Predicting Progression of Endometrial Cancer

DATA 602, DEEPIKA DILIP

Abstract

Background: Recent biomedical advances have pointed to the heterogeneity of cancer subtypes. Endometrial cancer primarily affects women, and progression is dependent on a series of factors

Methods: We used the 2018 Memorial Sloan Kettering cohort of uterine cancer cases (N = 187). Here, we apply three classification models (logistic regression, kNN-classification, and decision trees) to predict disease progression. We integrated DNA-sequenced data with clinically annotated tables.

Results: Both kNN and logistic regression indicated high accuracy rates (~94%). When examining feature importance, certain mutations appeared to drive disease progression (e.g. TP53, CCND1)

Conclusion: Further models should integrate clinical criteria alongside mutation data. Mutations appear to be significant predictors of disease, and assays can ultimately inform treatment options.

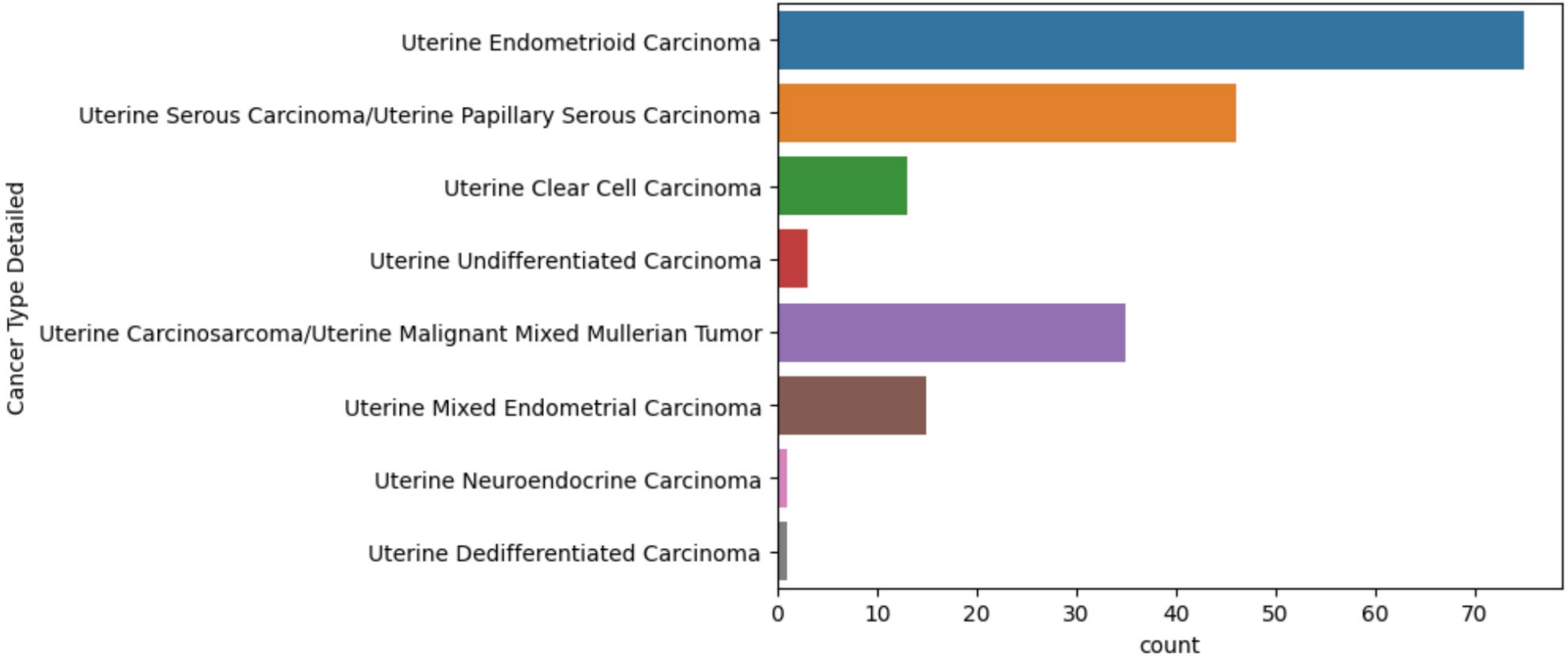
Introduction

Endometrial cancer is increasing in prevalence, yet treatment remains limited (1). As with many cancer types, a wider understanding of the cancer genomics and biomedical pathways can inform treatment options and cancer research. Certain clinical factors contribute towards progression (2), including age, co-morbidities, and pathologic risk factors. Mutation profiling is especially useful in cancer research: certain mutations are indicative of disease progression, prognosis, and dictate the best chemotherapy regimen.

Here we ask the following question: which mutations are most associated with endometrial cancer progression?

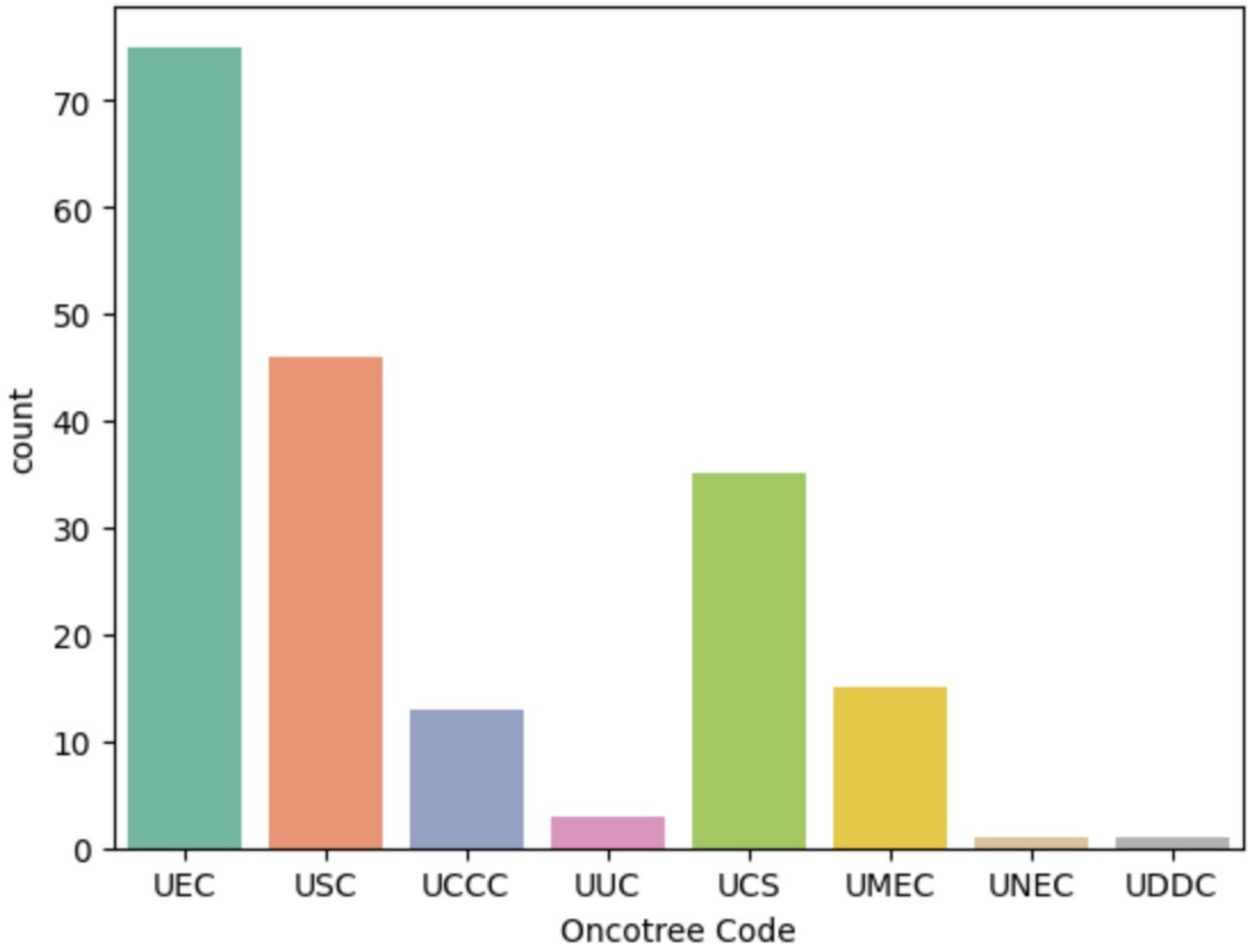
This project integrates mutation data with known clinical factors. By taking a data-driven approach, we can quantify the impact of certain biomarkers on disease progression. We used the 2018 Memorial Sloan Kettering Cohort, accessible via cBioPortal.

Exploratory Data Analysis: Clinical

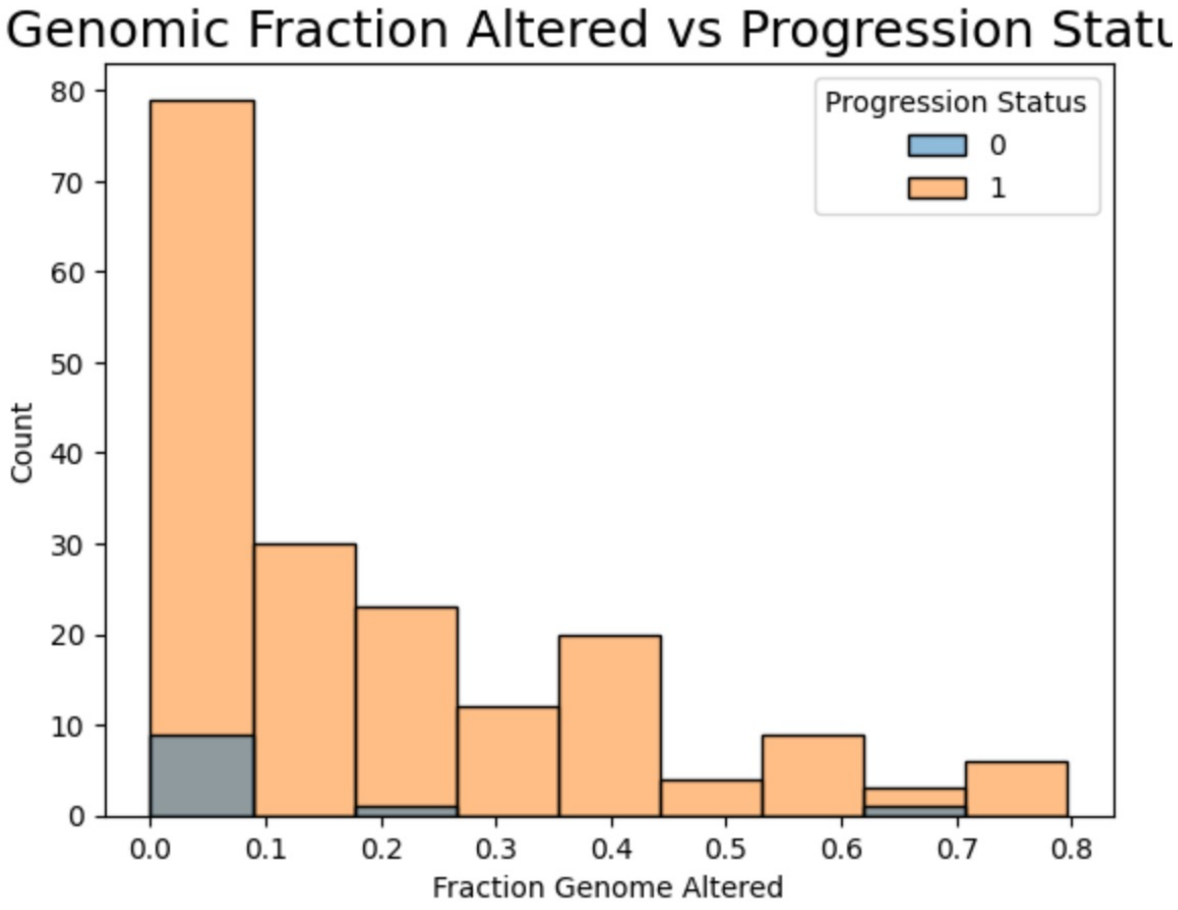


***Interpretation**: Various types of uterine cancer are contained in the cohort, could affect results*

Exploratory Data Analysis: Clinical

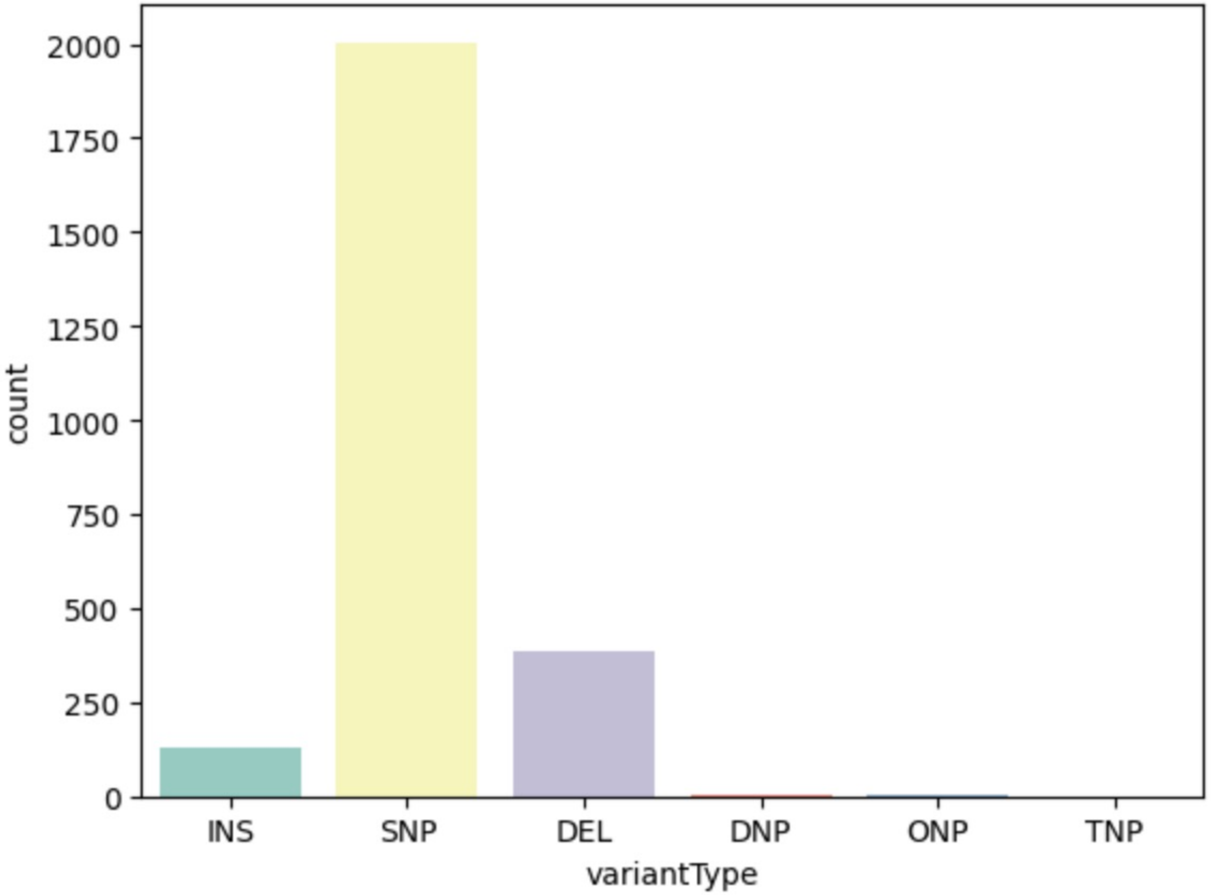


Exploratory Data Analysis: Clinical



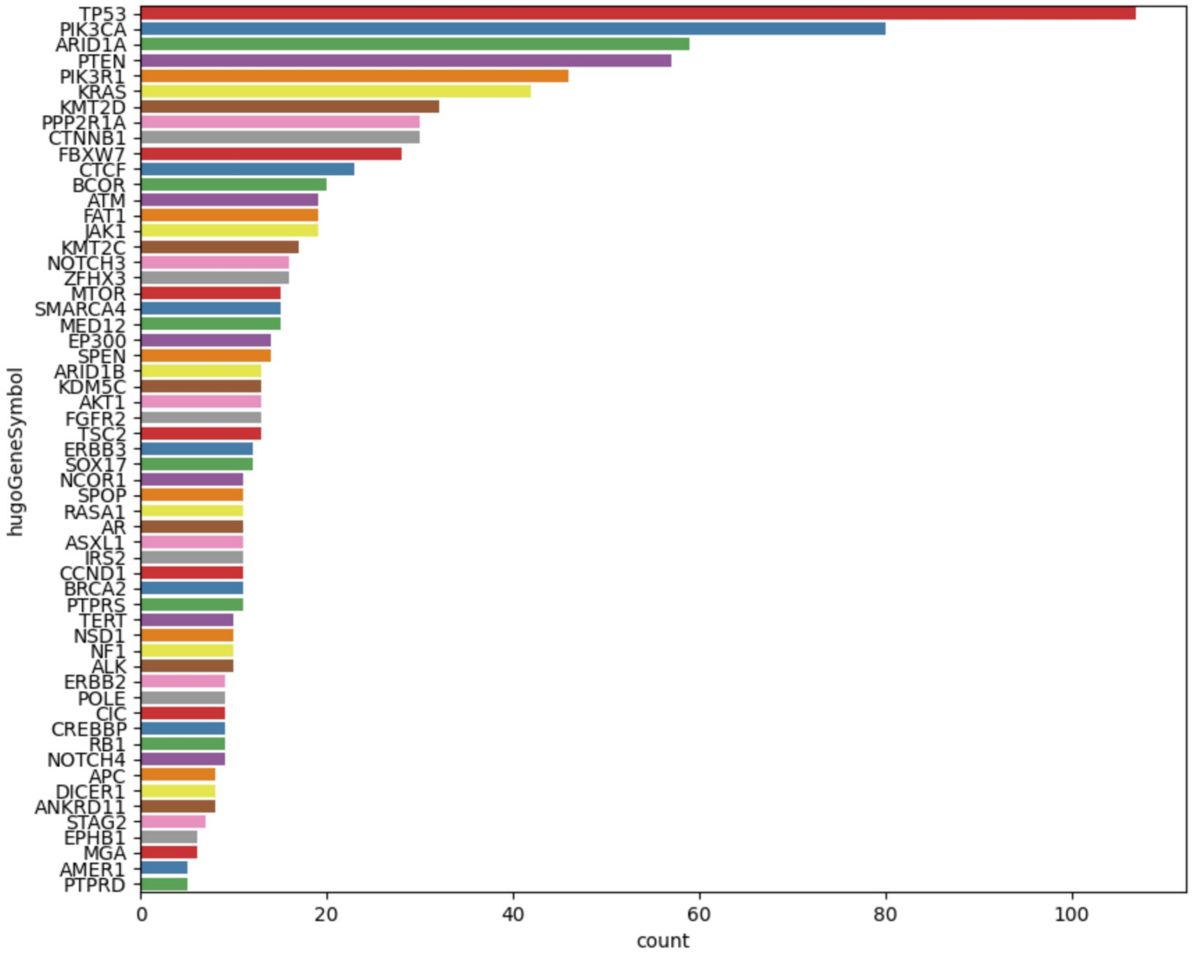
Interpretation*: The more genomic disruption, the more likely the cancer progresses!*

Exploratory Data Analysis: Mutations



***Interpretation**: Mutations seem to be consistent (single nucleotide polymorphisms). Note deletions are more disruptive and probably affect our outcome of interest*

Exploratory Data Analysis: Mutations

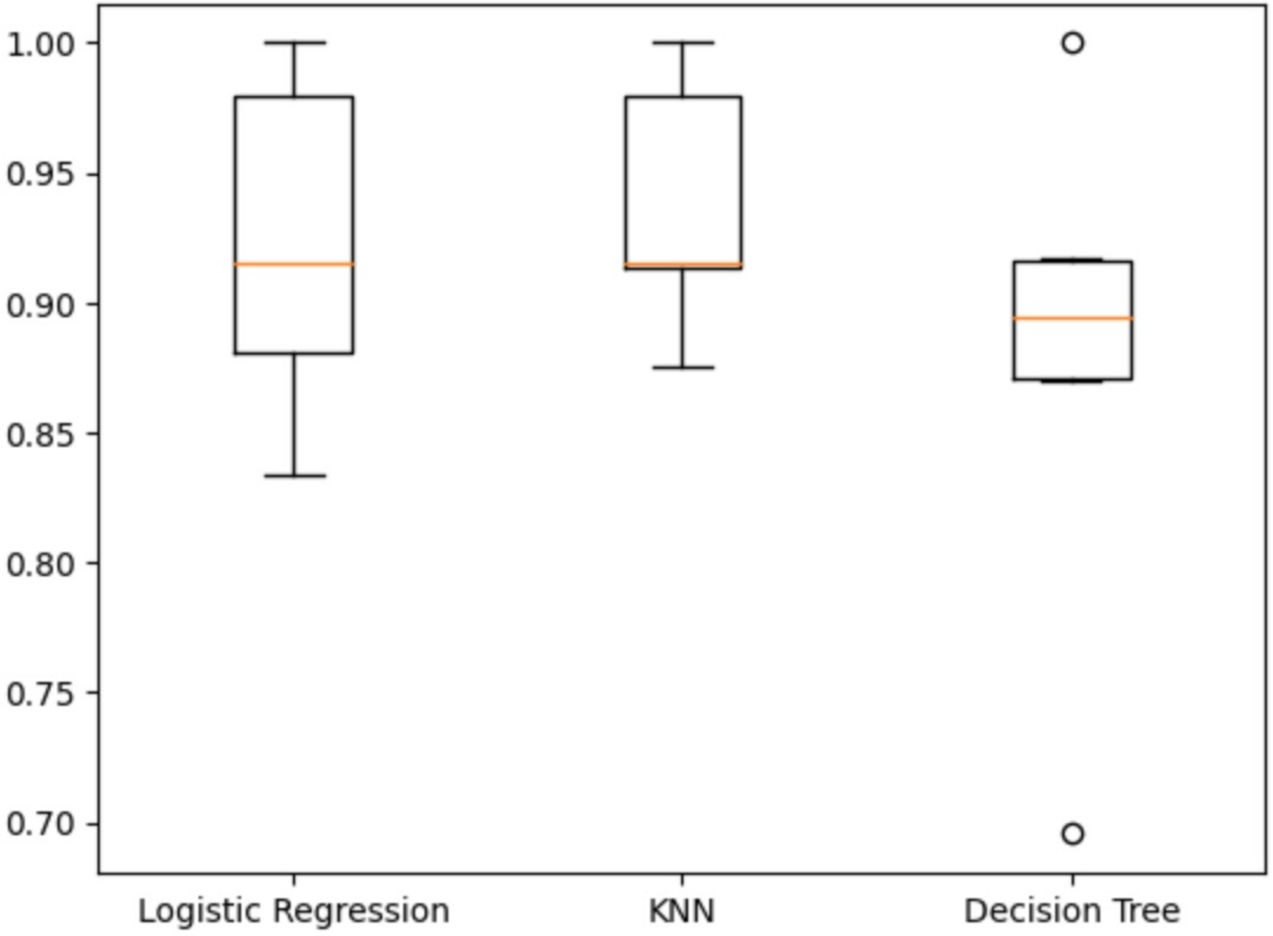


- Note: this figure only contains mutations with a frequency above 8
- Certain ones are more prevalent, with TP53 being #1
- This is unsurprising, given its role in cell division.

Model Building: Classification Models

- Model Statement: Disease Progression (1 or 0) ~ TP53 + …. (Mutations)
- We need to fit a **classifier** model.
- The following models were fit.
  - Logistic Regression
  - kNN-Classifier
  - Decision Tree
- Partition: 75% train, 25% test

Cross-Validation Accuracy (Training)



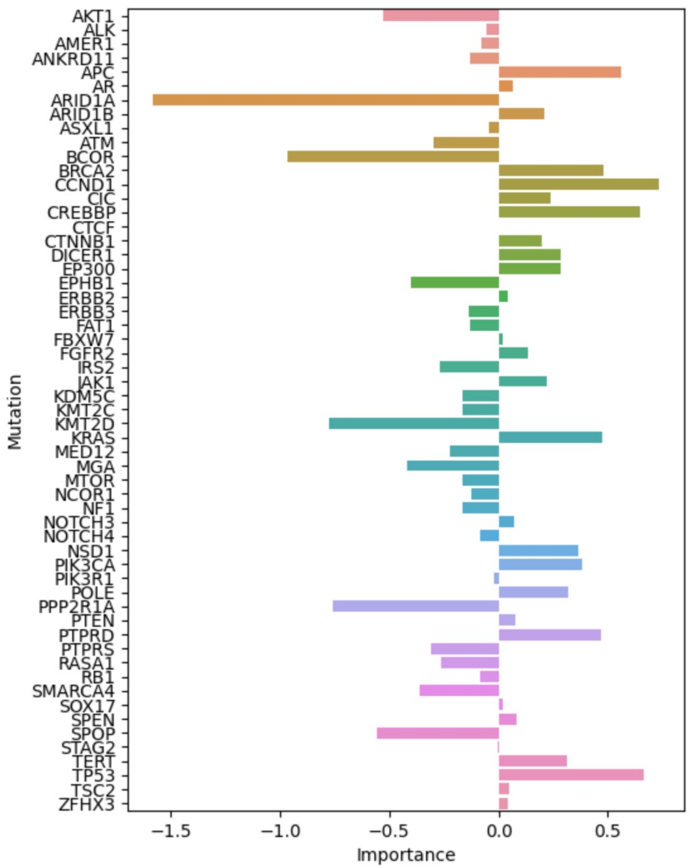
- High scores overall
- Range of logistic regression is the highest, compared to kNN.
- In aggregate, decision tree appears to be the lowest

Testing

Accuracy Scores

Model	Accuracy
Logistic Regression	0.9574468085106383
kNN classification	0.9574468085106383
Decision Trees	0.9148936170212766

Feature Ranking (Logistic)



Key Takeaways

- Molecular data can be a key determinant of clinical outcomes
- Certain mutations drive disease progression compared to others
- Future models should account for clinical criteria such as co-morbidities
- **The fancier model is not necessarily the best**