### Model Design

Using Caret package for developing a classifier in R. The model selected was Random Forrest since it yielded functional results, even though it is non-ideal to have limited transparency. The model predicts future recurrence of bladder cancer, with the outcome variable being a binary classification between “Recurrence likely” and “Recurrence unlikely”.
Github link to model: https://github.com/GiulioSP/520hw4giuli

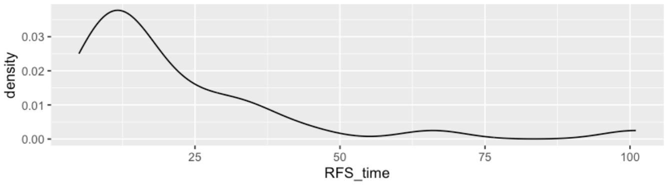
### Data cleaning and Feature Selection

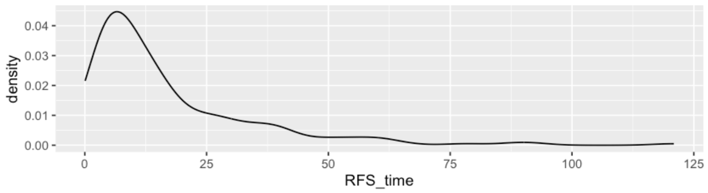
- Removing columns not in common between both datasets
  - In Uromol: UROMOL.ID, Smoking, Tumor.size, Incident.tumor, EAU.risk
  - In Knowles: knowles_ID
- Removing columns with the same information in all rows: Tumor.stage and Tumor.grade
- Removing PFS_time due its many missing entries in Knowles dataset, as there is no progression in the knowles dataset.
- Knowles dataset has “NA” at RFS_time data for cases without recurrence, so they were replaced with “0” values.

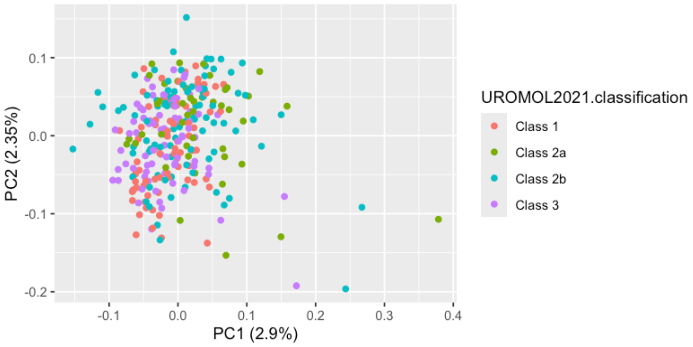
### Expression Data Feature Selection

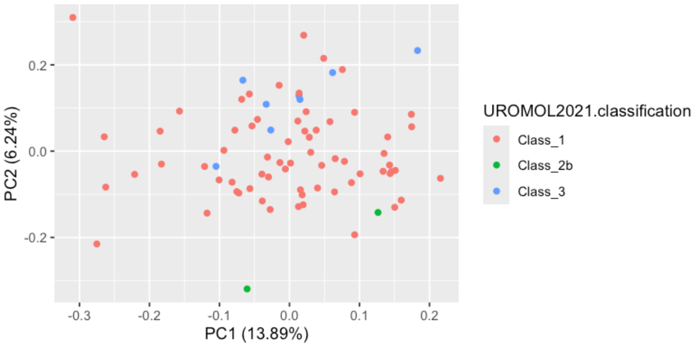
- Removed genes with near zero variance (~4k genes) using caret::nearZeroVar()
- Removed genes not present in both datasets (~10k genes)
- Custom correlation based filtering:
  - Generated a correlation matrix of all remaining genes with cor().
  - Summed the absolute values across columns to evaluate the overall correlation between genes.
  - Finally, selected the 500 genes with the smallest correlation sum values as they should be the most informative subset.

### Training Data Visualizations


Density of knowles RFS time


Density of uromol RFS time


Uromol PCA labelled by UROMOL2021 classification

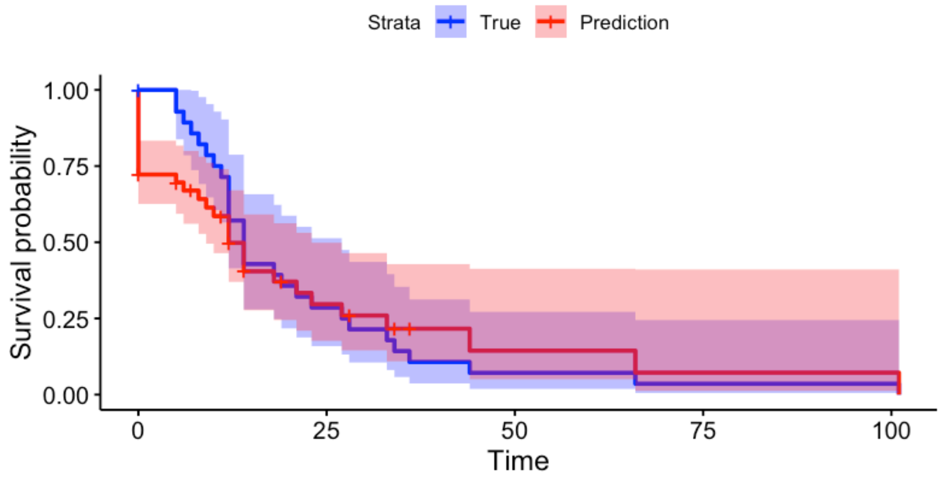

Knowles PCA labelled by UROMOL2021 classification

**Performance Metrics**

Confusion matrix:

| | | True values | |
|---|---|---|---|
| | | No Recurrence | Recurrence |
| Predicted values | No Recurrence | 24 | 9 |
| | Recurrence | 20 | 20 |

Kaplan-Meier (survival) plot:



Note that predicted false positives do not have equivalent RFS_time data in Knowles dataset, so they appear as a large dip at time 0 in the prediction plot (in red).

## Clinical Application

This model has some clinical applicability to evaluate which patients are likely to display recurrence, and therefore require more aggressive treatment. The model has better sensitivity than specificity, which is appropriate for avoiding undertreatment. Future iterations could be more useful if they classified cases by 3 categories: “long RFS time”, “short RFS time” and “Recurrence unlikely”.

## References

- Kuhn, Max (2008). “Building Predictive Models in R Using the caret Package.” *Journal of Statistical Software*, 28(5), 1–26. doi:10.18637/jss.v028.i05, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.
- Lindskrog, S.V., Prip, F., Lamy, P. *et al.* An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat Commun* 12, 2301 (2021). https://doi.org/10.1038/s41467-021-22465-w