Joosep Lember, Norman Tolmats, Marten Mathias Jaani
Institute of Computer Science, University of Tartu

Cancer Survival Prediction with Synthetic Data

Goals and mission

- **Objective:** Predict the likelihood of a patient's mortality based on medical data.
- **Workflow Design:** Structured to identify crucial features/create new ones for an accurate classifier.
- **Initial Implementation:** Applied the workflow using synthetic and eventually real data for lung cancer.
- **End result:** Develop a workflow adaptable for predicting outcomes in various other diseases.

Approach

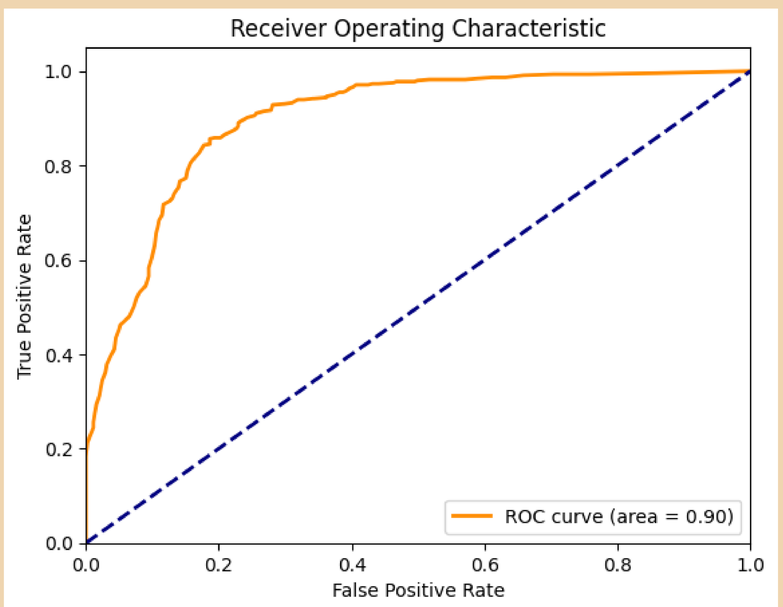
Workflow 1

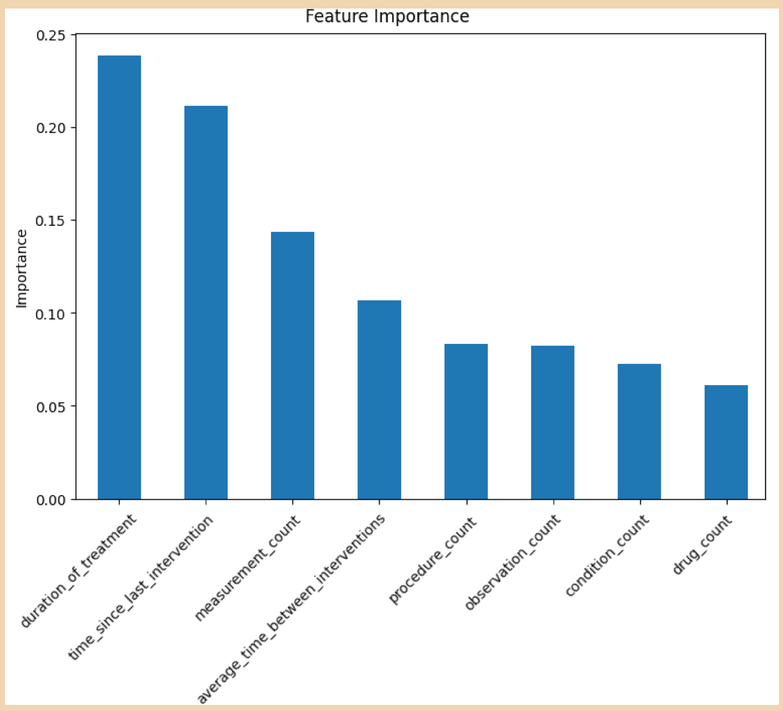
- Create new time features and change dataset's provided definitions for more general ones.
- Random forest classifier on modified data.

Workflow 2

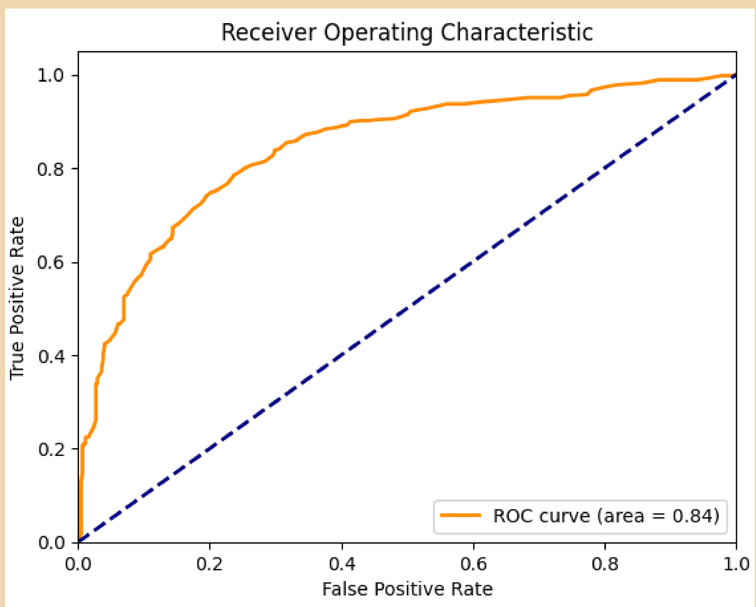
- Graph-based method to identify the most crucial features.
- Random forest with only the key features.

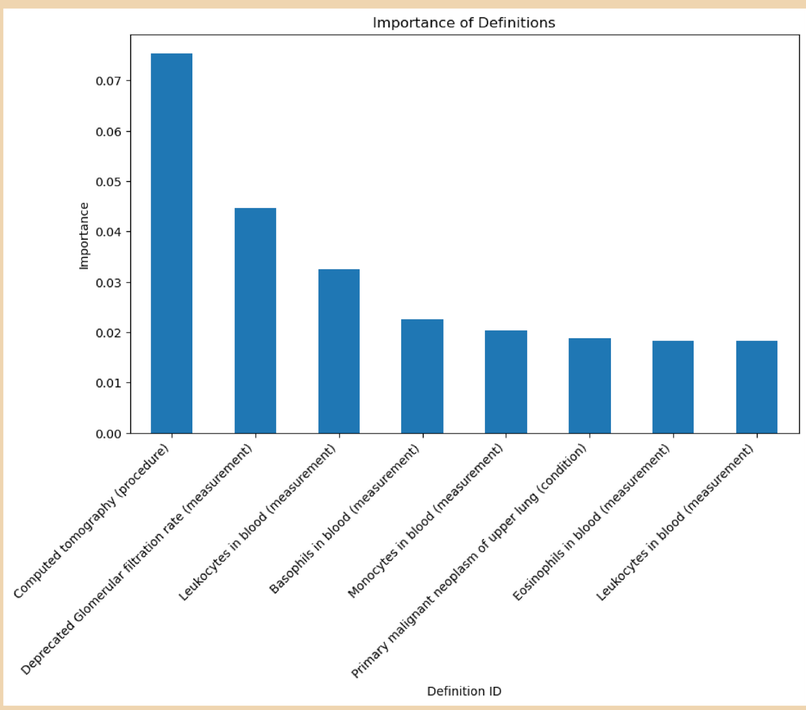
Workflow 1 Graphs





Workflow 2 Graphs





Data

- Synthetic data for development, real data for final results.
- Three columns: SUBJECT_ID, DEFINITION_ID, TIME
- SUBJECT_ID is a unique identifier for patients.
- DEFINITION_ID categorizes the type of medical intervention (drug, measurement, condition, procedure etc).
- TIME represents the time in years since diagnosis at which the intervention occurred.

Conclusion

- Different ways to modify data to train models
- **First workflow:** modified definitions for count-based features and new time based features. Also a predictive function for stakeholder use.
- **Second workflow:** used graphs to identify the key features.
- Best results with a random forest classifier.
- Hopefully these workflows can be useful with other diseases as well.