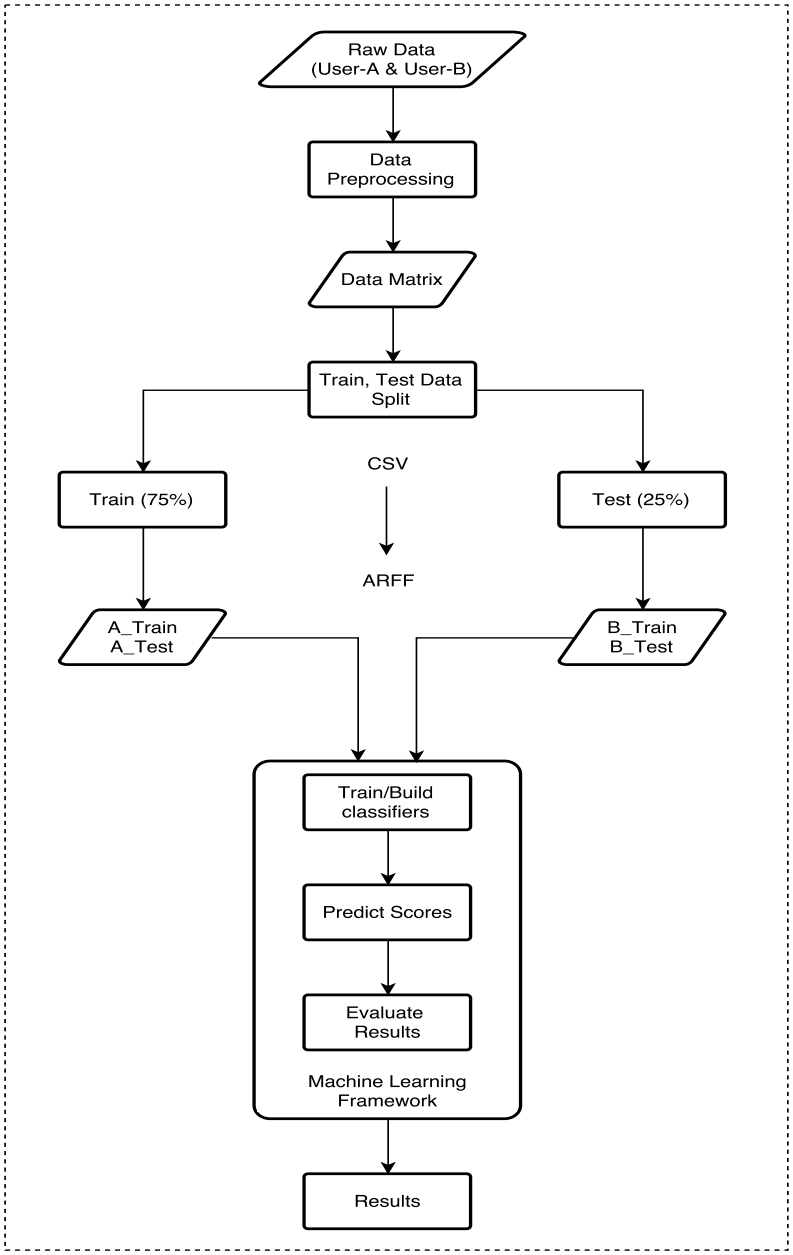
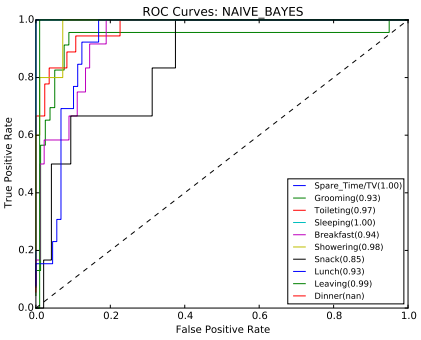
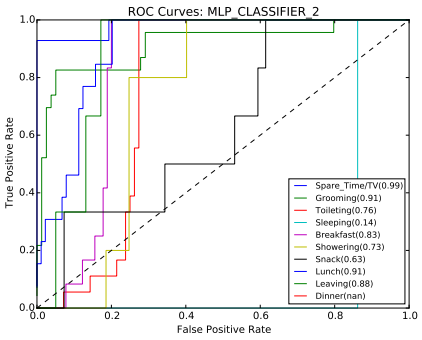


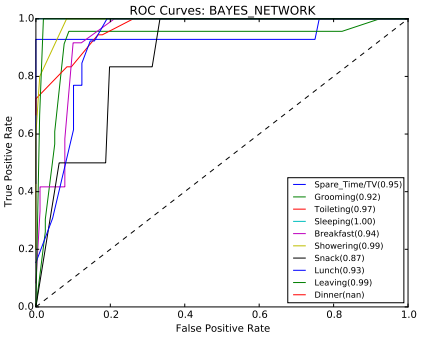
Figure 3.1: Flowchart representing the Methodology work flow.



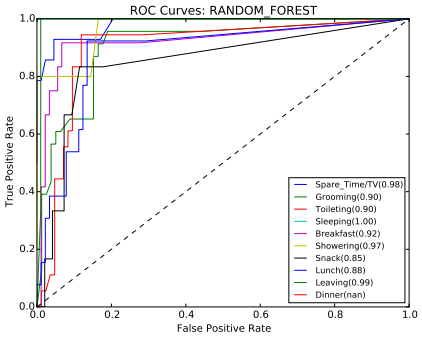
(a) Naive Bayes



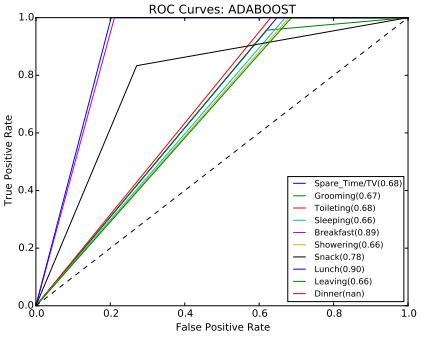
(b) MLPC-2



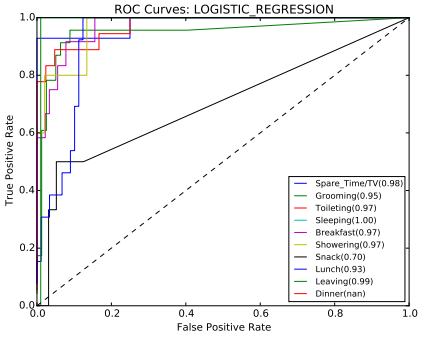
(c) Bayesian Network



(d) Random Forest

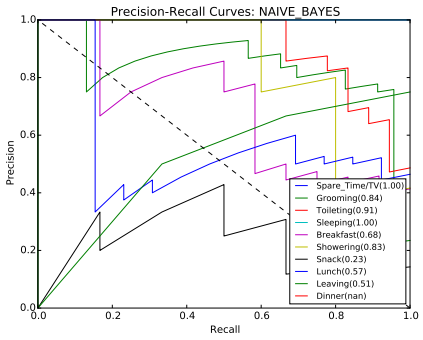


(e) Adaboost

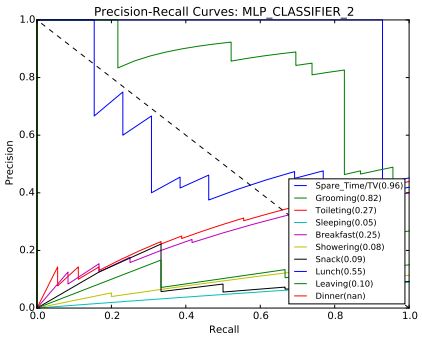


(f) Logistic Regression

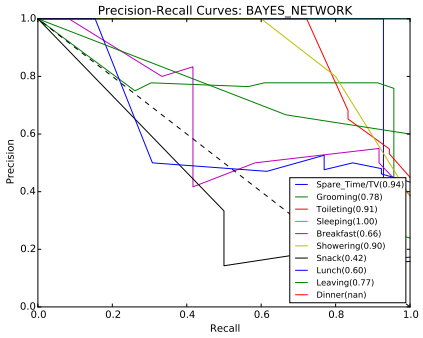
Figure 4.1: ROC curves for User A.



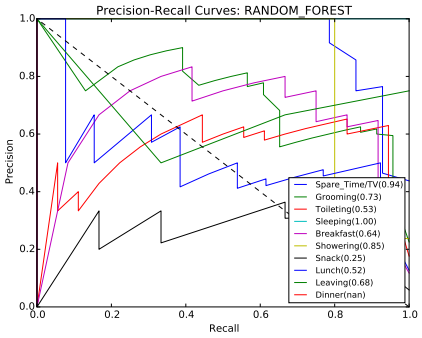
(a) Naive Bayes



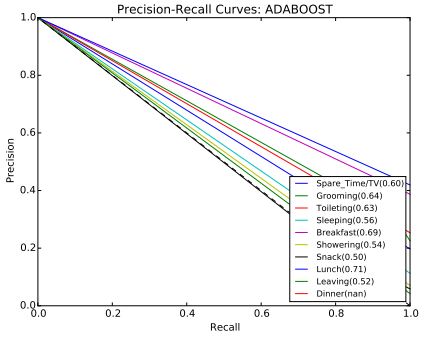
(b) MLPC-2



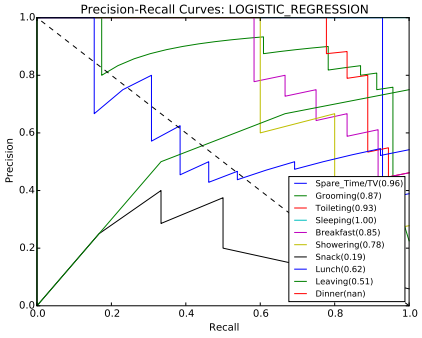
(c) Bayesian Network



(d) Random Forest

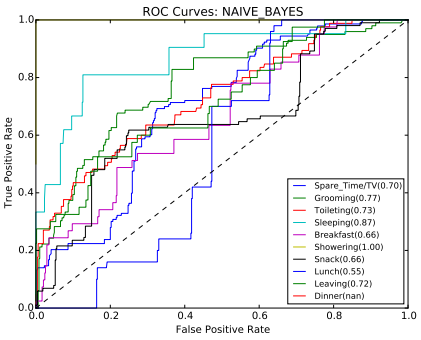


(e) Adaboost

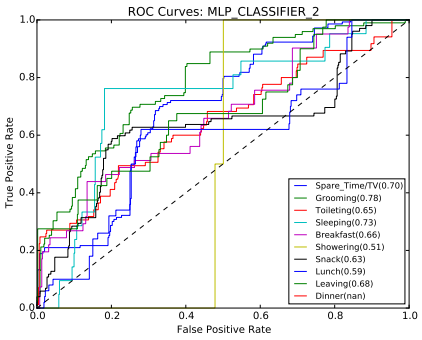


(f) Logistic Regression

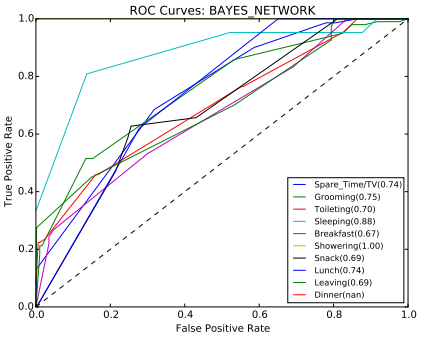
Figure 4.2: PR curves for User A.



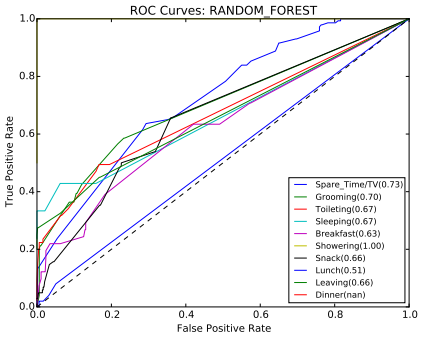
(a) Naive Bayes



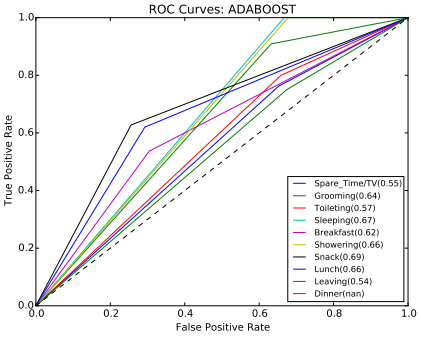
(b) MLPC-2



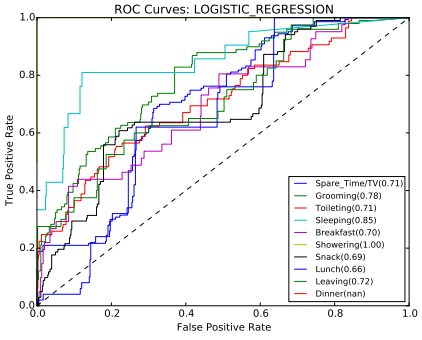
(c) Bayesian Network



(d) Random Forest

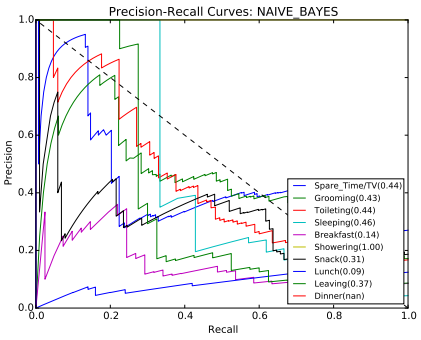


(e) Adaboost

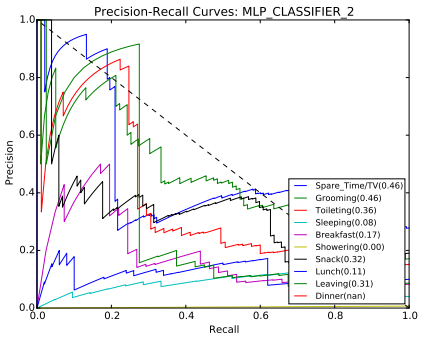


(f) Logistic Regression

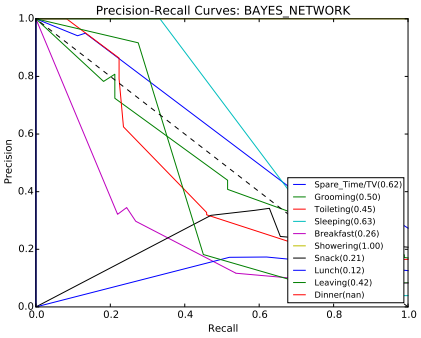
Figure 4.3: ROC curves for User B.



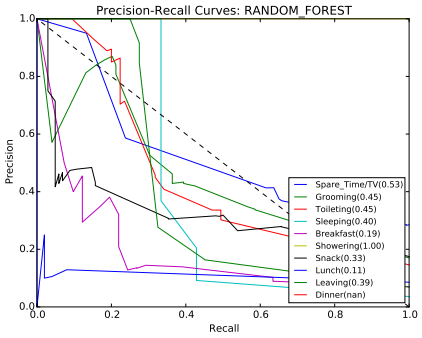
(a) Naive Bayes



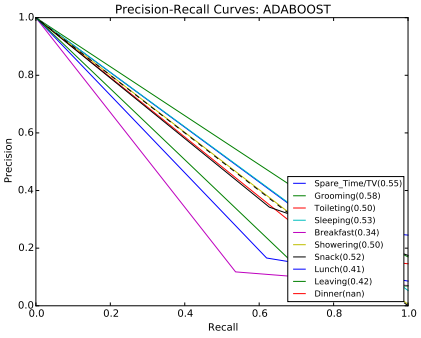
(b) MLPC-2



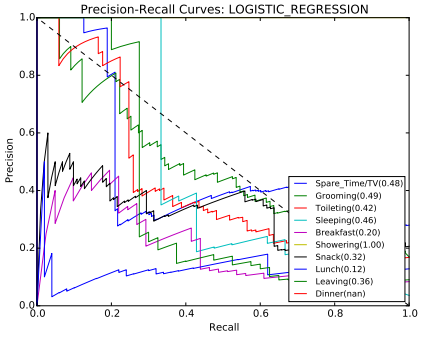
(c) Bayesian Network



(d) Random Forest



(e) Adaboost



(f) Logistic Regression

Figure 4.4: PR curves for User B.