Genome Genies

Bilal A., Thomas C., Alex L.

PharmaHacks 2024 - Genomics Challenge

Our High-Level Plan

Step 1: Extract relevant data from .rds file in RStudio

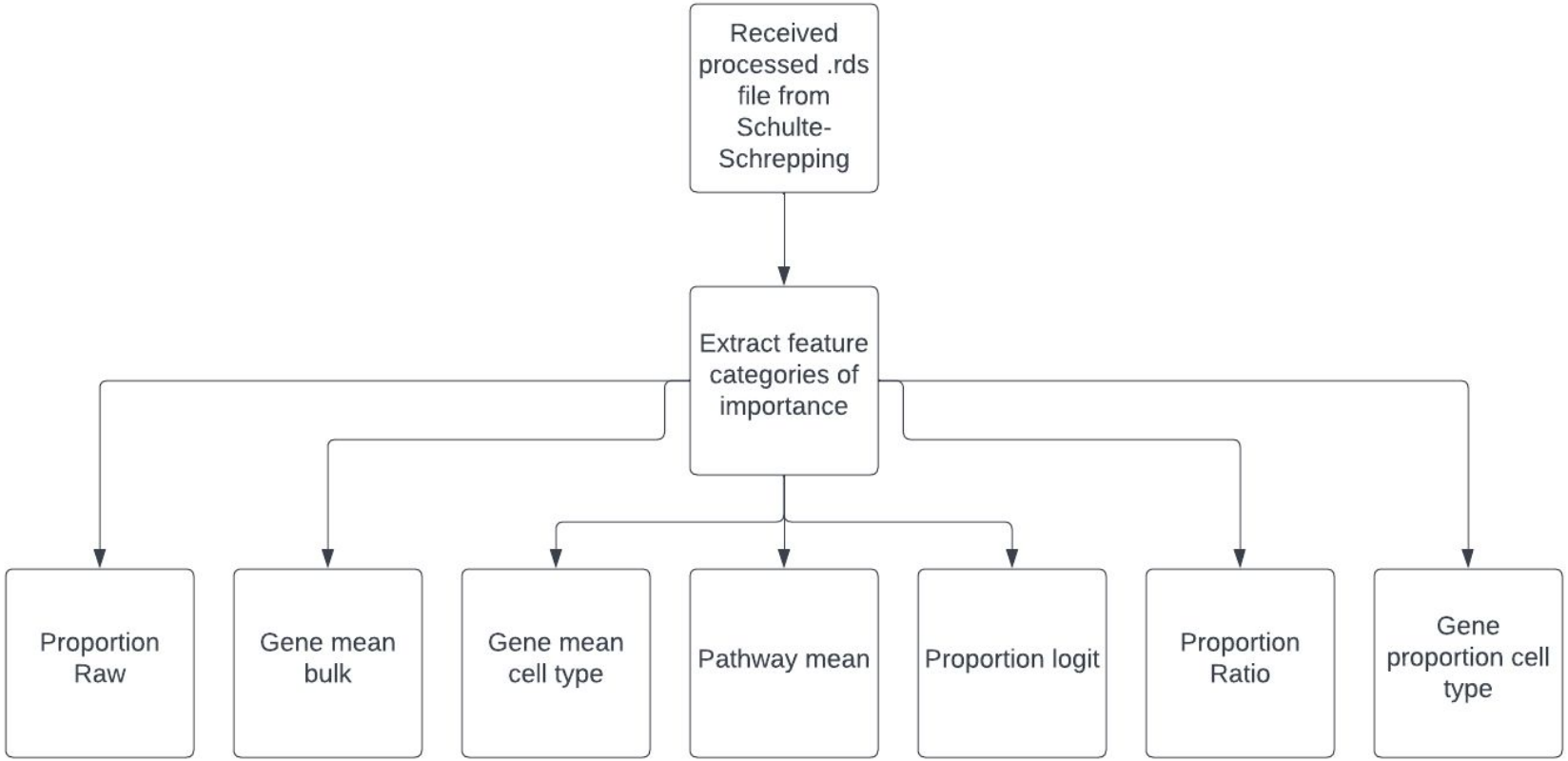
Step 2: Further process and analyze with R & Python (Pandas, SKLearn)

Step 3: Use findings and feed into a KNN

The Goal:

Train a model that accurately predicts COVID-19 patient outcomes from scRNA-seq data.

Extracting Relevant Data


Received processed .rds file from Schulte-Schrepping
Extract feature categories of importance
Proportion Raw
Gene mean bulk
Gene mean cell type
Pathway mean
Proportion logit
Proportion Ratio
Gene proportion cell type

Pre-Processing

- The Schulte-Schrepping scRNA-seq dataset had already gone through normalization and dimension reduction.
- No further modification of the dataset was necessary.

```
An object of class Seurat
46611 features across 99049 samples within 7 assays
Active assay: RNA (46584 features, 2000 variable features)
 3 layers present: counts, data, scale.data
 6 other assays present: HTOsPool1, HTOsPool2, HTOsPool3, HTOsPool4, HTOsPool5, HTOsPool6
 3 dimensional reductions calculated: pca, umap, harmony
```

Extracting Features

- Individual (patient-level) features were pulled from the dataset as-is using the “scFeatures” function in the “scFeatures” package.
- All necessary data and metadata was present in the Seurat object.

```
# Load features ["proportion_ratio", "proportion_logit", "pathway_gsva",
# "pathway_mean", "gene_mean_aggregated", "gene_mean_celltype"]
data <- rna_data@assays$RNA@data
celltype <- rna_data$id.celltype
sample <- rna_data$sampleID
scFeatures <- scFeatures(data, celltype = celltype, sample = sample, type = "scrna",
                         feature_types = c("proportion_ratio", "proportion_logit", "pathway_gsva",
                                           "pathway_mean", "gene_mean_aggregated", "gene_mean_celltype"))
```

Exporting Feature Data

- Individual feature data was generated as a list of dataframes, one dataframe per individual feature.
- At this point, there was nothing more to do in R, and individual feature data was exported in csv format for further manipulation in Python.

```
19  # Write extracted individual features to csv for further analysis in python.
20  write.csv(scFeatures$proportion_ratio, "/Volumes/EXTERNAL/proportion_ratio.csv", row.names=TRUE, col.names=TRUE)
21  write.csv(scFeatures$proportion_logit, "/Volumes/EXTERNAL/proportion_logit.csv", row.names=TRUE, col.names=TRUE)
22  write.csv(scFeatures$pathway_mean, "/Volumes/EXTERNAL/pathway_mean.csv", row.names=TRUE, col.names=TRUE)
23  write.csv(scFeatures$gene_mean_bulk, "/Volumes/EXTERNAL/gene_mean_bulk.csv", row.names=TRUE, col.names=TRUE)
24  write.csv(scFeatures$gene_mean_celltype, "/Volumes/EXTERNAL/gene_mean_celltype.csv", row.names=TRUE, col.names=TRUE)
```

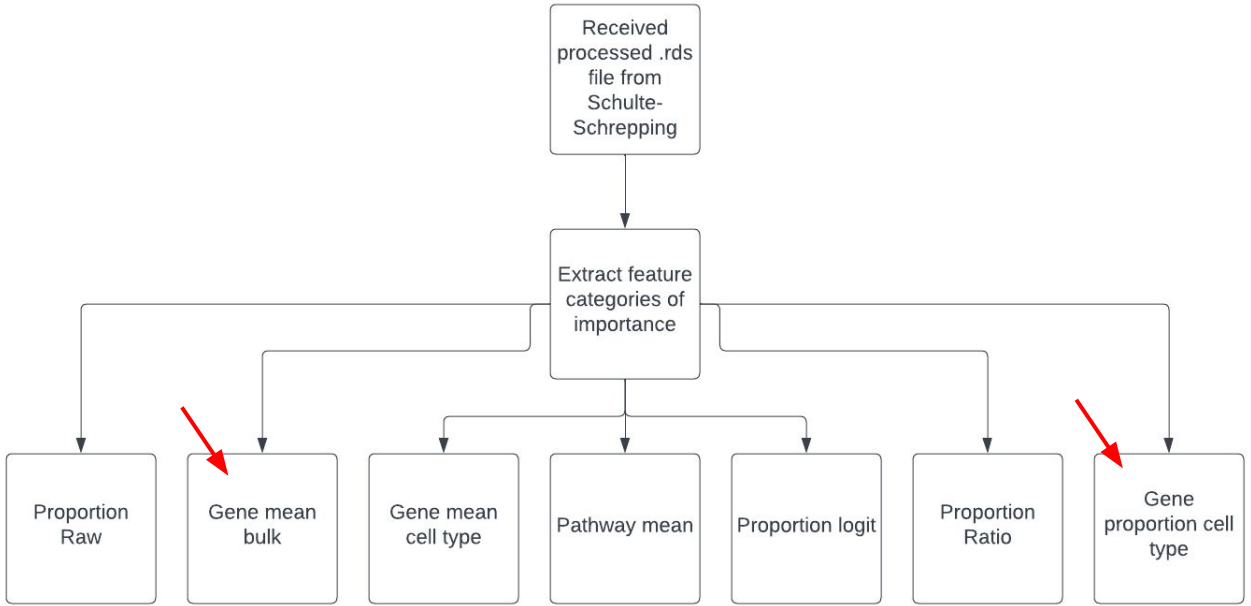
Exporting Patient Labels

- The Seurat object metadata contained unique patient identification codes and patient outcome linked to each sample.
- Again, we used R to extract these data frames to csv format for use in model training.
- Patient IDs and their outcomes were compiled into a unique set in python.

```
31  # Write sampleIDs matched to patient names and outcomes to extract patient ID -> outcome in python.
32  patient_outcomes <- rna_data$group_per_sample
33  patient_names <- rna_data$sampleID
34  write.csv(patient_outcomes, "./patient_outcomes.csv", row.names=TRUE, col.names=TRUE)
35  write.csv(patient_names, "/Volumes/EXTERNAL/patient_names.csv", row.names=TRUE, col.names=TRUE)
```

Further Processing with Python

Narrowing of Gene Scope



Narrowing of Gene Scope

Certain cellular biomarkers are common among patients with severe cases of COVID-19:

- Elevated leukocyte and neutrophil counts;
- Suppressed lymphocyte count (Huang C. et al., 2020; Qin et al., 2020)

Additionally, in a meta-analysis of nine studies including 1779 patients, 399 of them with severe disease, low platelet count was significantly associated with disease severity. (Lippi et al., 2020)

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395 (10223), 497–506. doi: 10.1016/s0140-6736(20)30183-5

Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., et al. (2020). Dysregulation of immune response in patients with COVID-19 in Wuhan, China. Clin. Infect. Dis. doi: 10.1093/cid/ciaa248

Lippi, G., Plebani, M., & Henry, B. M. (2020). Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis. Clinica Chimica Acta, 506, 145–148. https://doi.org/10.1016/j.cca.2020.03.022

Platelet Count as a Risk-Factor for Disease

Based on this information, looked for studies that explored genes associated with platelet counts:

A meta-analysis of platelet count data from Genome-Wide Association Studies (GWAS) done on 536,974 Europeans:

- 577 gene SNPs were found to impact platelet counts. (Mikaelsdottir, E. et al., 2021)

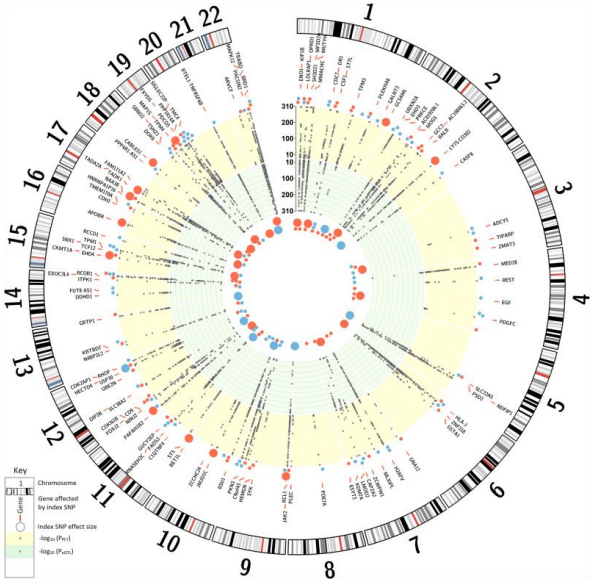


Image Source: Retrieved from study linked below.

Mikaelsdottir, E., Thorleifsson, G., Stefansdottir, L., Halldorsson, G., Sigurdsson, J. K., Lund, S. H., Tragante, V., Melsted, P., Rognvaldsson, S., Norland, K., Helgadottir, A., Magnusson, M. K., Ragnarsson, G. B., Kristinsson, S. Y., Reykdal, S., Vidarsson, B., Gudmundsdottir, I. J., Olafsson, I., Onundarson, P. T., ... Stefansson, K. (2021). Genetic variants associated with platelet count are predictive of human disease and physiological markers. Communications Biology, 4(1). https://doi.org/10.1038/s42003-021-02642-9

Genetic Interactions Linked to COVID-19 Severity

Ensemble models trained using Whole Exome Sequencing (WES) data from a cohort of 2000 Italian patients:

- several key genes that increase the risk of COVID-19 severity were identified. (Onoja, A. et al., 2022)

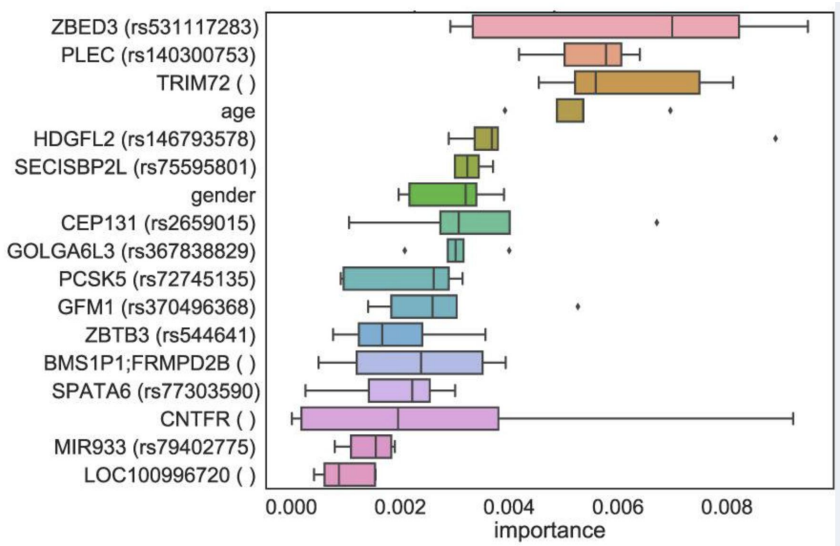


Image Source: Retrieved from study linked below.

Onoja, A., Picchiotti, N., Fallerini, C., Baldassarri, M., Fava, F., Mari, F., Daga, S., Benetti, E., Bruttini, M., Palmieri, M., Croci, S., Amitrano, S., Meloni, I., Frullanti, E., Doddato, G., Lista, M., Beligni, G., Valentino, F., Zguro, K., ... Raimondi, F. (2022). An explainable model of host genetic interactions linked to COVID-19 severity. Communications Biology, 5(1). https://doi.org/10.1038/s42003-022-04073-6

```
df_common_genes = df_gene_mean_bulk.loc[:, common_genes]
features = df_common_genes
labels = df_labels["Outcome"]

X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.2, random_state=42)

knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
scores = cross_val_score(knn, features, labels, cv=13)
print("Mean cross-validation score:", scores.mean())
print(classification_report(y_test, y_pred))
```

[99] ✓ 0.2s

```
Accuracy: 0.8
Mean cross-validation score: 0.9358974358974359
              precision    recall  f1-score   support

     control       1.00      1.00      1.00         5
        mild       0.00      0.00      0.00         2
      severe       0.60      1.00      0.75         3

    accuracy                           0.80        10
   macro avg       0.53      0.67      0.58        10
weighted avg       0.68      0.80      0.72        10
```

Limiting the number of considered genes in the data used to train our models let to a significant increase in accuracy

70% (~3500 genes) → 80% (136 genes)

Further Processing with Python

Goals in mind:

- Figure out which feature category was the most relevant
- Rank the features within the feature categories in order of statistical relevance

Significance of features (Gene mean cell type)

We assumed that the gene mean cell feature category would produce the most significant and interpretable results.

As mentioned earlier, elevated leukocyte and neutrophil counts are a key cellular biomarker in patients with severe COVID-19. We:

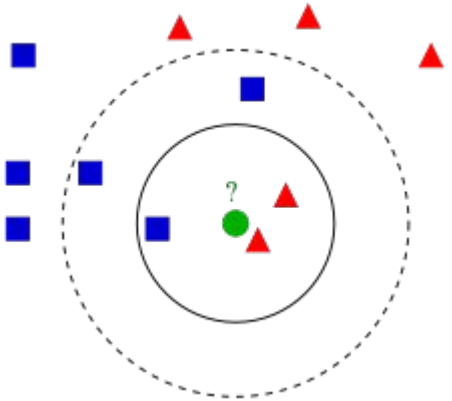
- Ranked features based on chi-squared test.
- Printed the 10 best features based on their chi-squared values
- Also printed cross validation (CV) score

```
Mean CV Score: 89.00%
Standard Deviation of CV Scores: 18.14%
[0.89]
```

	Feature	Score
15103	6: Immature Neutrophils--LTF	97.567372
15200	6: Immature Neutrophils--S100A8	94.453252
15112	6: Immature Neutrophils--RETN	87.539055
15107	6: Immature Neutrophils--LCN2	86.658265
15098	6: Immature Neutrophils--DEFA3	75.656955
15201	6: Immature Neutrophils--S100A9	71.586954
15426	6: Immature Neutrophils--ATP5F1E	69.198740
15214	6: Immature Neutrophils--S100P	65.769064
15111	6: Immature Neutrophils--MMP8	64.722158
15100	6: Immature Neutrophils--CAMP	61.515777

Making Predictions

ML Model



K-Nearest-Neighbours

- Compares new pieces of data with existing data to classify the data point (with euclidean distance)

Advantages:

- Simplicity and intuitiveness
- No assumptions of data distribution
- Effectiveness in small datasets

Cons:

- Sensitivity to irrelevant or redundant features

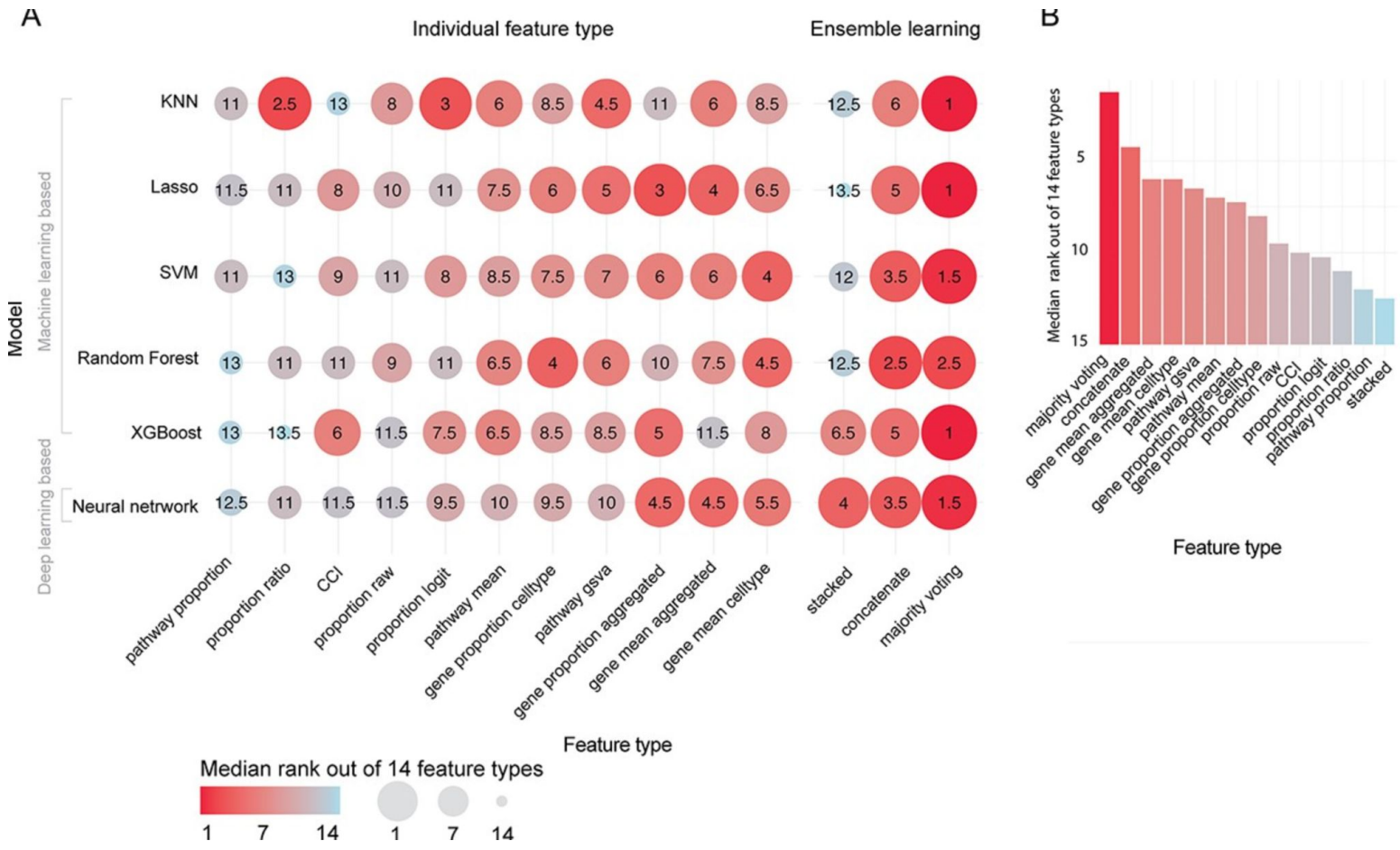
Evaluation metric

K-Fold Cross Validation

- Chosen metric because of its robustness, specifically for smaller data sets (we only had access to 50 patients)
- Chosen K=13 because some of the members only have 13 populated cells, so 13 was the maximum without having splits without encountering issues

Results

```
gene_mean_bulk Accuracy: 0.7
Mean cross-validation score: 0.8333333333333334
---------------------------------
gene_mean_celltype Accuracy: 0.9
Mean cross-validation score: 0.846153846153846
---------------------------------
pathway_mean Accuracy: 0.9
Mean cross-validation score: 0.7884615384615383
---------------------------------
proportion_logit Accuracy: 0.9
Mean cross-validation score: 0.9423076923076923
---------------------------------
proportion_ratio Accuracy: 0.9
Mean cross-validation score: 0.9358974358974359
---------------------------------
gene_proportion_celltype Accuracy: 0.9
Mean cross-validation score: 0.891025641025641
---------------------------------
proportion_raw Accuracy: 0.8
Mean cross-validation score: 0.8717948717948717
---------------------------------
Majority Voting Accuracy: 0.8
```


A
Individual feature type
Ensemble learning
Model
Machine learning based
Deep learning based
KNN 11 2.5 13 8 3 6 8.5 4.5 11 6 8.5 12.5 6 1
Lasso 11.5 11 8 10 11 7.5 6 5 3 4 6.5 13.5 5 1
SVM 11 13 9 11 8 8.5 7.5 7 6 6 4 12 3.5 1.5
Random Forest 13 11 11 9 11 6.5 4 6 10 7.5 4.5 12.5 2.5 2.5
XGBoost 13 13.5 6 11.5 7.5 6.5 8.5 8.5 5 11.5 8 6.5 5 1
Neural netrwork 12.5 11 11.5 11.5 9.5 10 9.5 10 4.5 4.5 5.5 4 3.5 1.5
pathway proportion
proportion ratio
CCI
proportion raw
proportion logit
pathway mean
gene proportion celltype
pathway gsva
gene proportion aggregated
gene mean aggregated
gene mean celltype
stacked
concatenate
majority voting
Feature type
Median rank out of 14 feature types
1 7 14 1 7 14
B
Median rank out of 14 feature types
5
10
15
majority voting
concatenate
gene mean aggregated
gene mean celltype
pathway gsva
pathway mean
gene proportion aggregated
gene proportion celltype
proportion raw
CCI
proportion logit
proportion ratio
pathway proportion
stacked
Feature type

Conclusions

What we *might* have done differently next time

- Focus more on data processing as opposed to plugging into models straight away
- Create a clearer picture of goals and what we hoped to achieve
- Take more time to explore the data manually to get a feel for it
- Brush up on R....
- Played around more with hyperparameters