# **Table of contents**

[1. Introduction 1](#_Toc508328293)

[1.1. Algorithm inputs 1](#_Toc508328294)

[1.2. Algorithm output 1](#_Toc508328295)

[1.1.1. Single-target feature ranking 1](#_Toc508328296)

[1.1.2. Multi-target feature ranking 3](#_Toc508328297)

[2. Tests 5](#_Toc508328298)

[3. Integration tests output 7](#_Toc508328299)

# Introduction

Methods for feature ranking are used in many domains with many descriptive variables and high-dimensional problems. The obtained rankings (feature importance scores) provide an additional insight about the importance of the variables for the target and/or reduce the dimensionality of the problem. Many real-life problems have structured targets that need to be predicted. However, the task of feature ranking in the context of predicting structured target variables is more complex than the same task for simple classification or regression. Typical approaches for this task decompose the output to primitive components, perform feature ranking on these smaller problems, and then aggregate the resulting rankings into a single ranking. Such an approach ignores the dependencies between components of the structured target variable. This approach is based on ensembles of predictive clustering trees and treats structured variables directly.

The algorithm takes the following parameters:

* *Descriptive variables:* A set of descriptive variables, whose rankings (importance scores) we are interested in. They can be numeric or categoric or mix of both types.
* *Target variables:* A set of one or more (dependent) variables in relation to which we are estimating the importance of descriptive variables. They can either be all numeric or all categoric.
* *Number of decision trees in ensemble:* In general, more trees in the ensemble improves the accuracy, but at the expense of a higher computational cost.

## Algorithm inputs

The algorithm supports feature ranking for input attributes predicting (multiple) numeric or (multiple) nominal attributes.

## Algorithm output

As needed by the MIP, our algorithm produces a tabular data resource output.

### Single-target feature ranking

#### Tabular data resource

|  |
| --- |
| {  "data": [  {  "Root mean squared error (RMSE)-overall": 0.05816588189863536,  "Root mean squared error (RMSE)-overallRank": 1,  "attributeDatasetIndex": 9,  "attributeName": "input9",  "id": 1  },  {  "Root mean squared error (RMSE)-overall": 0.039558951770772885,  "Root mean squared error (RMSE)-overallRank": 2,  "attributeDatasetIndex": 1,  "attributeName": "input1",  "id": 2  },  {  "Root mean squared error (RMSE)-overall": 0.03287201595938097,  "Root mean squared error (RMSE)-overallRank": 3,  "attributeDatasetIndex": 8,  "attributeName": "input8",  "id": 3  },  {  "Root mean squared error (RMSE)-overall": 0.02032144054422691,  "Root mean squared error (RMSE)-overallRank": 4,  "attributeDatasetIndex": 5,  "attributeName": "input5",  "id": 4  },  {  "Root mean squared error (RMSE)-overall": 0.015513243213930005,  "Root mean squared error (RMSE)-overallRank": 5,  "attributeDatasetIndex": 2,  "attributeName": "input2",  "id": 5  },  {  "Root mean squared error (RMSE)-overall": 0.01149432661300421,  "Root mean squared error (RMSE)-overallRank": 6,  "attributeDatasetIndex": 7,  "attributeName": "input7",  "id": 6  },  {  "Root mean squared error (RMSE)-overall": 0.009394524080608193,  "Root mean squared error (RMSE)-overallRank": 7,  "attributeDatasetIndex": 3,  "attributeName": "input3",  "id": 7  },  {  "Root mean squared error (RMSE)-overall": -0.0031947642109536032,  "Root mean squared error (RMSE)-overallRank": 8,  "attributeDatasetIndex": 6,  "attributeName": "input6",  "id": 8  },  {  "Root mean squared error (RMSE)-overall": -0.0140351392743078,  "Root mean squared error (RMSE)-overallRank": 9,  "attributeDatasetIndex": 4,  "attributeName": "input4",  "id": 9  }  ],  "name": "feature-importances",  "profile": "tabular-data-resource",  "schema": {  "fields": [  {  "name": "id",  "type": "integer"  },  {  "name": "attributeDatasetIndex",  "type": "integer"  },  {  "name": "attributeName",  "type": "string"  },  {  "name": "Root mean squared error (RMSE)-overallRank",  "type": "integer"  },  {  "name": "Root mean squared error (RMSE)-overall",  "type": "double"  }  ],  "primaryKey": "id"  }  } |

### Multi-target feature ranking

#### Tabular data resource

|  |
| --- |
| {  "data": [  {  "Root mean squared error (RMSE)-overall": 0.001715702477428658,  "Root mean squared error (RMSE)-overallRank": 1,  "attributeDatasetIndex": 16,  "attributeName": "input16",  "id": 1  },  {  "Root mean squared error (RMSE)-overall": -0.004471161474187219,  "Root mean squared error (RMSE)-overallRank": 2,  "attributeDatasetIndex": 14,  "attributeName": "input14",  "id": 2  },  {  "Root mean squared error (RMSE)-overall": -0.005468231685397082,  "Root mean squared error (RMSE)-overallRank": 3,  "attributeDatasetIndex": 9,  "attributeName": "input9",  "id": 3  },  {  "Root mean squared error (RMSE)-overall": -0.005498700425691784,  "Root mean squared error (RMSE)-overallRank": 4,  "attributeDatasetIndex": 13,  "attributeName": "input13",  "id": 4  },  {  "Root mean squared error (RMSE)-overall": -0.0061655296113774264,  "Root mean squared error (RMSE)-overallRank": 5,  "attributeDatasetIndex": 10,  "attributeName": "input10",  "id": 5  },  {  "Root mean squared error (RMSE)-overall": -0.007283615718304835,  "Root mean squared error (RMSE)-overallRank": 6,  "attributeDatasetIndex": 15,  "attributeName": "input15",  "id": 6  },  {  "Root mean squared error (RMSE)-overall": -0.008112668983072748,  "Root mean squared error (RMSE)-overallRank": 7,  "attributeDatasetIndex": 7,  "attributeName": "input7",  "id": 7  },  {  "Root mean squared error (RMSE)-overall": -0.008432520185961518,  "Root mean squared error (RMSE)-overallRank": 8,  "attributeDatasetIndex": 5,  "attributeName": "input5",  "id": 8  },  {  "Root mean squared error (RMSE)-overall": -0.008668725576155566,  "Root mean squared error (RMSE)-overallRank": 9,  "attributeDatasetIndex": 6,  "attributeName": "input6",  "id": 9  },  {  "Root mean squared error (RMSE)-overall": -0.009648170634604289,  "Root mean squared error (RMSE)-overallRank": 10,  "attributeDatasetIndex": 2,  "attributeName": "input2",  "id": 10  },  {  "Root mean squared error (RMSE)-overall": -0.010484475560180697,  "Root mean squared error (RMSE)-overallRank": 11,  "attributeDatasetIndex": 17,  "attributeName": "input17",  "id": 11  },  {  "Root mean squared error (RMSE)-overall": -0.01092530202275275,  "Root mean squared error (RMSE)-overallRank": 12,  "attributeDatasetIndex": 8,  "attributeName": "input8",  "id": 12  },  {  "Root mean squared error (RMSE)-overall": -0.013538697204628535,  "Root mean squared error (RMSE)-overallRank": 13,  "attributeDatasetIndex": 3,  "attributeName": "input3",  "id": 13  },  {  "Root mean squared error (RMSE)-overall": -0.013847485276771085,  "Root mean squared error (RMSE)-overallRank": 14,  "attributeDatasetIndex": 1,  "attributeName": "input1",  "id": 14  },  {  "Root mean squared error (RMSE)-overall": -0.0150264808485986,  "Root mean squared error (RMSE)-overallRank": 15,  "attributeDatasetIndex": 4,  "attributeName": "input4",  "id": 15  },  {  "Root mean squared error (RMSE)-overall": -0.015308931104010364,  "Root mean squared error (RMSE)-overallRank": 16,  "attributeDatasetIndex": 12,  "attributeName": "input12",  "id": 16  },  {  "Root mean squared error (RMSE)-overall": -0.017934486412840504,  "Root mean squared error (RMSE)-overallRank": 17,  "attributeDatasetIndex": 11,  "attributeName": "input11",  "id": 17  }  ],  "name": "feature-importances",  "profile": "tabular-data-resource",  "schema": {  "fields": [  {  "name": "id",  "type": "integer"  },  {  "name": "attributeDatasetIndex",  "type": "integer"  },  {  "name": "attributeName",  "type": "string"  },  {  "name": "Root mean squared error (RMSE)-overallRank",  "type": "integer"  },  {  "name": "Root mean squared error (RMSE)-overall",  "type": "double"  }  ],  "primaryKey": "id"  }  } |

# Tests

We have also prepared an integration test of the algorithm. The tests run the algorithm on the datasets which are available in the algorithm-factory-demo database. To setup the testing environment, the following docker configuration was used:

|  |
| --- |
| ---  version: '2'  services:  db:  image: postgres:9.6.5-alpine  hostname: db  environment:  POSTGRES\_PASSWORD: test  wait\_dbs:  image: "waisbrot/wait"  restart: "no"  environment:  TARGETS: "db:5432"  TIMEOUT: 60  create\_dbs:  image: "hbpmip/create-databases:1.0.0"  restart: "no"  environment:  DB\_HOST: db  DB\_PORT: 5432  DB\_ADMIN\_USER: postgres  DB\_ADMIN\_PASSWORD: test  DB1: features  USER1: features  PASSWORD1: featurespwd  DB2: woken  USER2: woken  PASSWORD2: wokenpwd  depends\_on:  - db  sample\_data\_db\_setup:  image: "hbpmip/sample-data-db-setup:0.5.0"  container\_name: "data-db-setup"  restart: "no"  environment:  FLYWAY\_DBMS: postgresql  FLYWAY\_HOST: db  FLYWAY\_PORT: 5432  FLYWAY\_DATABASE\_NAME: features  FLYWAY\_USER: postgres  FLYWAY\_PASSWORD: test  depends\_on:  - db  woken\_db\_setup:  image: "hbpmip/woken-db-setup:latest"  container\_name: "woken-db-setup"  restart: "no"  environment:  FLYWAY\_DBMS: postgresql  FLYWAY\_HOST: db  FLYWAY\_PORT: 5432  FLYWAY\_DATABASE\_NAME: woken  FLYWAY\_USER: postgres  FLYWAY\_PASSWORD: test  depends\_on:  - db  clus\_fr\_regression\_st:  image: "hbpmip/java-jsi-clus-fr:latest"  container\_name: "clus-fr-regression-st"  restart: "no"  environment:  NODE: job\_test  JOB\_ID: 1  IN\_JDBC\_DRIVER: org.postgresql.Driver  IN\_JDBC\_URL: jdbc:postgresql://db:5432/features  IN\_JDBC\_USER: features  IN\_JDBC\_PASSWORD: featurespwd  OUT\_JDBC\_DRIVER: org.postgresql.Driver  OUT\_JDBC\_URL: jdbc:postgresql://db:5432/woken  OUT\_JDBC\_USER: woken  OUT\_JDBC\_PASSWORD: wokenpwd  PARAM\_variables: "score\_test1"  PARAM\_covariables: "stress\_before\_test1,iq,cognitive\_task2,practice\_task2,response\_time\_task2,college\_math,score\_math\_course1,score\_math\_course2"  PARAM\_query: "SELECT stress\_before\_test1,score\_test1,iq,cognitive\_task2,practice\_task2,response\_time\_task2,college\_math,score\_math\_course1,score\_math\_course2 FROM SAMPLE\_DATA"  FUNCTION: java-jsi-clus-fr  PARAM\_MODEL\_size: 100  links:  - "db:db"  clus\_fr\_regression\_mt:  image: "hbpmip/java-jsi-clus-fr:latest"  container\_name: "clus-fr-regression-mt"  restart: "no"  environment:  NODE: job\_test  JOB\_ID: 2  IN\_JDBC\_DRIVER: org.postgresql.Driver  IN\_JDBC\_URL: jdbc:postgresql://db:5432/features  IN\_JDBC\_USER: features  IN\_JDBC\_PASSWORD: featurespwd  OUT\_JDBC\_DRIVER: org.postgresql.Driver  OUT\_JDBC\_URL: jdbc:postgresql://db:5432/woken  OUT\_JDBC\_USER: woken  OUT\_JDBC\_PASSWORD: wokenpwd  PARAM\_variables: "score\_test1,stress\_before\_test1,iq,cognitive\_task2"  PARAM\_covariables: "practice\_task2,response\_time\_task2,college\_math,score\_math\_course1,score\_math\_course2"  PARAM\_query: "SELECT stress\_before\_test1,score\_test1,iq,cognitive\_task2,practice\_task2,response\_time\_task2,college\_math,score\_math\_course1,score\_math\_course2 FROM SAMPLE\_DATA"  FUNCTION: java-jsi-clus-fr  PARAM\_MODEL\_size: 100  links:  - "db:db" |

# Integration tests output

|  |
| --- |
| [java-jsi-clus-fr] ./tests/test.sh [ 2:06AM]  Starting the databases...  Creating network "tests\_default" with the default driver  Creating tests\_db\_1 ...  Creating tests\_db\_1 ... done  Waiting for db:5432 .... up!  Everything is up  Starting tests\_db\_1 ... done  PLAY [localhost] \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  TASK [Create the new database(s)"] \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  changed: [localhost] => (item={'password': 'featurespwd', 'db': 'features', 'user': 'features'})  changed: [localhost] => (item={'password': u'wokenpwd', 'db': u'woken', 'user': u'woken'})  TASK [Create user(s)] \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  changed: [localhost] => (item={'password': 'featurespwd', 'db': 'features', 'user': 'features'})  changed: [localhost] => (item={'password': 'wokenpwd', 'db': 'woken', 'user': 'woken'})  PLAY RECAP \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  localhost : ok=2 changed=2 unreachable=0 failed=0  Initialise the databases...  Starting tests\_db\_1 ... done  2018/03/09 02:06:26 Waiting for: tcp://db:5432  2018/03/09 02:06:26 Connected to tcp://db:5432  Flyway 4.2.0 by Boxfuse  Database: jdbc:postgresql://db:5432/features (PostgreSQL 9.6)  Successfully validated 8 migrations (execution time 00:00.066s)  Creating Metadata table: "public"."schema\_version"  Current version of schema "public": << Empty Schema >>  Migrating schema "public" to version 1.0 - create  Migrating schema "public" to version 1.1 - churn  Migrating schema "public" to version 1.2 - iris  Migrating schema "public" to version 1.3 - dummy ldsm  Migrating schema "public" to version 1.4 - dummy federation  Migrating schema "public" to version 1.5 - synthetic datasets  Migrating schema "public" with repeatable migration Create view  Migrating schema "public" with repeatable migration Setup datasets linreg\_sample,churn,iris,desd\_synth,nida\_synth,qqni\_synth  Mar 09, 2018 2:06:27 AM eu.humanbrainproject.mip.migrations.R\_\_SetupValues migrate  INFO: Migrating dataset linreg\_sample...  Mar 09, 2018 2:06:27 AM eu.humanbrainproject.mip.migrations.R\_\_SetupValues migrate  INFO: Migrating dataset churn...  Mar 09, 2018 2:06:28 AM eu.humanbrainproject.mip.migrations.R\_\_SetupValues migrate  INFO: Migrating dataset iris...  Mar 09, 2018 2:06:28 AM eu.humanbrainproject.mip.migrations.R\_\_SetupValues migrate  INFO: Migrating dataset desd\_synth...  Mar 09, 2018 2:06:29 AM eu.humanbrainproject.mip.migrations.R\_\_SetupValues migrate  INFO: Migrating dataset nida\_synth...  Mar 09, 2018 2:06:30 AM eu.humanbrainproject.mip.migrations.R\_\_SetupValues migrate  INFO: Migrating dataset qqni\_synth...  Successfully applied 8 migrations to schema "public" (execution time 00:03.629s).  2018/03/09 02:06:30 Command finished successfully.  Starting tests\_db\_1 ... done  2018/03/09 02:06:32 Waiting for: tcp://db:5432  2018/03/09 02:06:32 Connected to tcp://db:5432  Flyway 4.2.0 by Boxfuse  Database: jdbc:postgresql://db:5432/woken (PostgreSQL 9.6)  Successfully validated 1 migration (execution time 00:00.008s)  Creating Metadata table: "public"."schema\_version"  Current version of schema "public": << Empty Schema >>  Migrating schema "public" to version 1.0 - create  Successfully applied 1 migration to schema "public" (execution time 00:00.079s).  2018/03/09 02:06:33 Command finished successfully.  Run the CLUS feature ranking algorithm for single-target regression...  Starting tests\_db\_1 ... done  Mar 09, 2018 2:06:34 AM eu.humanbrainproject.mip.algorithms.jsi.Main run  INFO: Starting experiment  Mar 09, 2018 2:06:35 AM com.github.fommil.netlib.ARPACK <clinit>  WARNING: Failed to load implementation from: com.github.fommil.netlib.NativeSystemARPACK  Mar 09, 2018 2:06:35 AM com.github.fommil.netlib.ARPACK <clinit>  WARNING: Failed to load implementation from: com.github.fommil.netlib.NativeRefARPACK  Clus v2.11 - Software for Predictive Clustering  Copyright (C) 2007, 2008, 2009, 2010  Katholieke Universiteit Leuven, Leuven, Belgium  Jozef Stefan Institute, Ljubljana, Slovenia  This program is free software and comes with ABSOLUTELY NO  WARRANTY. You are welcome to redistribute it under certain  conditions. Type 'clus -copying' for distribution details.  Clustering attributes check ==> #nominal: 0 #numeric: 1  For Feature Ranking RForest, OOB estimate of error should also be performed.  OOB Error Estimate is set to true.  Memory And Time Optimization = false  Out-Of-Bag Estimate of the error = true  Perform Feature Ranking = true  Ensemble Method: Random Forest  Feature importances written to: experimentTrees100.fimp  Output written to: experiment.out  Mar 09, 2018 2:06:38 AM eu.humanbrainproject.mip.algorithms.jsi.Main run  INFO: Reading feature importances  Run the CLUS feature ranking algorithm for multi-target regression...  Starting tests\_db\_1 ... done  Mar 09, 2018 2:06:40 AM eu.humanbrainproject.mip.algorithms.jsi.Main run  INFO: Starting experiment  Mar 09, 2018 2:06:40 AM com.github.fommil.netlib.ARPACK <clinit>  WARNING: Failed to load implementation from: com.github.fommil.netlib.NativeSystemARPACK  Mar 09, 2018 2:06:40 AM com.github.fommil.netlib.ARPACK <clinit>  WARNING: Failed to load implementation from: com.github.fommil.netlib.NativeRefARPACK  Clus v2.11 - Software for Predictive Clustering  Copyright (C) 2007, 2008, 2009, 2010  Katholieke Universiteit Leuven, Leuven, Belgium  Jozef Stefan Institute, Ljubljana, Slovenia  This program is free software and comes with ABSOLUTELY NO  WARRANTY. You are welcome to redistribute it under certain  conditions. Type 'clus -copying' for distribution details.  Clustering attributes check ==> #nominal: 0 #numeric: 4  For Feature Ranking RForest, OOB estimate of error should also be performed.  OOB Error Estimate is set to true.  Memory And Time Optimization = false  Out-Of-Bag Estimate of the error = true  Perform Feature Ranking = true  Ensemble Method: Random Forest  Feature importances written to: experimentTrees100.fimp  Output written to: experiment.out  Mar 09, 2018 2:06:44 AM eu.humanbrainproject.mip.algorithms.jsi.Main run  INFO: Reading feature importances  Stopping the containers...  Stopping tests\_db\_1 ... done  Removing tests\_clus\_fr\_regression\_mt\_run\_1 ... done  Removing tests\_clus\_fr\_regression\_st\_run\_1 ... done  Removing tests\_woken\_db\_setup\_run\_1 ... done  Removing tests\_sample\_data\_db\_setup\_run\_1 ... done  Removing tests\_create\_dbs\_run\_1 ... done  Removing tests\_wait\_dbs\_run\_1 ... done  Removing tests\_db\_1 ... done  Removing network tests\_default  Stopping the containers...  Removing network tests\_default  WARNING: Network tests\_default not found.  manager@bl8vbox[java-jsi-clus-fr] |