Project Overview

**The data set is explored through below steps:**

1. Data frame building and cleaning
2. Feature engineering
3. Modeling building
4. Model selecting
5. Feature importance analysis
6. Data frame building and cleaning:

* The data frame was joined based on the sample ID of mrna\_data, seq\_data and patient\_data after basic string cleaning. Redundant features in each data set were removed from the initial cleaning based on their distribution and explanatory characters.
* The data frame is attached in the folder and named as df.csv. It’s a data frame with 531 observations and 195 features.
* Tested kNN, EM and removing strategies on missing value imputation. The kNN strategy is selected based on the performance of value estimation.

1. Feature engineering:

* Categorical data 'American Joint Committee on Cancer Tumor Stage Code' was recategorized as four main categories: ‘T1’, ‘T2’, ‘T3’, ‘T4’ to boost model performance and enhance feature significance.
* Checked the distribution of each feature and removed features with extreme imbalanced distribution or strong correlation.
* The features correlation matrix indicates that numeric features mostly have low correlations.

1. Modeling building:

* Applied the most fundamental classification models: Logistic Regression, kNN, Random Forest and Gradient Boosting Decision Tree to test model performance on testing set through patients/mrna/sequence features against patient’s 'Overall Survival Status\_DECEASED'.
* Conducted a 10-folds cross validation to double verify models’ bias and variance.

1. Model selecting:

* Tuned each model’s hyper parameters based on 10-folds validation performance. Logistic Regression and Gradient Boosting Decision Tree generally have the best performance in terms of models’ precision and recall rate.
* Applied the parameter-tuned model on the testing set and checked the confusion matrix and ROC curve.
* Worked on the trends between threshold and precision/recall to decide the best performing threshold of maximizing precision/recall. In this case, the model is designed to weight more on recall rate since detecting all Positive Deceased is more important to our findings.

1. Feature importance analysis:

* Conducted feature importance analysis via L1-regularized Logistic Regression coefficient and Random Forest feature importance.
* Find the common important features and adjust the input data to re-test models.

**Model Performance:**

After the parameter tuning, here are the scores of each model:

A picture containing screenshot

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The best model is Logistic Regression under this training and testing case.

A close up of a map

Description automatically generated

A screenshot of a map

Description automatically generated

From the graphs above, we optimize the threshold to get a recall-weighted model based on the recall curve derivate. The optimized threshold is around 0.3.

We take the deceased\_rate as the threshold. The performance of the model is:

A screenshot of a cell phone

Description automatically generated

Here we have the final optimized model with Recall = 0.846 and Precision = 0.537.

L1-regularized Logistic Regression coefficient:

A screenshot of a cell phone

Description automatically generated

Random Forest feature importance:

A screenshot of a cell phone

Description automatically generated

From the data above, we can conclude that the most important features to the model prediction are 'recategorized American Joint Committee on Cancer Tumor Stage Code', 'Neoplasm Histologic Grade' and some other mrna features like: 'ROS1', 'HIST1H2AM' and etc.