CODECHECK certificate 2020-001

Item	Value
Title	ShinyLearner: A containerized benchmarking tool for machine-learning classification of tabular data.
Authors	Terry J Lee; Erica Suh; Kimball Hill; Stephen R Piccolo
Reference	Paper to appear in Gigascience.
Codechecker	Stephen J. Eglen https://orcid.org/0000-0001-8607-8025
Date of check:	2019-02-14 10:00:00
Summary:	Only visualiation steps performed, rather than machine learning (which could take several hours/days). The created figures match those in the article. The content of other output files was not checked.
Repository:	https://github.com/codecheckers/Piccolo-2020

Table 1: CODECHECK summary

File	Comment	Size
Figures/Datasets_Basic_AUROC.pdf	Figure 2 of manuscript	8078
Figures/Predictions_Histograms.pdf	Figure 3 of manuscript	8727
Figures/Algorithms_ParamsImprovement_AUROC.pdf	Figure 4 of manuscript	7837
Figures/Algorithms_FSImprovement_AUROC.pdf	Figure 5 of manuscript	8190
Figures/FS_vs_CL.pdf	Figure 6 of manuscript	6521
Figures/FS_NumFeatures.pdf	Figure 7 of manuscript	5810
Tables/Basic_DiffFromMedian.tsv	Example output table 1 (not in manuscript)	619
Tables/ParamOpt_Improvement.tsv	Example output table 2 (not in manuscript)	14216

Table 2: Summary of output files generated

Summary

The reproduction of the figures in the manuscript was straightforward given that the authors provided a Rmarkdown document that processed the results data files. The results data files were not independently reproduced at this stage because of the long compute time.

CODECHECKER notes

Data and Code

After creating an empty repository, the following data and results folders were copied from CODE OCEAN capsule (https://codeocean.com/capsule/5449763/tree) for the paper: `Datasets`, `Results_Basic`, `Results_FeatureSelection` and `Results_ParamsOptimized`. One of the data files `Results_ParamsOptimized/diabetes/Nested_Predictions.tsv.gz` needed to be compressed to reduce the file size from 160 Mb to 9 Mb (Github has limit on individual files of 100 Mb.)

The main Rmd and helper script were downloaded from https://github.com/srp33/ShinyLearner/. The Rmd file was edited to read the compressed version of the .tsv file.

As no MANIFEST was provided by the author, SJE made one.

Extra software installations

To run this script, I needed to install the following R packages

```
install.packages("tidyverse")
install.packages("cowplot")
```

Running the software to regenerate outputs.

A few small edits were made to the `Analyze_Results.Rmd`, see the git history,

To regenerate the Figures/ folder, we simply ran `make run` in the Makefile, created by SJE.

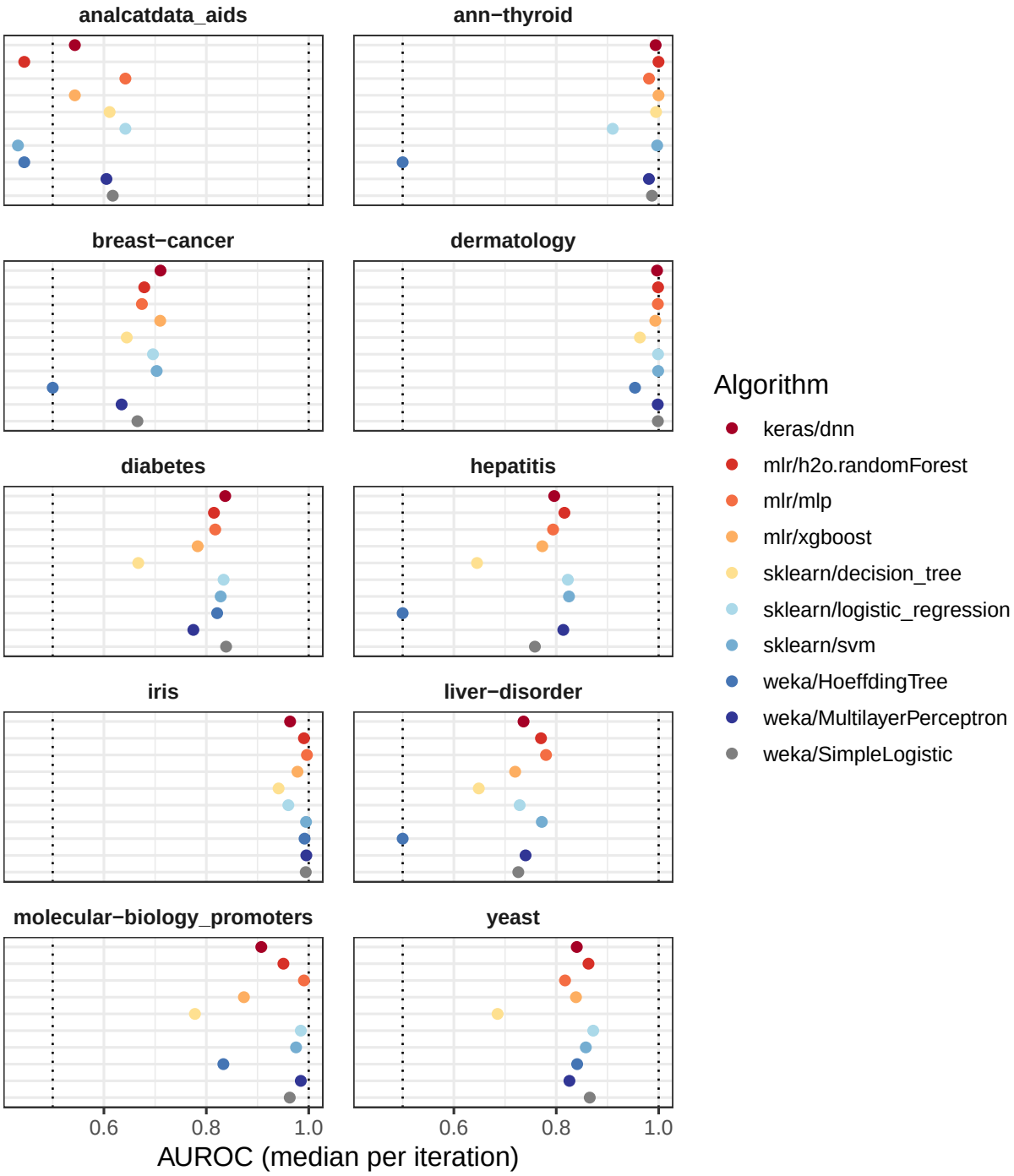

Figure 1: Figure 2 of manuscript

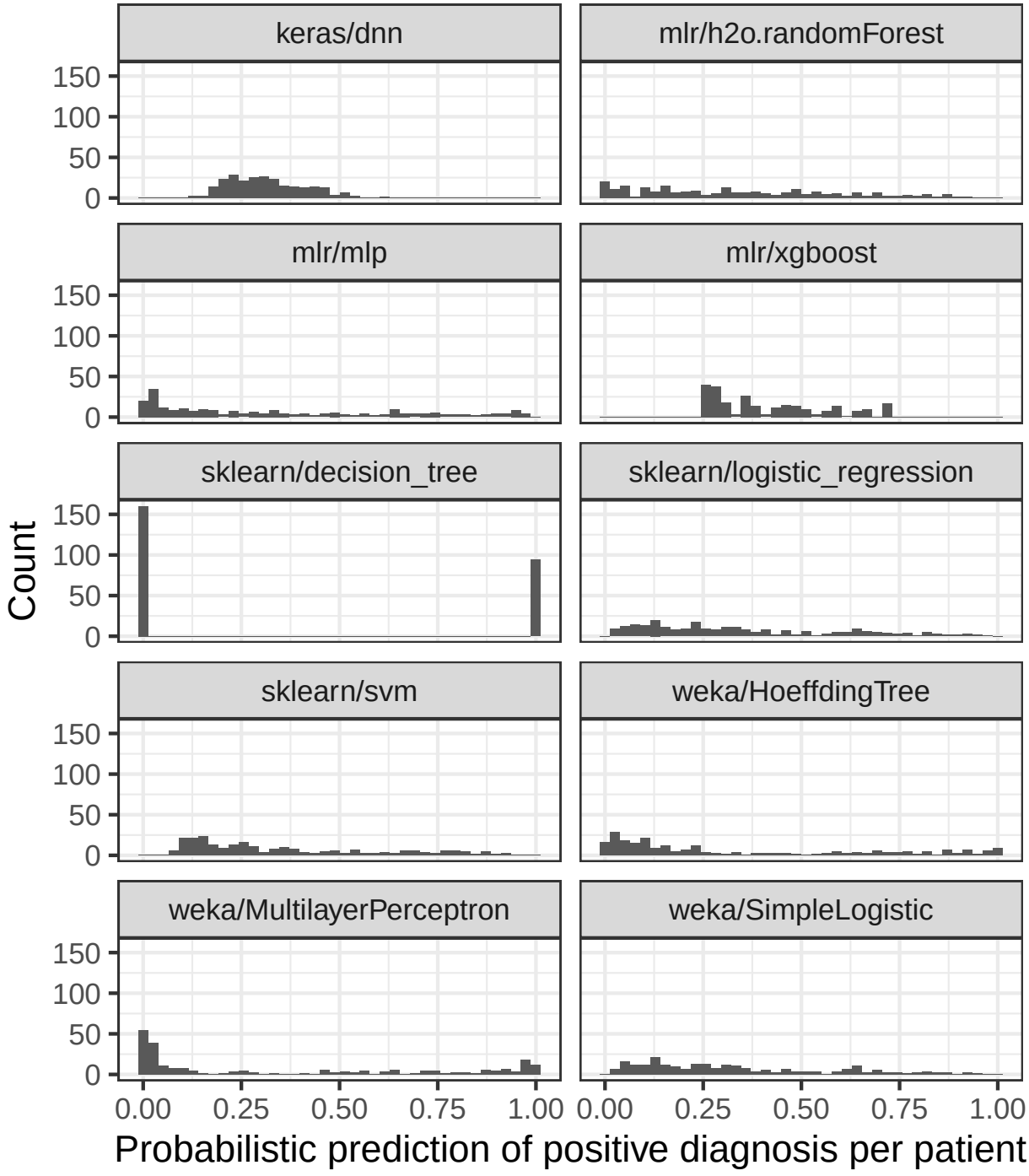

Figure 2: Figure 3 of manuscript

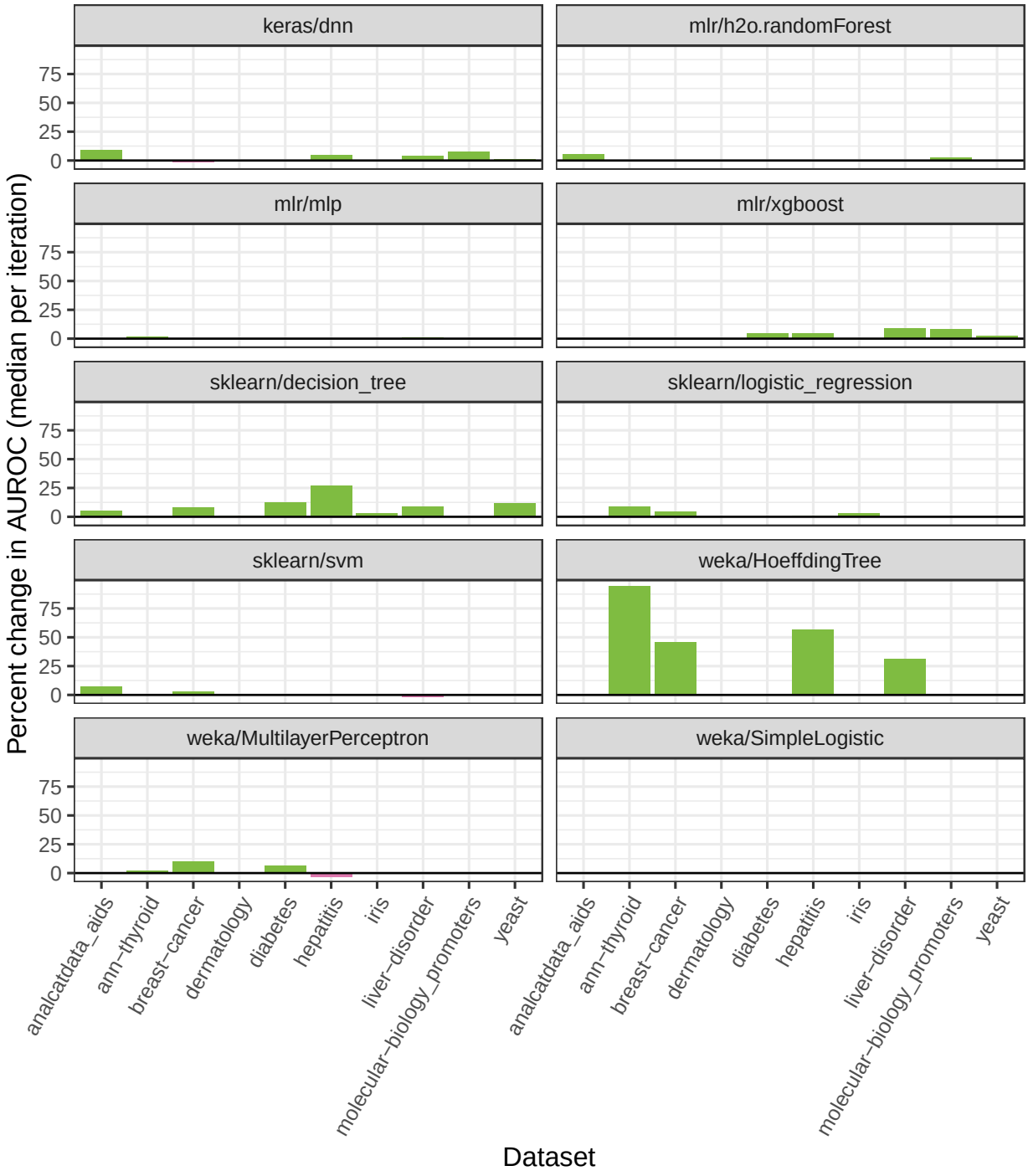

Figure 3: Figure 4 of manuscript

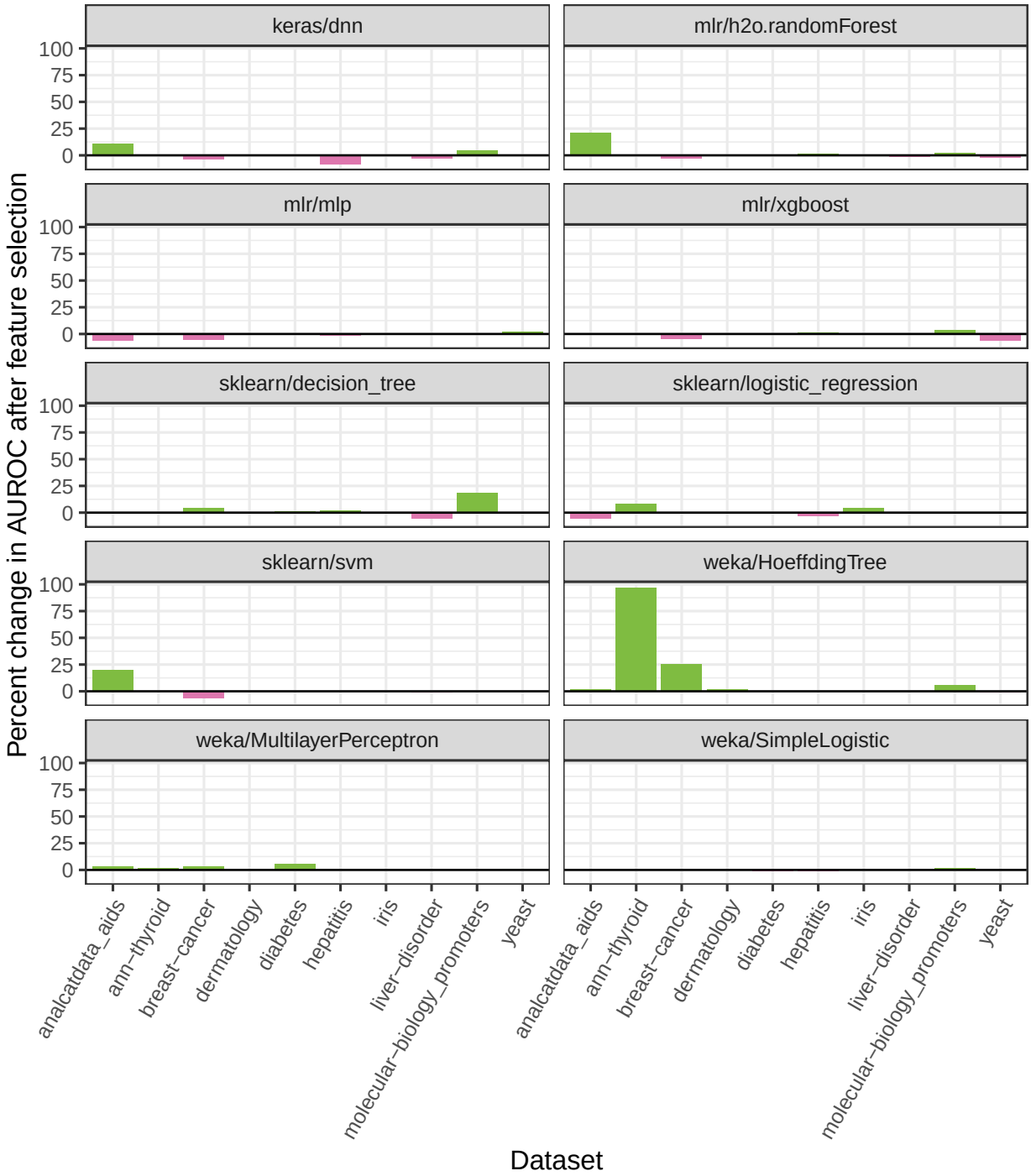

Figure 4: Figure 5 of manuscript

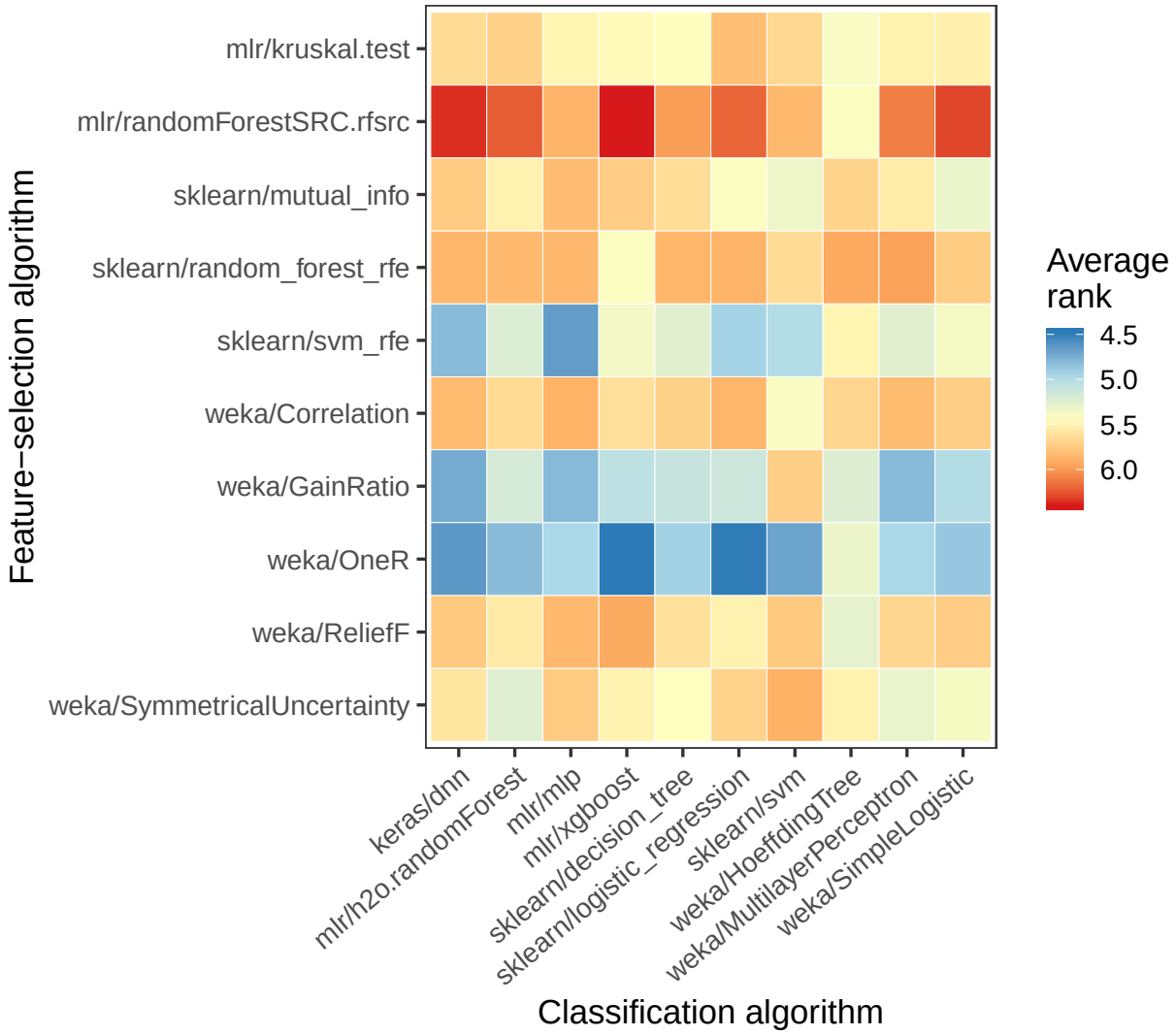

Figure 5: Figure 6 of manuscript

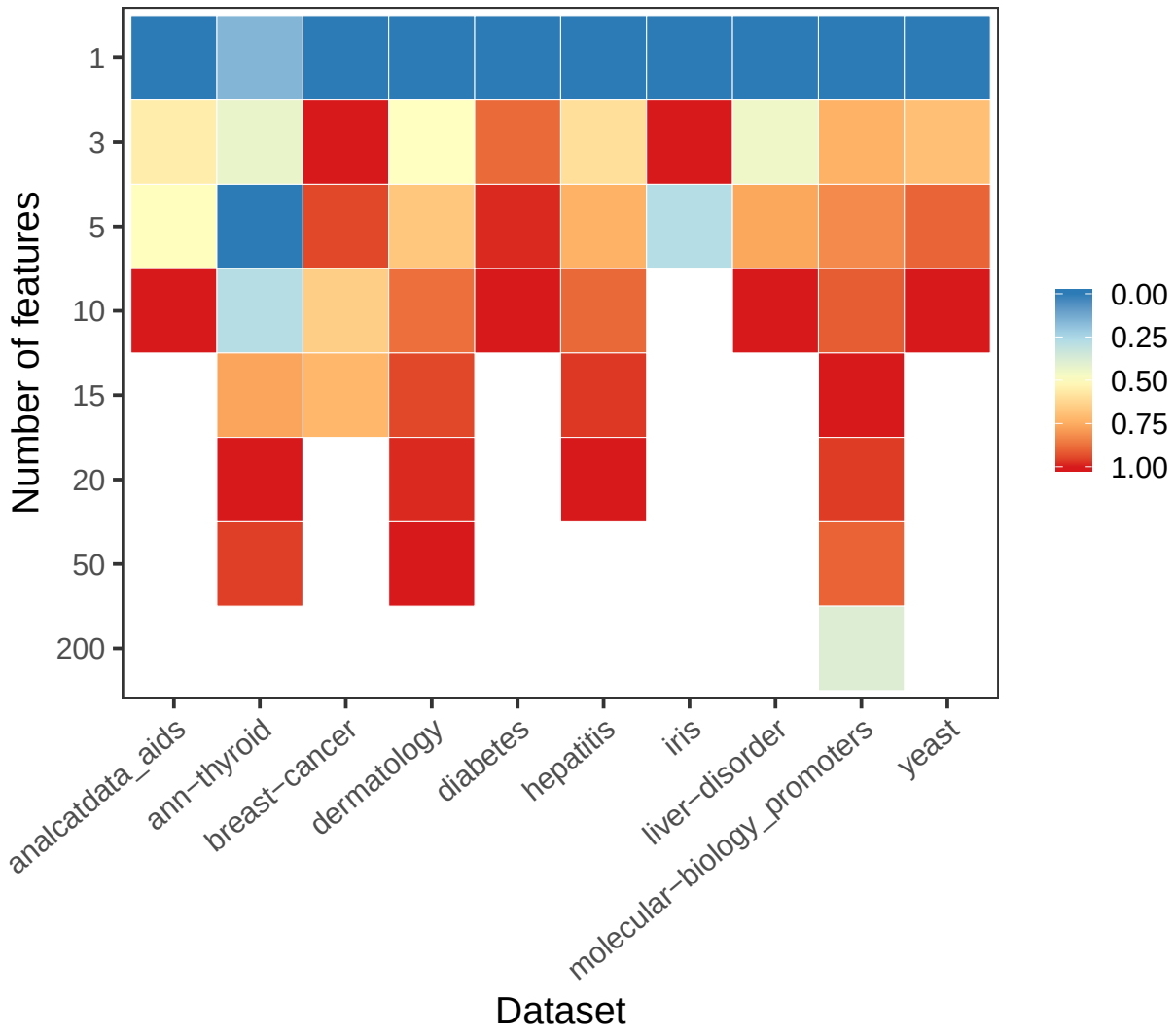

Figure 6: Figure 7 of manuscript

Table 1

```
read_tsv(dest_files[7])
```

```
## Parsed with column specification:
## cols(
##   Algorithm = col_character(),
##   Min_Diff = col_double(),
##   Max_Diff = col_double()
## )
```

```
## # A tibble: 10 x 3
##    Algorithm                   Min_Diff Max_Diff
##    <chr>                          <dbl>    <dbl>
##  1 keras/dnn                    -0.0494  0.0340
##  2 mlr/h2o.randomForest         -0.130   0.0379
##  3 mlr/mlp                      -0.0233  0.0679
##  4 mlr/xgboost                  -0.0833  0.0335
##  5 sklearn/decision_tree        -0.179   0.0370
##  6 sklearn/logistic_regression  -0.0803  0.0679
##  7 sklearn/svm                  -0.142   0.0395
##  8 weka/HoeffdingTree           -0.491   0.00187
##  9 weka/MultilayerPerceptron    -0.0446  0.0309
## 10 weka/SimpleLogistic          -0.0366  0.0432
```

Table 2

```
print(read_tsv(dest_files[8]))
```

```
## Parsed with column specification:
## cols(
##   Params = col_character(),
##   analcatdata_aids = col_double(),
##   `ann-thyroid` = col_double(),
##   `breast-cancer` = col_double(),
##   dermatology = col_double(),
##   diabetes = col_double(),
##   hepatitis = col_double(),
##   iris = col_double(),
##   `liver-disorder` = col_double(),
##   `molecular-biology_promoters` = col_double(),
##   yeast = col_double()
## )
```

```
## # A tibble: 53 x 11
##    Params analcatdata_aids `ann-thyroid` `breast-cancer`
##    <chr>             <dbl>         <dbl>           <dbl>
##  1 dropo~          -0.0833      0.00266         -0.00819
##  2 dropo~           0.0833     -0.000223        -0.0251
##  3 dropo~          -0.0278     -0.00301          0.0158
##  4 dropo~          -0.0278      0.00221         -0.00234
##  5 dropo~          -0.139      -0.00152         -0.0327
##  6 dropo~          -0.0278      0.00259         -0.00994
##  7 dropo~          -0.0278     -0.00405         -0.0655
##  8 dropo~          -0.0278     -0.00734         -0.0965
##  9 dropo~           0          -0.00800         -0.0550
## 10 dropo~          -0.0833     -0.0112          -0.0690
## # ... with 43 more rows, and 7 more variables:
## #   dermatology <dbl>, diabetes <dbl>, hepatitis <dbl>,
## #   iris <dbl>, `liver-disorder` <dbl>,
## #   `molecular-biology_promoters` <dbl>, yeast <dbl>
```

About this document

This document was created using Rmarkdown. `make codecheck.pdf` will regenerate the file.

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Manjaro Linux
##
## Matrix products: default
## BLAS:   /usr/lib/libopenblasp-r0.3.7.so
## LAPACK: /usr/lib/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## other attached packages:
## [1] rprojroot_1.3-2 readr_1.3.1     tibble_2.1.3
## [4] yaml_2.2.0      xtable_1.8-3    knitr_1.26
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       magrittr_1.5     hms_0.4.2
##  [4] R6_2.4.1         rlang_0.4.2      fansi_0.4.0
##  [7] stringr_1.4.0    highr_0.8        tools_3.6.2
## [10] xfun_0.11        utf8_1.1.4       cli_1.1.0
## [13] htmltools_0.4.0  assertthat_0.2.1 digest_0.6.23
## [16] crayon_1.3.4     vctrs_0.1.0      zeallot_0.1.0
## [19] evaluate_0.14    rmarkdown_1.18   stringi_1.4.3
## [22] compiler_3.6.2   pillar_1.4.1     backports_1.1.4
## [25] pkgconfig_2.0.2
```