MACHINE LEARNING ALGORITHMS TO PREDICT & DIAGNOSE BREAST CANCER

Dhruvi Agarwal
Shambhavi Dhakal
Sejal Bhargava

AGENDA

- Introduction
- Dataset Description
- Modelling Technique
- Exploring Variables
- Results/ Analysis

INTRODUCTION

- BREAST CANCER CONTINUES TO BE A SIGNIFICANT HEALTH CONCERN FOR PEOPLE ALL OVER THE WORLD.
- EARLY DETECTION OF BREAST CANCER CAN INCREASE THE CHANCES OF SUCCESSFUL TREATMENT AND IMPROVE SURVIVAL.
- OUR PROJECT AIMS TO PREDICT THE SURVIVAL STATUS OF WOMEN BASED ON TUMOR SIZE AND OTHER VARIABLES WITH THE CONCEPTS OF MACHINE LEARNING.

DATASET DESCRIPTION

The dataset used for this project was a public available dataset released on Kaggle

There are 16 columns in the data set including age, tumor size & status of the patient. In total, there are 4024 observation of patients.

Data Source: https://www.kaggle.com/code/koustavghosh149/data-visualization-using-breast-cancer-dataset/input

DATA DICTIONARY

Field	Description
Age	Age of the patient
Race	Race of the patient
Marital Status	Marital status (Divorced, married, separated, single & windowed)
T stage	Tumor size and how far the cancer is spread
N stage	"N stage" in cancer refers to the extent of spread of cancer cells to nearby lymph nodes.
A stage	Regional or Distant (How far the Neoplasm is spread)
Differentiate	This describes how much or how little tumor tissue looks like the normal tissue it came from.
Grade	Grade determines the growth speed of the cancer
Tumor Size	Size of the tumor in millimeters
Regional node examined	Refers to the number of nearby lymph nodes that have been tested for the presence of cancer cells.
Regional node positive	Refers to the number of nearby lymph nodes that have tested positive for the presence of cancer cells
Survival months	Survival months of patients
Status	Dead or Alive

RESEARCH QUESTION

RQ1: Is there a correlation between the size of the tumor and the likelihood of surviving breast cancer?

RQ2: Is there a correlation between a woman's race and the risk of developing breast cancer?

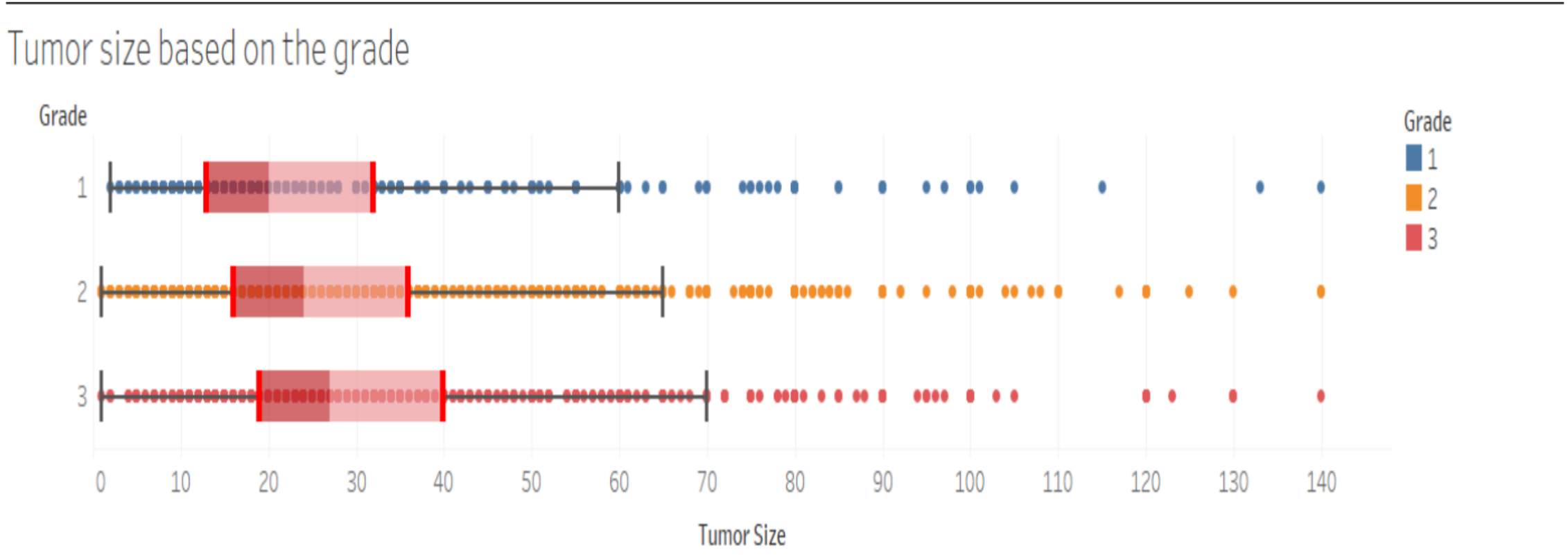
HYPOTHESIS

Ho: There is no correlation between the tumor size and the likelihood of survival
H1: There is a significant correlation between the tumor size and the likelihood of survival

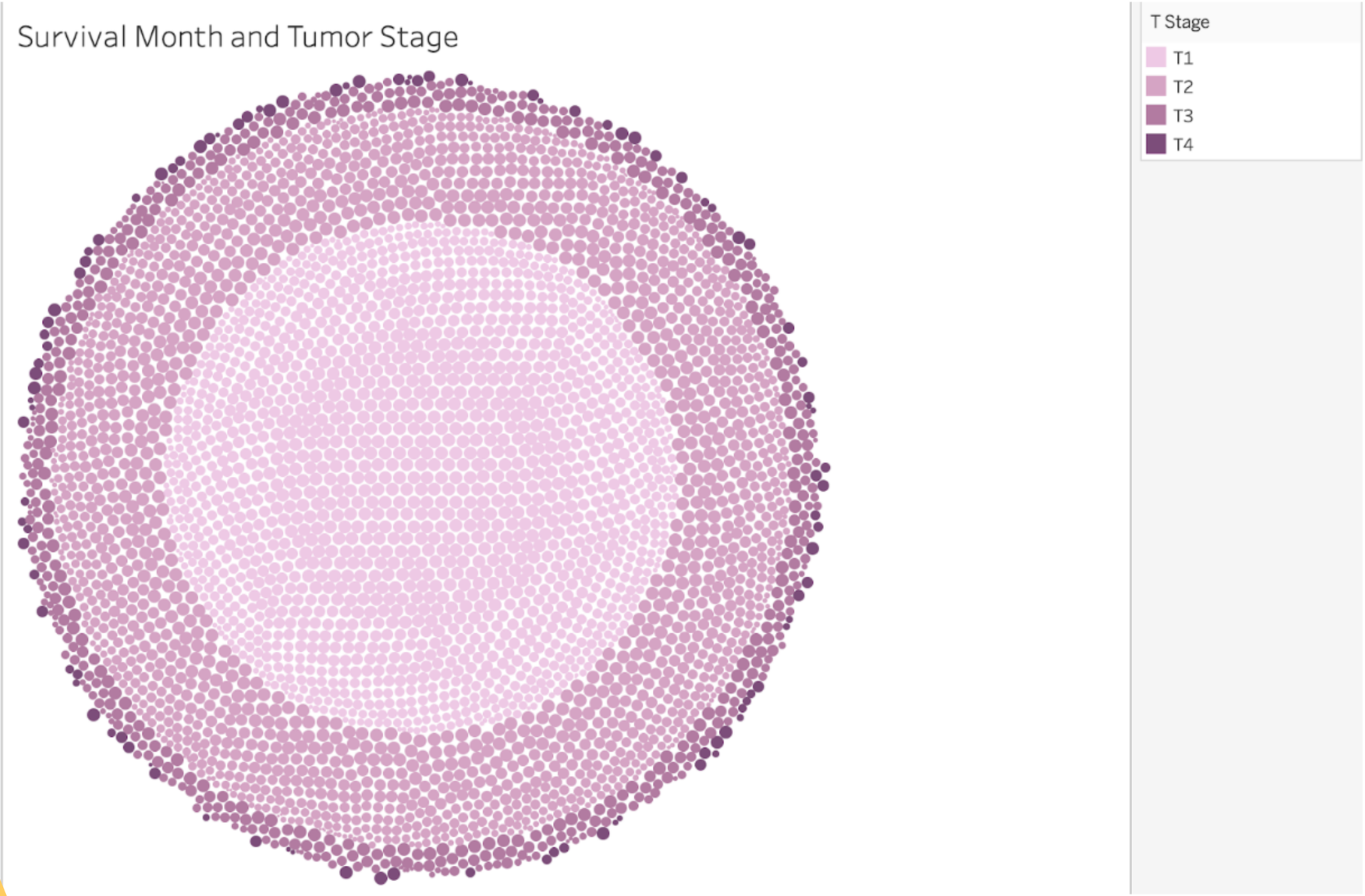
Ho:There is no correlation between a woman's race and survival status.
H1: There is a correlation between a woman's race and the survival status.

Descriptive Analysis - Tableau & Python

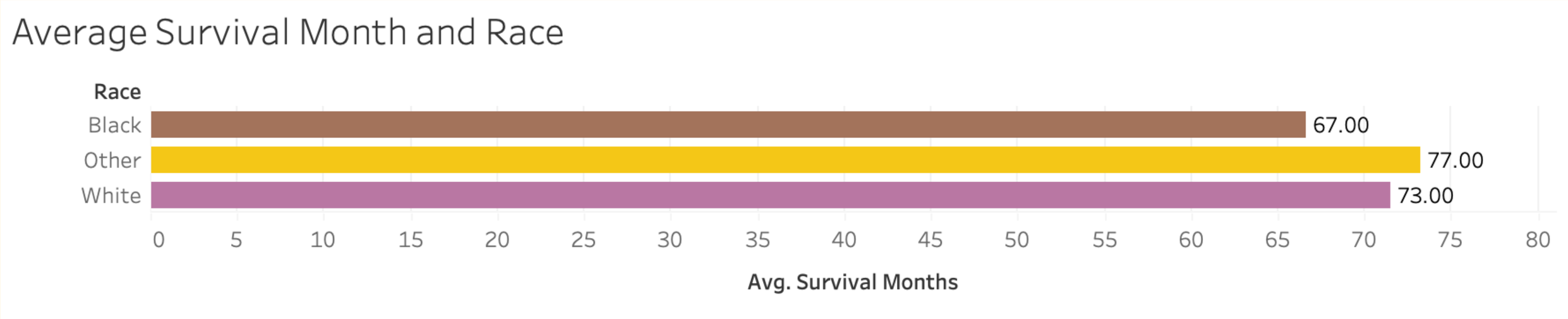
Predictive Analysis - Knime



The visualization displays the tumor size of cancer patients based on the grade of tumor



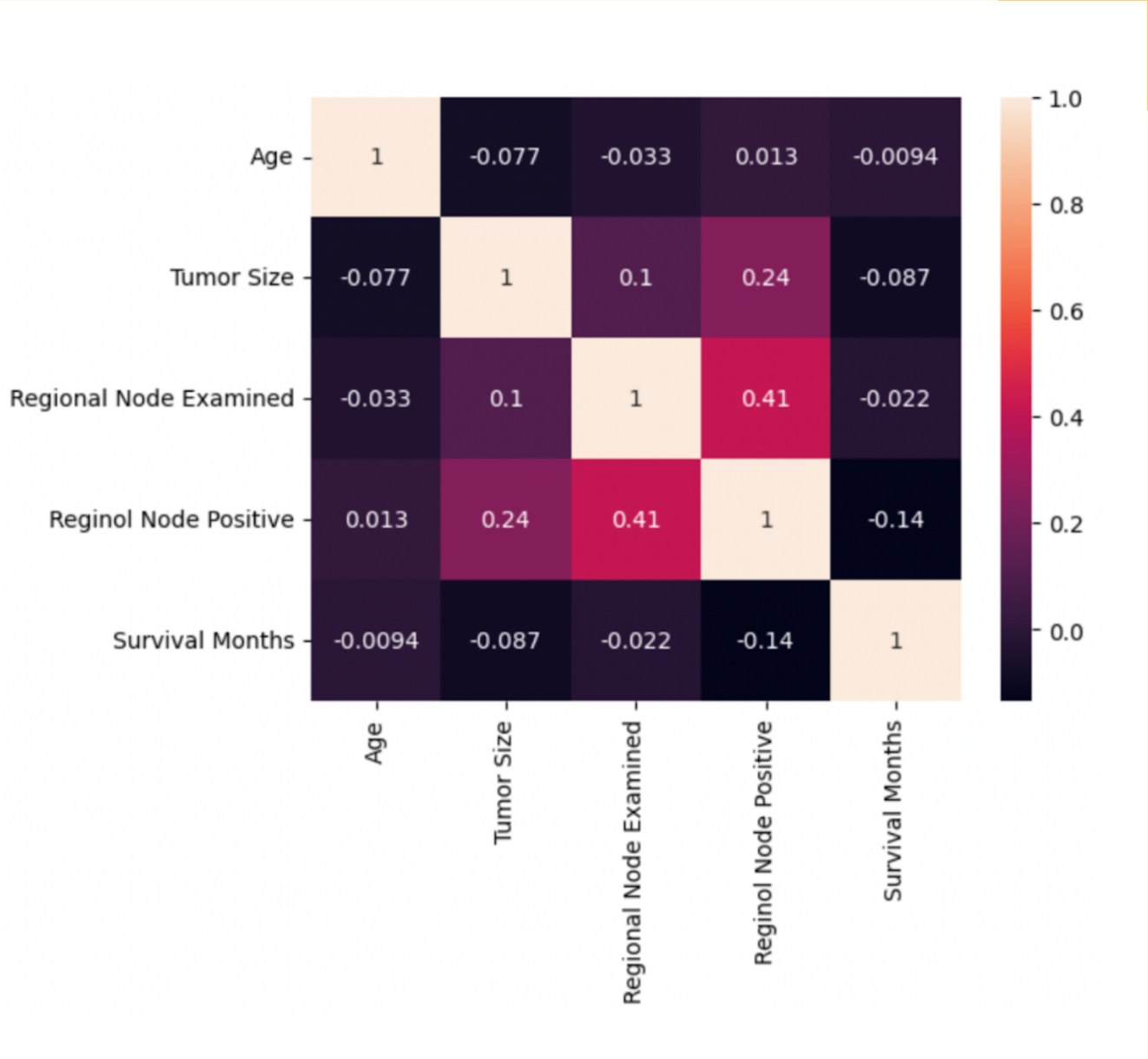
The visualization displays the relation between survival month and tumor stage



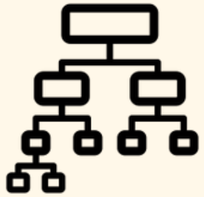
The visualization displays the Avergae Survival Months based on Race

HEATMAP

- THE HEATMAP HELPS US IDENTIFY WHICH VARIABLES ARE RELATED TO EACH OTHER.
- THE LIGHTER COLORS SHOW THE VARIABLES THAT GREATLY RELATE TO EACH OTHER.
- THE DARKER COLORS REPRESENT VARIABLES THAT ARE NOT CLOSELY RELATED TO ONE ANOTHER.



MODELLING TECHNIQUES



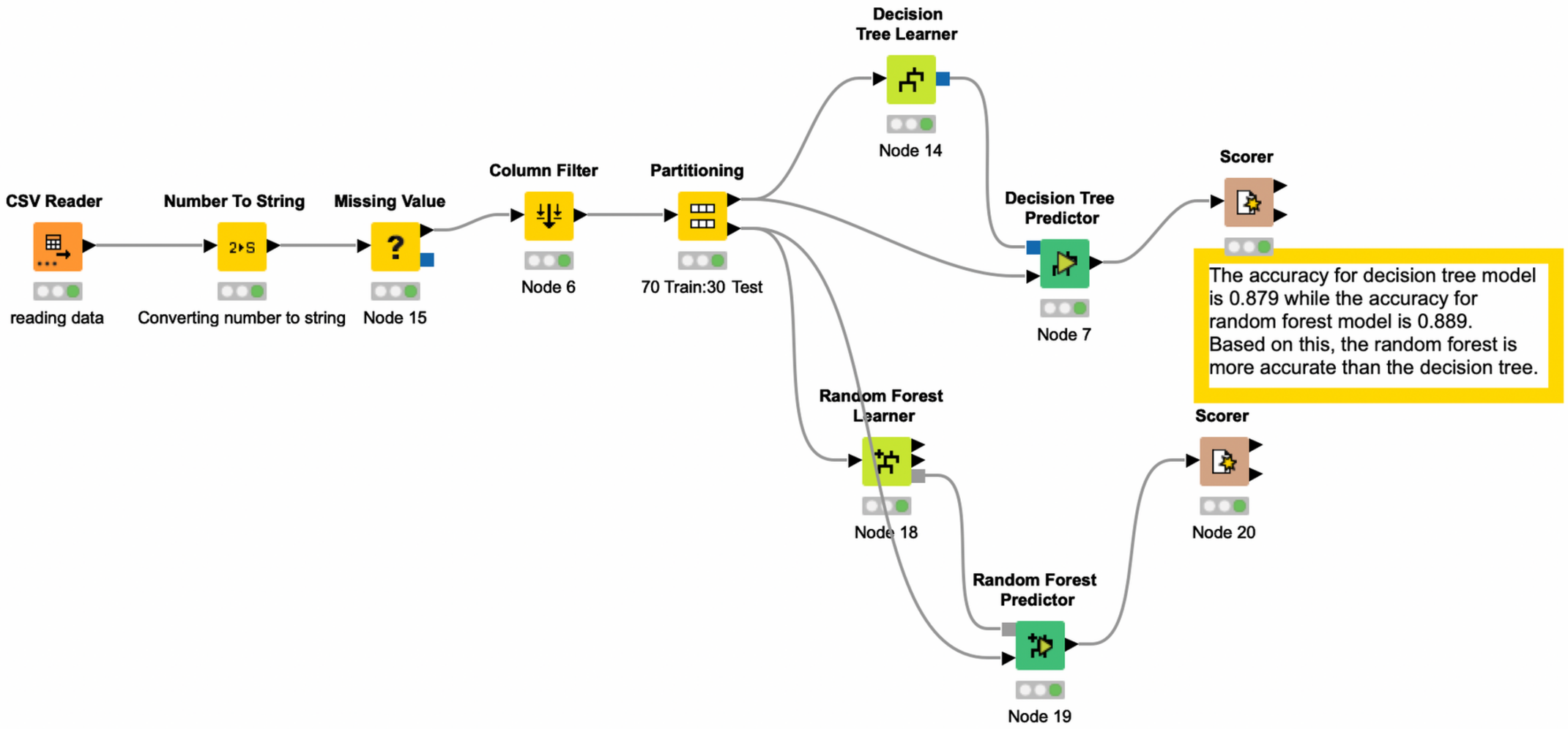
Decision Tree

Random Forest

Logistics Regression

Naive Bayes

If there is a correlation between the tumor size and the likelihood of survival



The accuracy for decision tree model is 0.879 while the accuracy for random forest model is 0.889. Based on this, the random forest is more accurate than the decision tree.

Decision Tree

Table "default" – Rows: 3 | Spec – Columns: 11 | Properties | Flow Variables

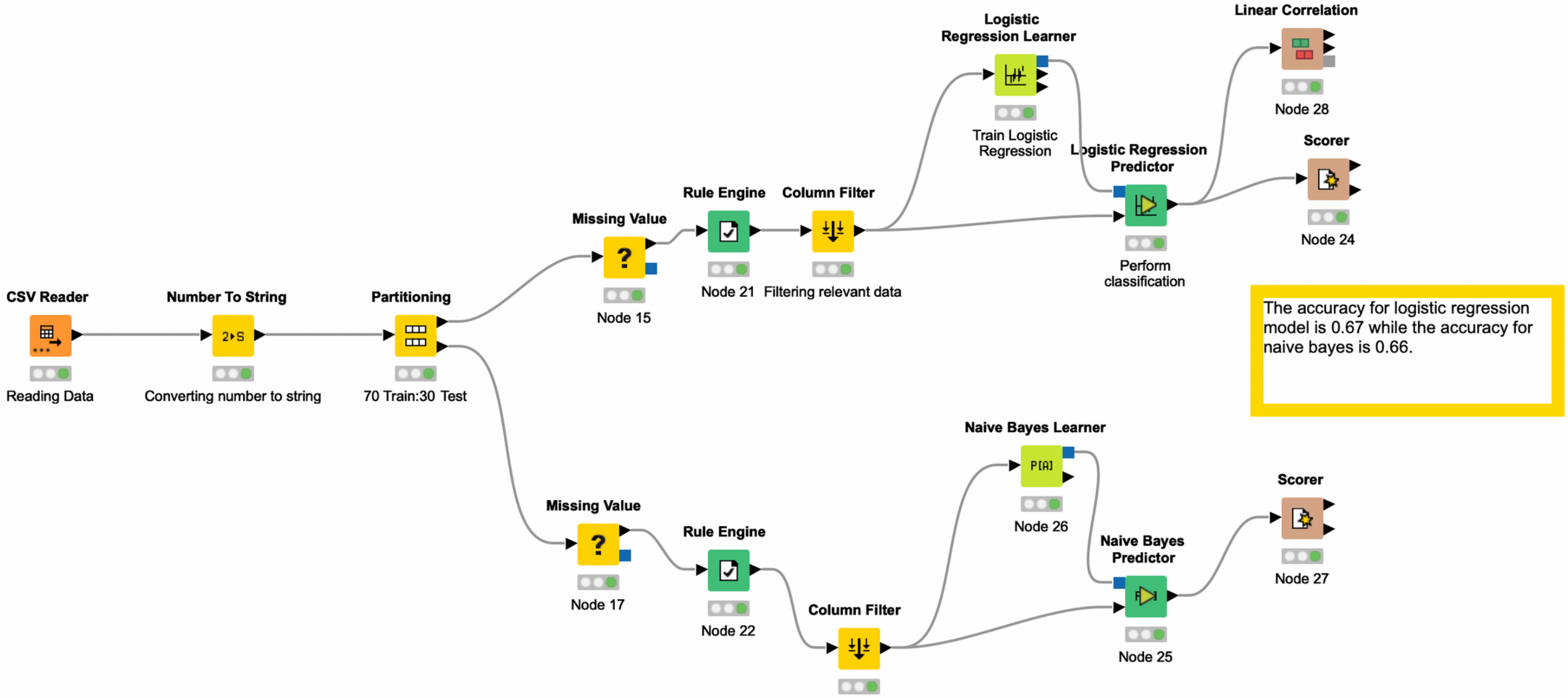
Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifi...	D F-me...	D Accuracy	D Cohen's kappa
Alive	2337	302	137	40	0.983	0.886	0.983	0.312	0.932	?	?
Dead	137	40	2337	302	0.312	0.774	0.312	0.983	0.445	?	?
Overall	?	?	?	?	?	?	?	?	?	0.879	0.39

Random Forest

Table "default" – Rows: 3 | Spec – Columns: 11 | Properties | Flow Variables

Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifi...	D F-me...	D Accuracy	D Cohen's kappa
Alive	1030	133	44	1	0.999	0.886	0.999	0.249	0.939	?	?
Dead	44	1	1030	133	0.249	0.978	0.249	0.999	0.396	?	?
Overall	?	?	?	?	?	?	?	?	?	0.889	0.358

If there is correlation between a women's race and survival status



The accuracy for logistic regression model is 0.67 while the accuracy for naive bayes is 0.66.

Logistics Regression

Table "default" – Rows: 3 | Spec – Columns: 11 | Properties | Flow Variables

Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifi...	D F-me...	D Accuracy	D Cohen's kappa
Likely to survive	1564	816	311	125	0.926	0.657	0.926	0.276	0.769	?	?
Likely to not survive	311	125	1564	816	0.276	0.713	0.276	0.926	0.398	?	?
Overall	?	?	?	?	?	?	?	?	?	0.666	0.225

Naive Bayes

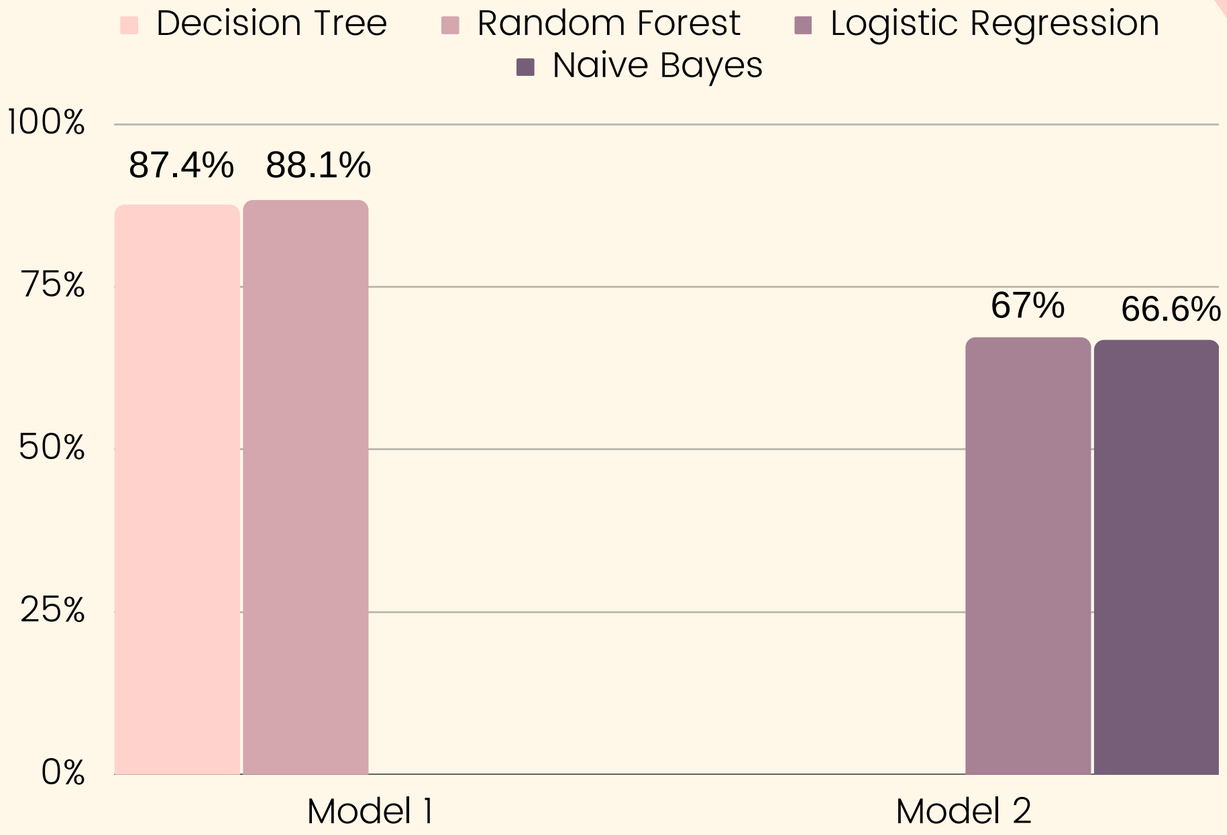
Table "default" – Rows: 3 | Spec – Columns: 11 | Properties | Flow Variables

Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifi...	D F-me...	D Accuracy	D Cohen's kappa
Likely to su...	663	46	134	365	0.645	0.935	0.645	0.744	0.763	?	?
Likely to no...	134	365	663	46	0.744	0.269	0.744	0.645	0.395	?	?
Overall	?	?	?	?	?	?	?	?	?	0.66	0.225

CONCLUSION

- **There is a correlation between tumor size and the likelihood of survival. Therefore, we reject H0.**

- **There is no correlation between a women's race and the survival status. Therefore, we accept H0.**



LIMITATIONS

- Data Quality
- Data Scope
- Representativeness
- Gap in medical knowledge

Any questions?

THANK YOU

RESOURCE PAGE

- https://doi.org/10.21147/j.issn.1000-9604.2021.06.05
- https://doi.org/10.31661/jbpe.v0i0.2109-1403
- https://www.researchgate.net/publication/344035493_Breast_Cancer_Prediction_A_Comparative_Study_Using_Machine_Learning_Techniques
- https://pubmed.ncbi.nlm.nih.gov/?term=Ayyoubzadeh%20SM%5BAuthor%5D