Approach for Analyzing the cancer data:

The entire methodology was broken down to the following parts:

1) Initial research

2) Data extraction

3) Data exploration

4) Data processing and Feature engineering

5) Modeling, validation and prediction

These steps are explained as follows:

1) Initial research

From the problem description, we could understand that data is related to prostate cancer and the below are a few lines corresponding to the research we did on the internet.

- We read that age can be a crucial factor as risk of cancer increases after the age of 50 years and we could also see that most of the patients in the data are above 50.
- Even family members more specifically brothers having cancer can increase your chances of getting it.
- Race has high significance as African-Americans can have 60% more chance than Americans in getting a prostate cancer.
- PSA level in the blood would also contribute to the significance of possibility of getting cancer.
- Even people drinking tea more than 7 times a day would have 50% higher risk than the rest. However, I don't think I could see anyone in the data who consume more than once in a day as the scale is in week.
- Smoking is one of the primary causes of any type of cancer or at least a supporting factor in it's spread.
- Symptoms are definitely important.
- Height may not be of significance unlike weight.

2) Data extraction

- The train and test data were extracted from the participant files section
- Summary statistics about the data were observed

3) Data exploration

- The data exploration was carried out on the train data, so that we could have a general feeling about the texture of the data
- The response variable 'survival_7_years' was plotted against multiple predictors
- In each case, the variation against the predictor was observed and insights, if any, were noted. For example, people having more cups of tea had higher chance of survival. If the cancer has not spread to lymph nodes, patient has higher chance of survival

- All the insights are mentioned along with the respective codes in the file 'EnovaDataChallenge_DataProcessing.ipynb'
- These insights helped us to process data and create new variables from the existing ones

4) Data processing and Feature engineering

- After data exploration on the train data, both train and test data were combined for ease of processing. A new Identifier column was added to help separate train and test data in the future
- Correlation among the numeric attributes were observed. Not many variables were highly correlated so multicollinearity was not observed to be a major problem
- The attributes were processed based on the data dictionary, data exploration stage, correlation among the variables among other factors
- Multiple new features were created which we felt could better describe the response variable like 'mnths_from_diag', 'tumor_change_6mnths',' psa_change_1yr' among others. Several categorical variables were converted to binary variables so that they could be used in the predictive model
- The missing data treatment also differed from variable to variable. In few cases knn-imputation was used since some groups were strongly correlated based on correlation as well as the data dictionary; the % of missing values was high which meant replacing by mean or median would not have been an accurate choice. In other cases, the % of missing values were low. So, we replaced by the median value in case of numeric attributes and by a flag value (like -1 or U) to indicate unknown status for categorical variables
- It was mentioned in the problem description that the symptoms were predictive in nature but their meanings were removed. So, we created dummy variables for each symptom and included them in the model building process
- The description of the data processing carried out can be found in the 'Preprocessing' tab of 'Data Dictionary.xlsx' file as well in the 'EnovaDataChallenge_DataProcessing.ipynb' notebook along with the respective operations carried out
- After all the processing was complete, we scaled all the variables since they were all converted to continuous and binary forms and did not contain any missing data
- The processed data can be found in the file "processedData.csv"

5) Modeling, validation and prediction

- The modeling and validation code can be found in the notebook "EnovaDataChallenge_Modeling.ipynb"
- Although we tried different classification models, we zeroed in on Randomforest since it can handle large volume of data, is robust towards outliers and provided a better cross-validation accuracy compared to Logistic or SVM
- We first utilized the feature importance option available from Randomforest to understand the importance. Since there were large number of variables, this helped us to take a decision on which variables to be considered for building the final model
- We considered 80% of the variables based on their importance and carried out a grid-search and 5-fold cross-validation to obtain the most optimal parameters for the final model. This also ensured that overfitting was minimized in our approach

- Finally, based on the most optimal parameters, we built our model and generated the output for the test data. The output is stored in the file ‘Arkojyoti_score.csv’