Group 12 - Cancer Technical Challenge

Report:
Our findings, classification approach and general interpretation of the dataset

General interpretation of the dataset

Our data was synthetic data about cancer. It contained two data sets, one about patient characteristics and the other about tumour characteristics. Across the two data sets were 41 variables that gave information about the status of one hundred thousand cancer patients.

Our approach: decisions made after our interpretation
To gain a better understanding of the dataset we began by searching for definitions of unfamiliar terms. Using the NCRAS data dictionaries we acquired a good understanding of each variable and its importance on our approach. Once we had a firm understanding of each variable we could decide how to best clean the data and manage missing values.

Some of the variables had a relatively simple management strategy, others were more complex. In this section we will briefly discuss the major variables that were dropped, and the logic behind dropping them, as well as our strategies for filling missing values.

- **Gleason patterns:** Gleason patterns were swiftly dropped from our data as they are specific to prostate cancer and therefore would skew our model for all other types of cancer. Additionally, all values for the gleason score were missing, and it would be impossible to fill them in with the data that we had.
- **Breast cancer variables:** All variables specific to breast cancer (oestrogen, progesterone and human epidermal growth factor receptor 2) were dropped. This decision follows the same logic as the previous. Including variables that are specific to breast cancer will bias our model, and therefore decrease accuracy.
- **Death cause codes:** Variables that describe the cause of death (death cause codes) were the most difficult to deal with in our experience.With the understanding that death cause codes followed a step wise process we knew that 1a was the most relevant cause of death and that 1b was a contributor to 1a, in the same way that 1c was a contributor to 1b. We decided that in order to simplify death cause, the best course of action was to l create one variable (Death cause) that prioritises 1a. In addition we filled in an inapplicable death cause value to all those who were alive.
- **Staging:** TNM staging is good at separating cancer staging into three broad groups, however for our model it was not useful, as overall stage best encapsulated all of the data held in the TNM without missing values, therefore T, N and M were dropped from our data frame.
- **Tumour site:** Site codes are an important factor when exploring prognosis. In our case we had two variables that explored the site, one of the variables was more accurate than the other. The most accurate site variable had 49 missing values, and therefore we filled those values with the less accurate ones and then dropped the less accurate site column.

- **Morphology, behaviour and grade:** Morphology, behaviour and grade are not strictly interchangeable, but the three variables are correlated. Aggressive behaviours are correlated with higher grades of cancer, as are certain morphologies. Consequently we used morphology and behaviour to aid the filling of the 604 missing grade values.
- **CNS:** Although the use of data around the type of clinical nurse specialist could have been useful in our model, there were far too many missing values, and the data that we had would have been unable to fill in these values, therefore we thought it best to drop this variable.
- **ACE27:** This variable looks at other diseases that a patient may have. They are a useful indicator of a patient's overall health and life expectancy, however, it is not cancer specific and therefore irrelevant to our model.Additionally, there were many missing values for this variable that could not have been filled with the data that we were given.
- **Date of first surgery:** Cancer is broad in its treatments; some types will require surgery whilst others require a different approach. Acknowledging this, we thought it best to drop surgery as it would bias the model as it doesn't account for patients who receive other forms of treatment.
-

Findings:

Classification:

Machine Learning Report: Classification of Patient Vital Status

Introduction:
This report presents the classification of patient vital status using various machine learning algorithms. The dataset contains patient information, and the goal is to predict the patient's vital status. The report covers data preprocessing, model selection, feature selection, and evaluation.

Libraries Used:
The following libraries were used in this analysis:

- `category_encoders` for target encoding
- `sklearn.preprocessing` for data preprocessing
- `sklearn.pipeline` for creating pipelines
- `sklearn.compose` for column transformations
- `sklearn.impute` for data imputation
- `sklearn.model_selection` for train-test splitting and cross-validation
- `seaborn` and `matplotlib.pyplot` for data visualisation

Data Preprocessing:
The dataset was split into informative features (X_train and X_test) and target labels (y_train). Uninformative features such as primary and foreign keys were dropped from the datasets. Categorical features were one-hot encoded using `OneHotEncoder` with handling

of infrequent categories. Numerical features were imputed with mean values and standardised using `StandardScaler`.

Model Selection:
Three classifiers were chosen for this analysis:

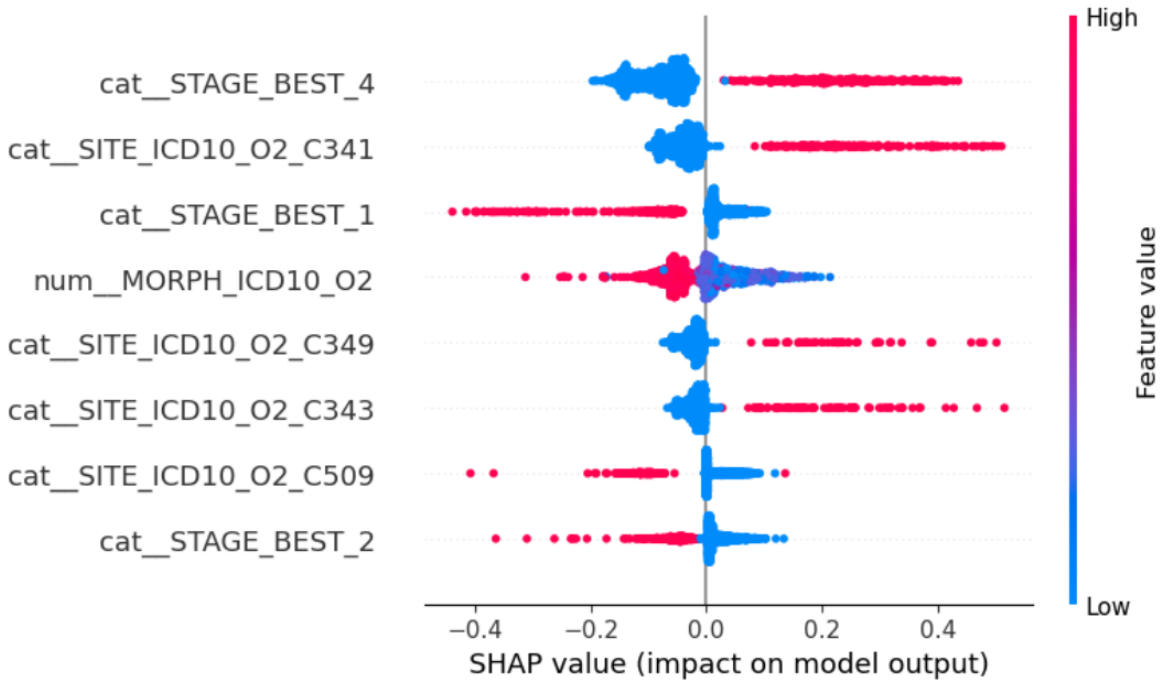
1. Support Vector Machine (SVM) with a linear kernel
2. k-Nearest Neighbors (KNN) Classifier
3. Random Forest Classifier

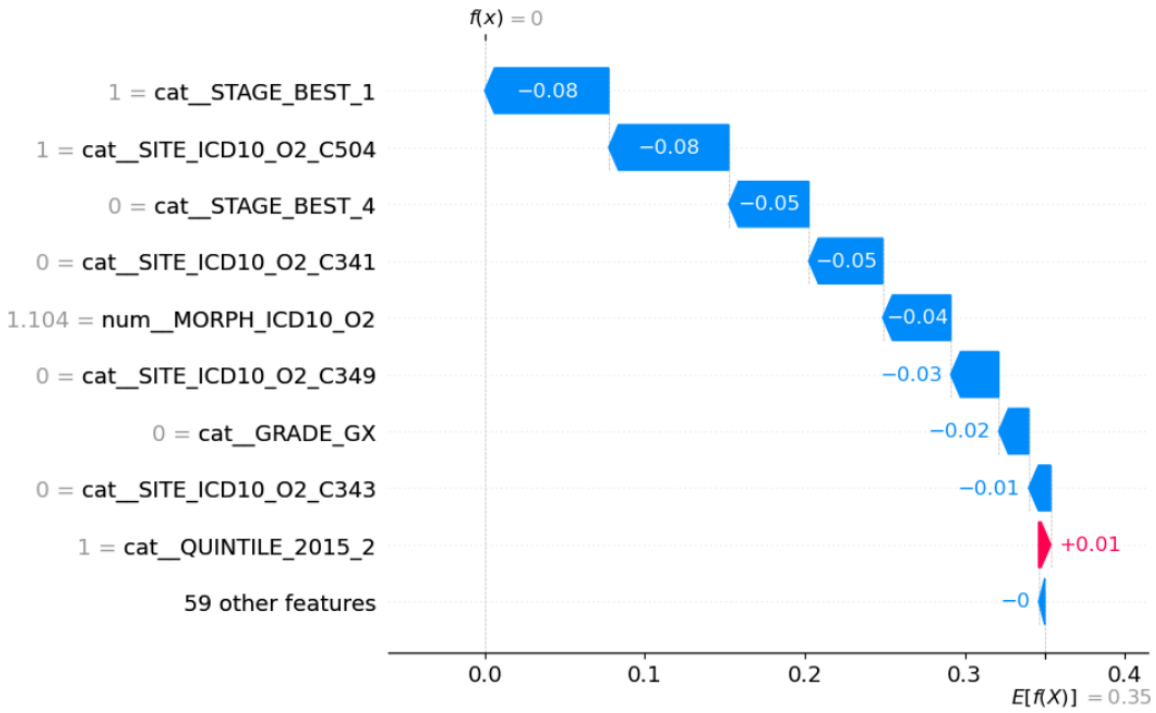
Feature Selection:
Recursive Feature Elimination (RFE) with a linear regression estimator was employed to select the most important features for prediction. Grid search was used to find the optimal number of features to retain.

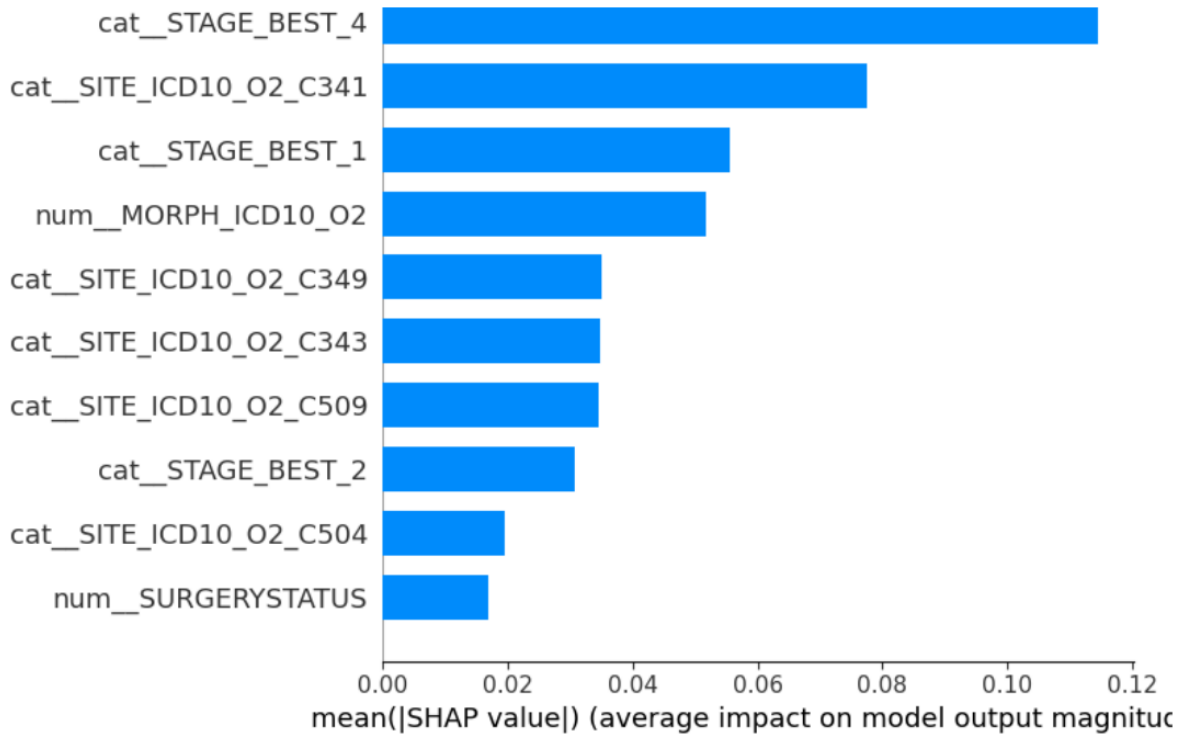
Evaluation:
The models were evaluated using accuracy as the metric. The accuracy of each model on the test data was as follows:

Model Name	**SVM**	**KNN**	**Random Forest**
Accuracy	77.81%	78.08%	78.85%

SHAP Analysis:
A random sample of size 1000 was taken from the test data for SHAP (SHapley Additive exPlanations) analysis. SHAP values provide insights into feature importance for individual predictions and global model behaviour.







SHAP Analysis and Feature Importance:
The SHAP analysis revealed important insights into the feature importance for predicting patient vital status. The most influential feature was found to be `STAGE_BEST_4`, indicating that the cancer stage at diagnosis plays a significant role in determining the patient's vital status. Following closely in importance was the feature

`SITE_ICD10_O2_C341`, which suggests that the specific location of the cancer has a notable impact on the prediction.

Suggestions for Further Work:
While this analysis provides valuable insights, there are opportunities for further exploration and improvement:

1. Feature Engineering: Investigate the creation of new features that capture interactions between existing features. Domain knowledge can be leveraged to engineer features that might enhance the model's predictive power.

2. Model Tuning: Experiment with hyperparameter tuning for the selected models. Different parameter combinations might lead to improved model performance.

3. Ensemble Methods: Explore ensemble methods like stacking or boosting to combine the predictions of multiple models. Ensembles often result in enhanced generalization and predictive accuracy.

4. Feature Importance Refinement: Continue refining the feature importance analysis using alternative methods such as permutation importance, which provides a complementary perspective on feature relevance.

5. Domain Expertise: Collaborate with domain experts to gain a deeper understanding of the medical factors influencing patient vital status. Their insights can lead to better feature selection and model understanding.

6. Handling Imbalanced Data: Investigate techniques to handle potential class imbalance in the dataset, which could lead to more balanced and accurate predictions.

7. Interpretable Models: Consider using models that offer more interpretability, such as decision trees or logistic regression, to enhance the understanding of how individual features impact predictions.

Conclusion:
In conclusion, this analysis successfully employed machine learning techniques to predict patient vital status. The models' performance was evaluated, and the Random Forest model demonstrated the highest accuracy. SHAP analysis highlighted the importance of the `STAGE_BEST_4` and `SITE_ICD10_O2_C341` features in predicting vital status. To further enhance the predictive power and interpretability of the models, future work should focus on feature engineering, model tuning, ensembling, and collaboration with domain experts. These efforts will contribute to better insights into patient outcomes and ultimately improve the healthcare decision-making process.