STAR TYPE PREDICTION

A Machine Learning Approach Using Star Data

Anuja Saira Abraham

TABLE OF CONTENTS

Topics Covered

INTRODUCTION

STAKEHOLDERS

Astronomical Research Institutions, To enhance understanding of stellar properties and evolution. Accurate classification of stars aids in research and discovery.

OBJECTIVE

Demonstrating that stars follow a specific pattern on the Hertzsprung-Russell (HR) Diagram. Using this pattern to classify stars accurately by plotting their features on the diagram.

METHODOLOGY

A PCA VS Full Model approach is developed with the goal of understanding wether a good trade of between computational time and performance can be achieved or not.
Different supervised methods are exploited and tuned with the aim of producing an accurate and reliable classification of the signal.

STAR TYPES

Red Dwarf

Small, relatively cool stars on the lower main sequence.

TEMPERATURE : 2,500-4,000 K

LUMINOSITY: 0.01 to 0.1 times of that of Sun

LIFETIME: Very long, potentially trillions of years

EXAMPLE: Proxima Centauri

STAR TYPES

Brown Dwarf

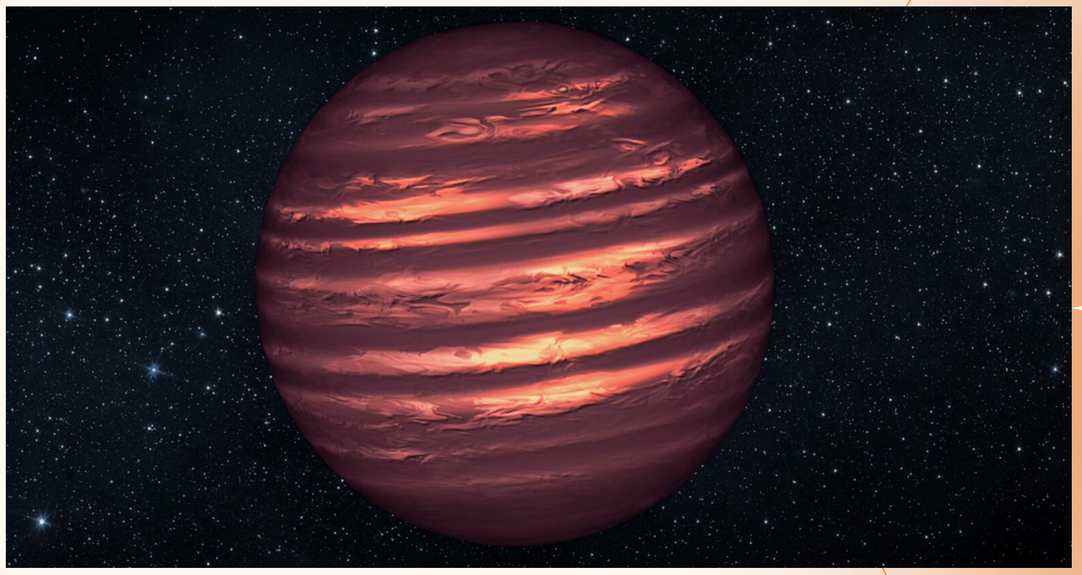
Substellar objects not massive enough to sustain hydrogen fusion.

TEMPERATURE : Below 2,500 K

LUMINOSITY: Very low, often less than 0.01 times that of the Sun

FORMATION: Bridge the gap between the largest planets and the smallest stars

EXAMPLE: Luhman 16



STAR TYPES

White Dwarf

Remnants of medium-sized stars after they have exhausted their nuclear fuel.

TEMPERATURE : Up to 100,000 K initially, cooling over time

LUMINOSITY: Low, but very dense

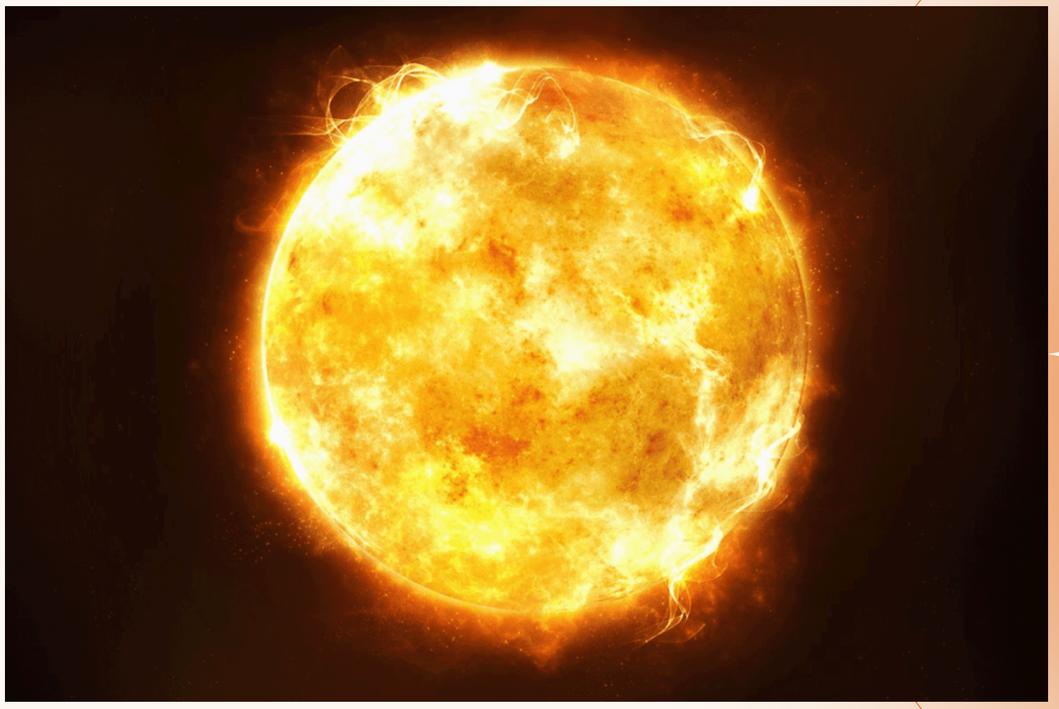
LIFETIME: Billions of years as they cool slowly

EXAMPLE: Sirius B

STAR TYPES

Main Sequence

Stars in the stable phase of hydrogen burning, forming a diagonal band on the HR Diagram.
TEMPERATURE : 3,000 - 30,000 K
LUMINOSITY: Wide range, from 0.1 to over 10 times that of the Sun
LIFETIME: Millions to billions of years
EXAMPLE: Sun, Alpha Centauri A



STAR TYPES

Supergiants

Extremely large and luminous stars, often found off the main sequence.

TEMPERATURE : 3,500 - 20,000 K

LUMINOSITY: Very high, can be tens of thousands of times that of the Sun

LIFETIME: Short, typically a few million years

EXAMPLE: Betelgeuse, Rigel

STAR TYPES

Hypergiants

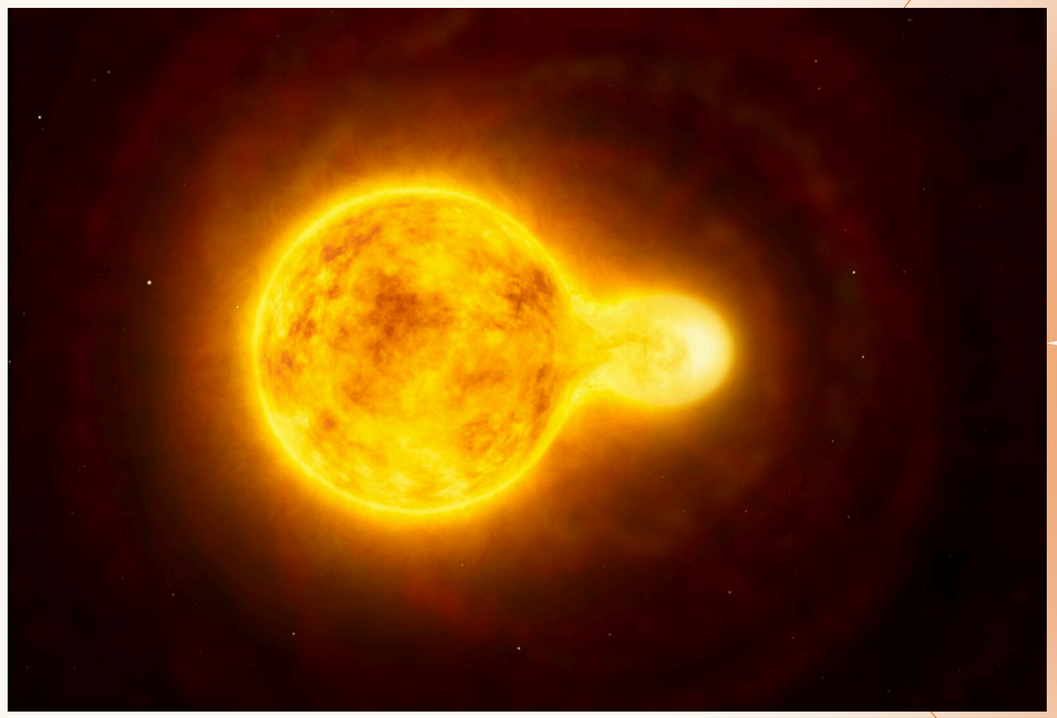
Among the most massive and luminous stars known.

TEMPERATURE : 5,000 - 40,000 K

LUMINOSITY: Extremely high, up to hundreds of thousands of times that of the Sun

LIFETIME: Very short, less than a few million years

EXAMPLE: VY Canis Majoris



HERTZSPRUNG-RUSSELL (HR) DIAGRAM.

We try to demonstrate that stars follow a specific pattern on the Hertzsprung-Russell (HR) Diagram.

A scatter plot that shows the relationship between stars' absolute magnitudes or luminosities and their spectral classifications or temperatures.

By plotting the dataset on the HR Diagram, we can observe distinct regions corresponding to different star types, confirming their classification.

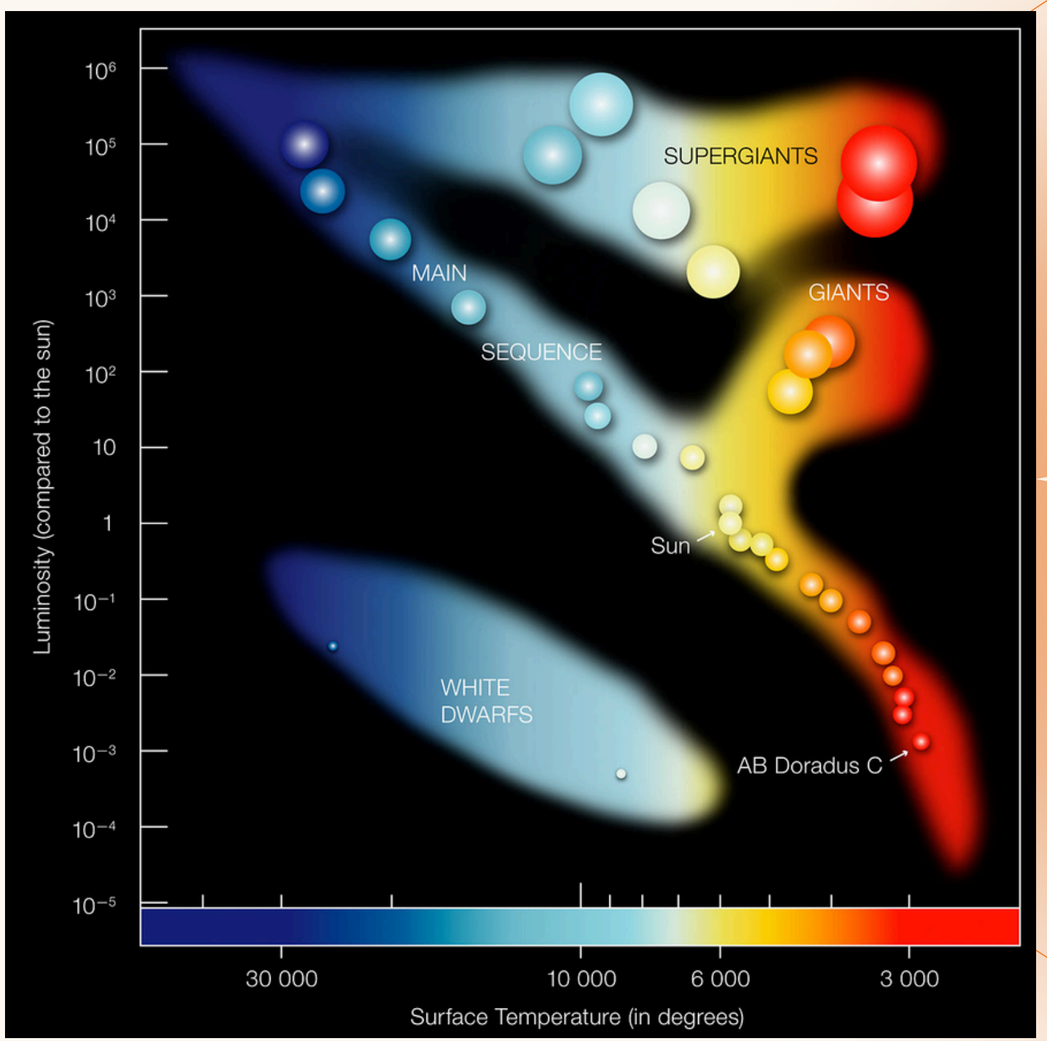


TABLE OF CONTENTS

Topics Covered

DATA COLLECTION TECHNIQUE

Stefan-Boltzmann Law

Determines **luminosity**.
This law relates the luminosity (L) of a star to its radius (R) and temperature (T), with σ being the Stefan-Boltzmann constant.

Wien's Displacement Law

Finds the **surface temperature** of a star using the peak wavelength of its emitted light.
This law relates the peak wavelength λmax of the emission spectrum to the temperature (T) of the star, with b being Wien's constant.

Absolute Magnitude Relation

Connects the **absolute magnitude** of a star to its **luminosity**.
This relation helps in determining the intrinsic brightness (absolute magnitude, M) of a star based on its luminosity (L).

Parallax Method

Determines **luminosity**.

This law relates the luminosity (L) of a star to its radius (R) and temperature (T), with σ\sigmaσ being the Stefan-Boltzmann constant.

DATASET SNAPSHOT

Temperature..K.	Luminosity..L.Lo.	Radius..R.Ro.	Absolute.magnitude..Mv.	Star.type	Star.category	Star.color	Spectral.Class
3068	0.002400	0.1700	16.12	1	Brown Dwarf	Red	M
3129	1.2200e-02	3.761e-01	11.790	2	Red Dwarf	Red	M
12990	8.5000e-05	9.840e-03	12.230	3	White Dwarf	Yellowish White	F
9700	7.4000e+01	2.890e+00	0.160	4	Main Sequence	Whitish	B
3600	3.2000e+05	2.900e+01	-6.600	5	Supergiant	Red	M

DATA SOURCE AND SPLITTING

Absolute Temperature (K)
Measurement of the star's surface temperature.

Relative Luminosity (L/Lo)
Luminosity compared to the Sun.

Relative Radius (R/Ro)
Radius compared to the Sun.

Absolute Magnitude (Mv)
Intrinsic brightness.

Star Color
Visual color.

white, red, blue, yellow, yellow-orange.

Spectral Class
Classification

O, B, A, F, G, K, M

Star Type
Category

Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, SuperGiants, HyperGiants

DATASET DETAILS

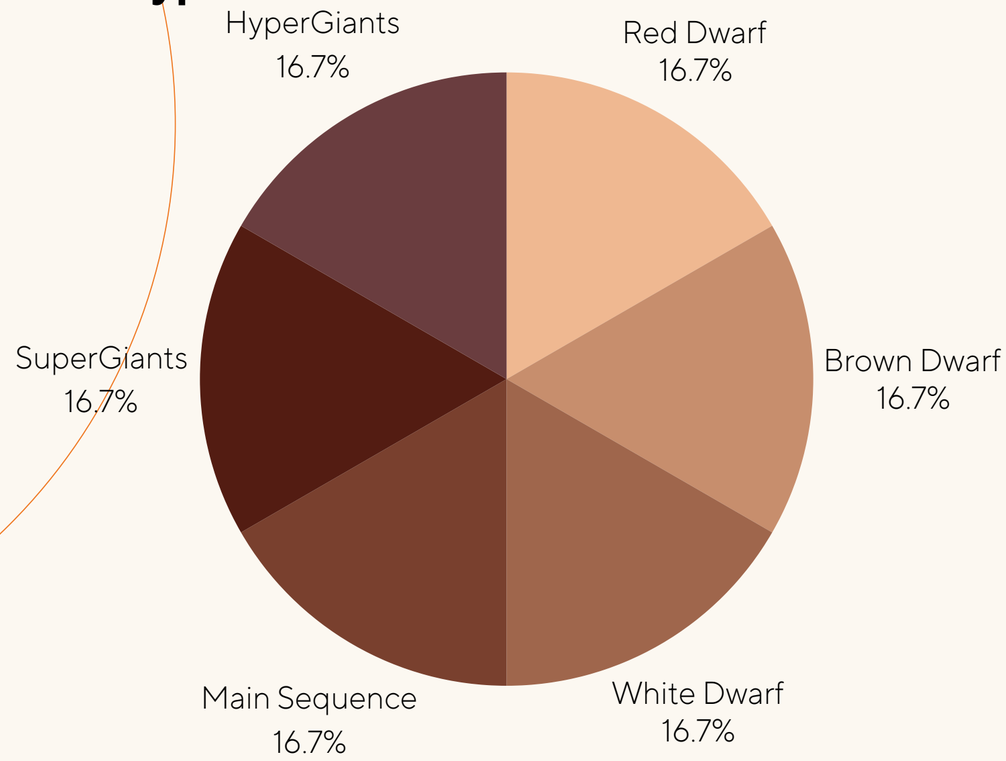
Source
Data downloaded from Kaggle.

Structure
Class “1” = Brown Dwarf, Class “2” =Red Dwarf, Class “3” = White Dwarf, Class “4” = Main Sequence, Class “5” = Supergiant, Class “6” = Hypergiant

Splitting
Data randomly split between train data and test data.
70% of observation used as train set and 30% as test set.

ABOUT THE DATASET

240 Observations
8 Variables
Response y : Categorical response variable that is the result of the type of star.
They are **Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, SuperGiants, HyperGiants.**



	1	2	3	4	5	6
TRAIN	30	29	26	29	23	31
TEST	10	11	14	11	17	9

TABLE OF CONTENTS

Topics Covered

ORIGINAL DATASET VS PCA

RANDOM FOREST

A random forest is an ensemble learning method that combines multiple decision trees to improve accuracy and robustness in predictions.

CLASSIFICATION TREE

A classification tree is a decision tree algorithm used to classify data into distinct categories based on input features.

K NEAREST NEIGHBOUR

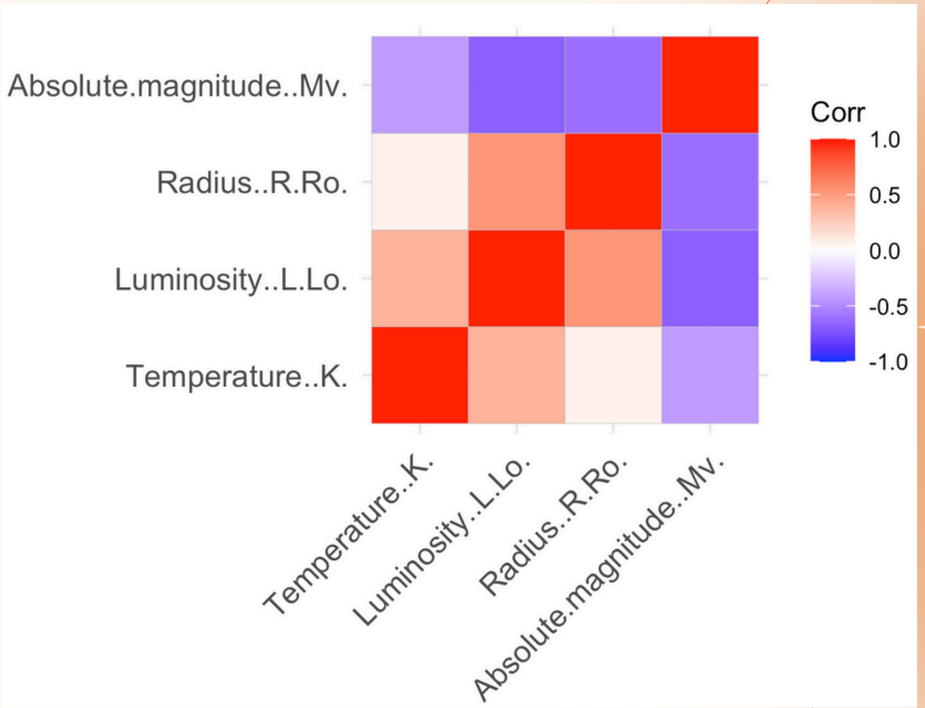
K-nearest neighbor is a simple algorithm that classifies data points based on the categories of the k closest points in the feature space.

PRINCIPAL COMPONENT ANALYSIS

- Dimension Reduction Method.
- Explore association between groups of variables.
- Consider only numerical and scaled variables

NUMBER OF PRINCIPAL COMPONENTS

Choose the number of principal components in PCA by selecting those that explain the most variance, typically determined by the cumulative explained variance plot. Often, components accounting for 80-95% of the total variance are retained.



THERE ARE NO CORRELATED VARIABLES

PRINCIPAL COMPONENT ANALYSIS

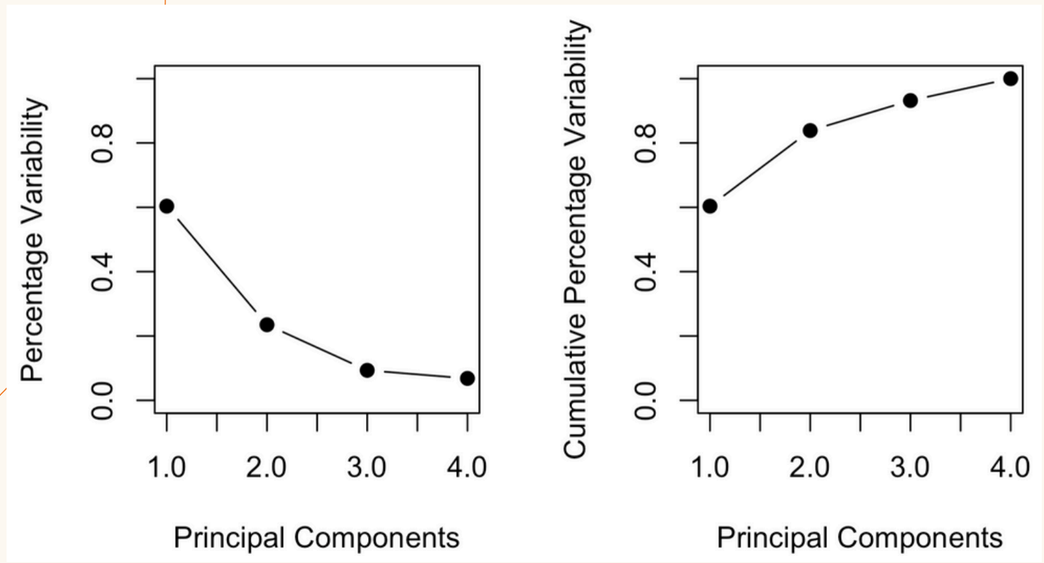
- **FIRST PRINCIPAL COMPONENT**

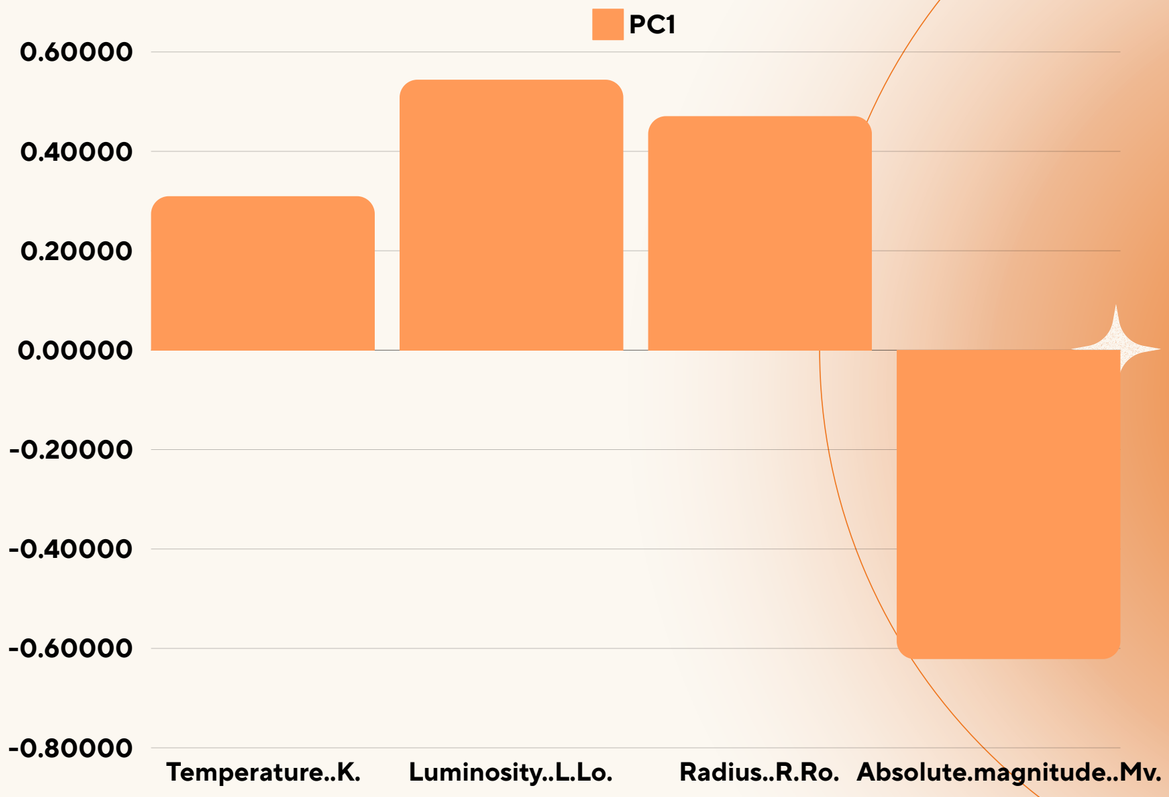
Explains around over 60% of the total variability.

- **SECOND PRINCIPAL COMPONENT**

Explains around over 20 % of the total variability.

The first two principal components together explains more than 80% of the total variability.





RANDOM FOREST

An ensemble learning method that builds multiple decision trees to enhance prediction accuracy and robustness.

Key Parameters

mtry
Number of features considered at each split.
Balances model bias and variance.
Optimal value is often determined through cross-validation.

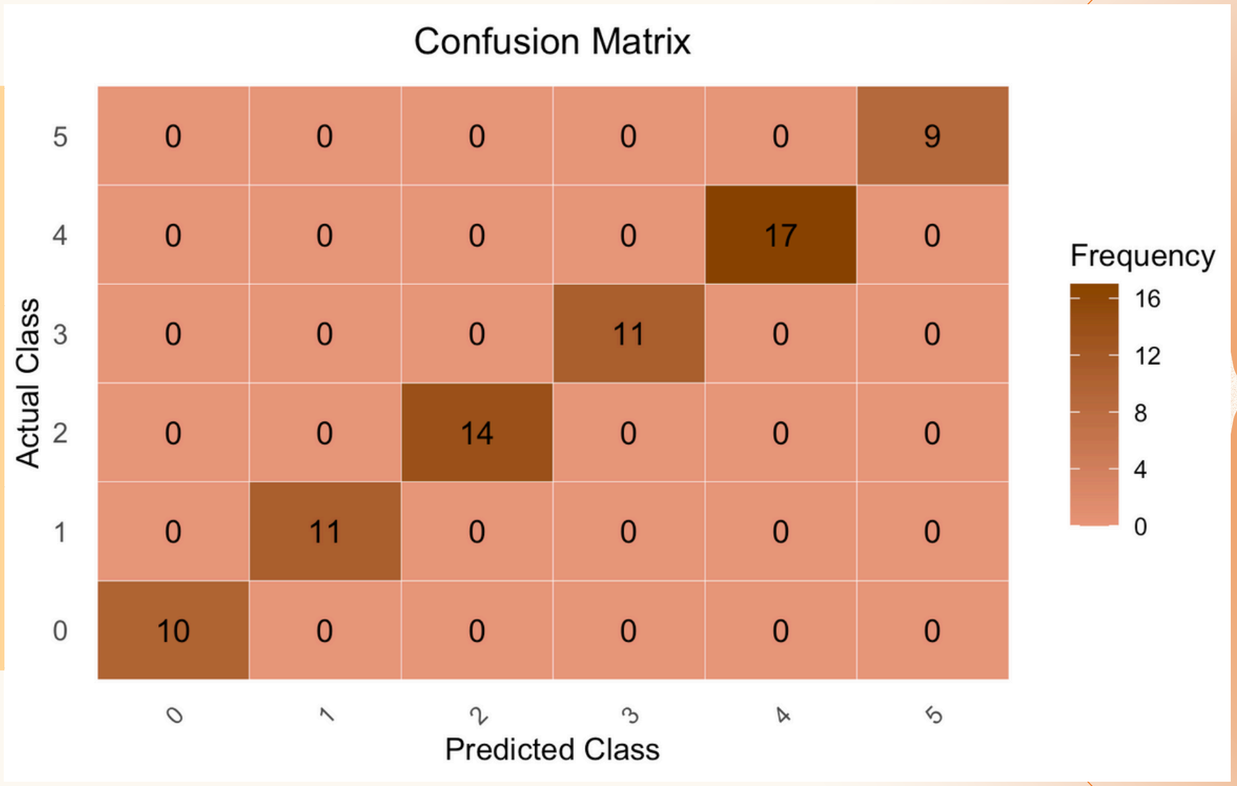
ntree
Number of trees in the forest.
More trees typically lead to better performance, up to a point.

RANDOM FOREST ON ORIGINAL DATA

Only quantitative variables are chosen

CLASS	1	2	3	4	5	6
SENSITIVITY	1	1	1	1	1	1
SPECIFICITY	1	1	1	1	1	1

ACCURACY : 1

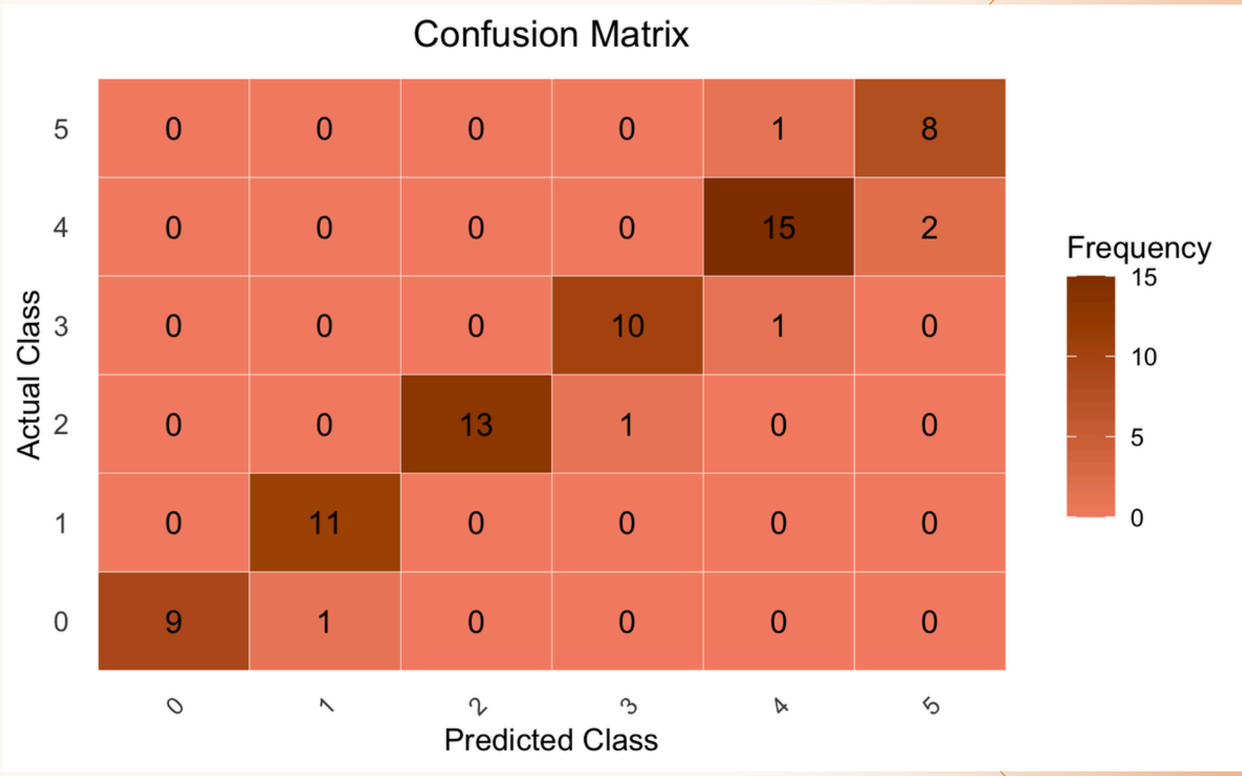


RANDOM FOREST ON PCA DATA

Only quantitative variables are chosen

CLASS	1	2	3	4	5	6
SENSITIVITY	0.9	1	0.92	1	0.82	0.88
SPECIFICITY	1	0.98	1	0.96	0.98	0.96

ACCURACY : 0.916



CLASSIFICATION TREE

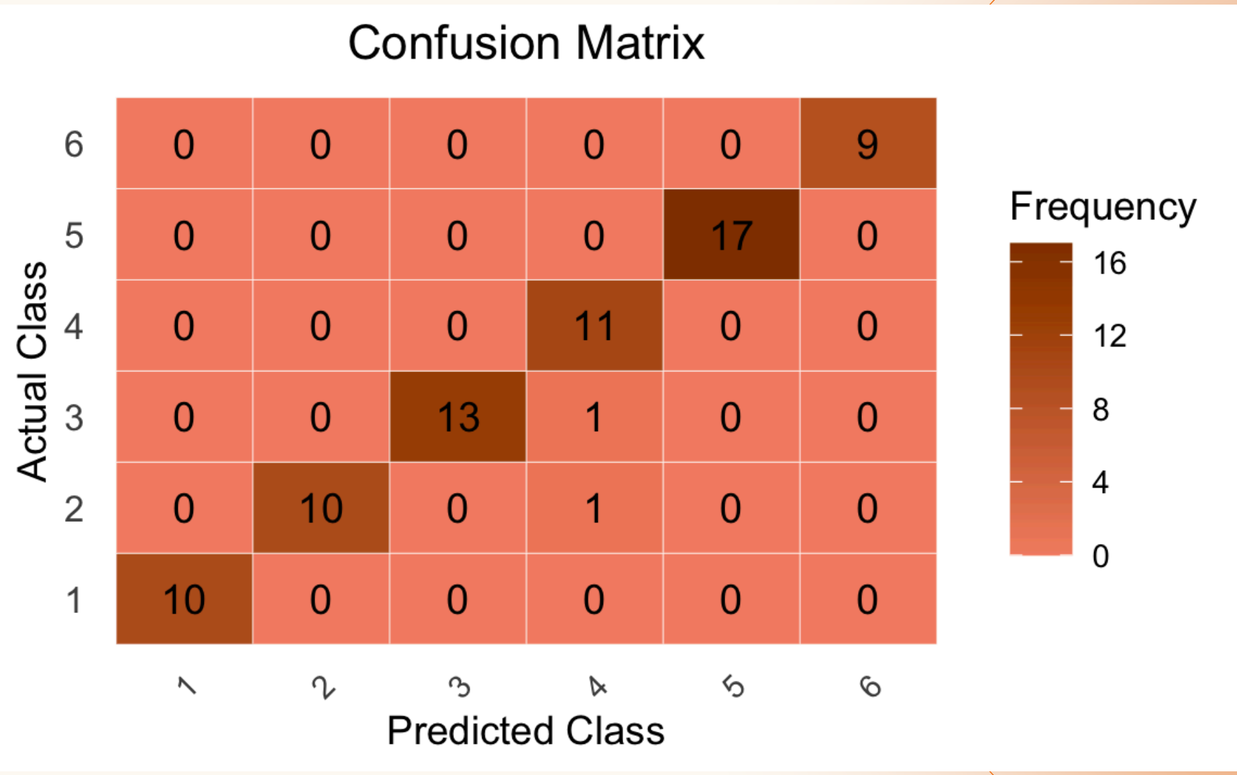
A classification tree is a type of decision tree used for categorical dependent variables. It is a predictive model which maps observations about an item to conclusions about the item's target value (class).

CLASSIFICATION TREE ON ORIGINAL DATA

Only quantitative variables are chosen

CLASS	1	2	3	4	5	6
SENSITIVITY	1	0.91	0.92	1	1	1
SPECIFICITY	1	1	1	0.97	1	1

ACCURACY : 0.972

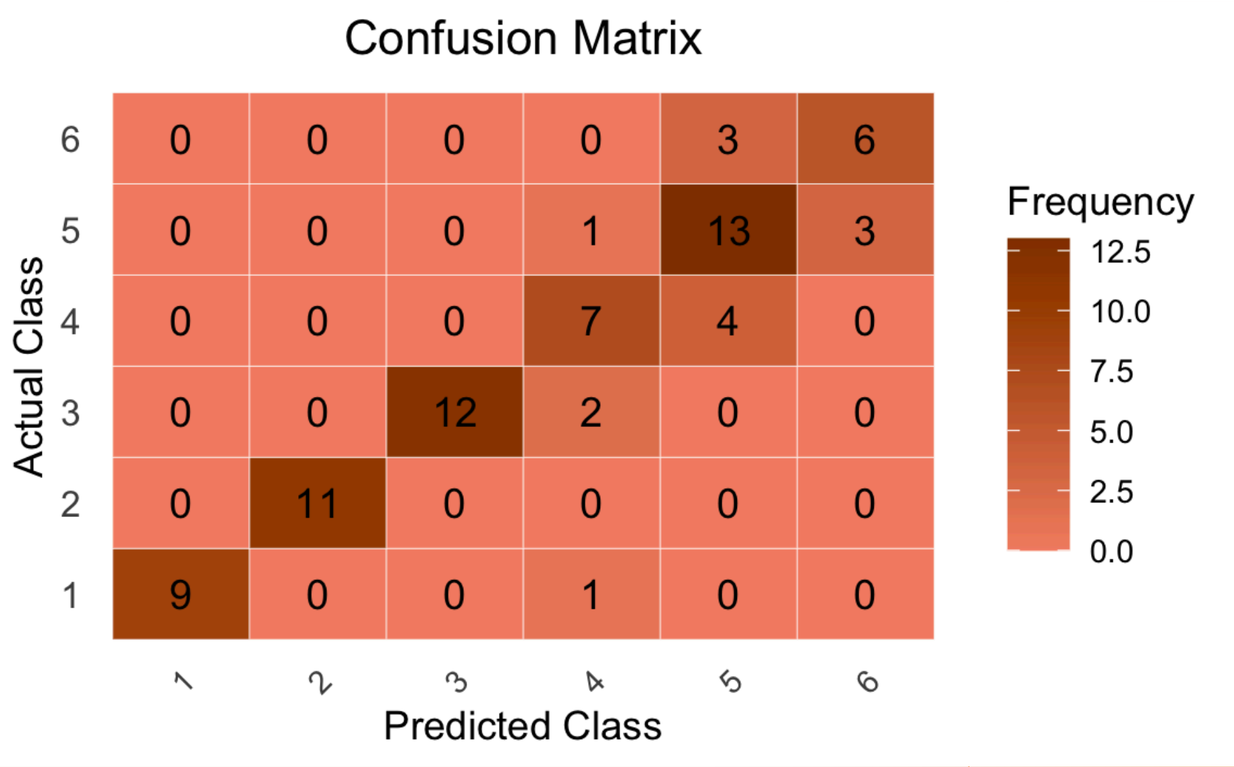


CLASSIFICATION TREE ON PCA DATA

Only quantitative variables are chosen

CLASS	1	2	3	4	5	6
SENSITIVITY	0.9	1	0.85	0.63	0.76	0.66
SPECIFICITY	1	1	1	0.93	0.87	0.95

ACCURACY : 0.805



K NEAREST NEIGHBOUR

KNN is a simple, non-parametric, lazy learning algorithm used for classification and regression. It classifies a data point based on how its neighbors are classified.

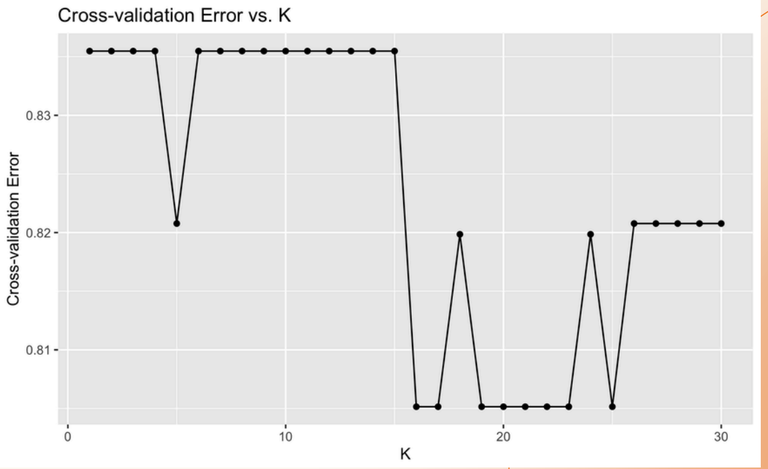
Key Parameters

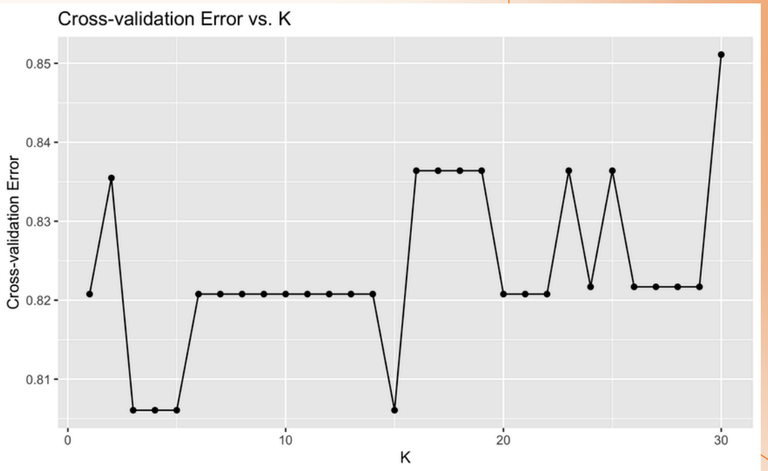
K in ORIGINAL DATA

Number of features considered at each split.
Balances model bias and variance.
Optimal value is often determined through cross-validation.

K in PCA DATA

Number of features considered at each split.
Balances model bias and variance.
Optimal value is often determined through cross-validation.



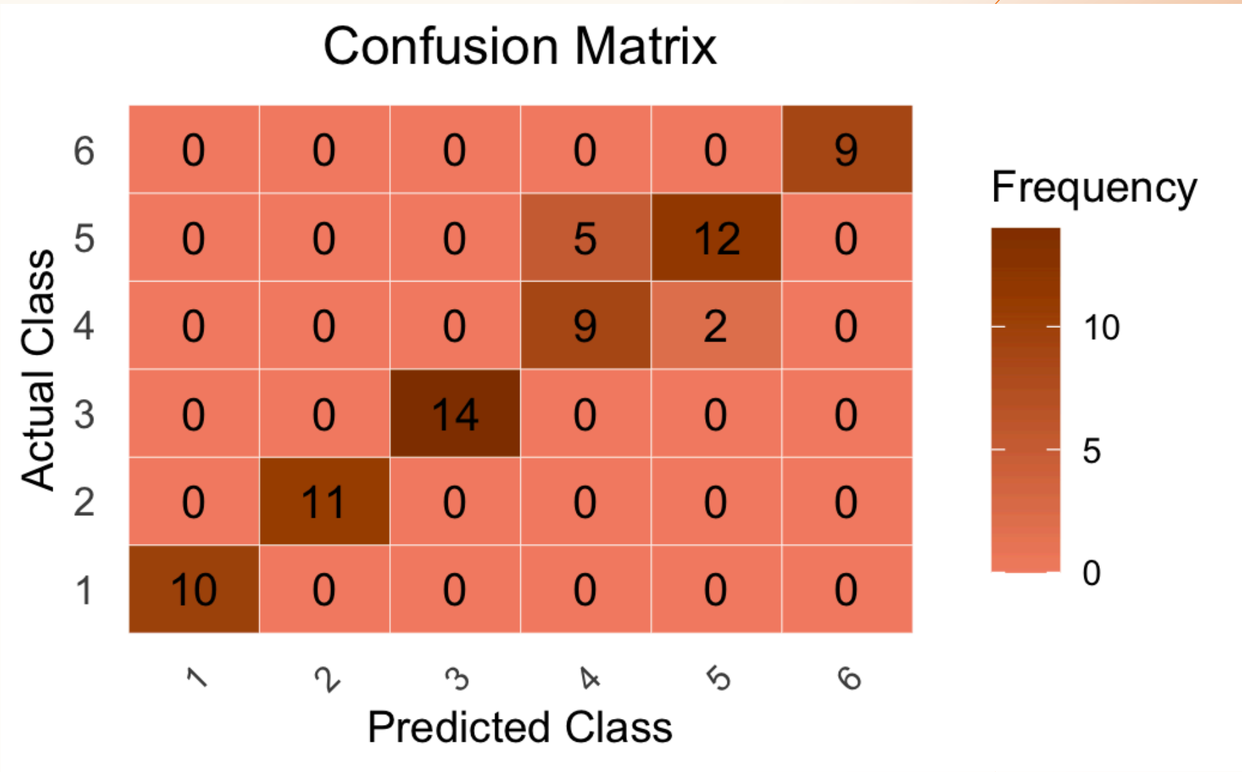


K NEAREST NEIGHBOUR ON ORIGINAL DATA

Only quantitative variables are chosen

CLASS	1	2	3	4	5	6
SENSITIVITY	1	1	1	0.81	0.70	1
SPECIFICITY	1	1	1	0.91	0.96	1

ACCURACY : 0.90



K NEAREST NEIGHBOUR ON PCA DATA

Only quantitative variables are chosen

CLASS	1	2	3	4	5	6
SENSITIVITY	1	1	1	1	0.82	0.88
SPECIFICITY	1	1	1	0.98	0.98	0.96

ACCURACY : 0.94

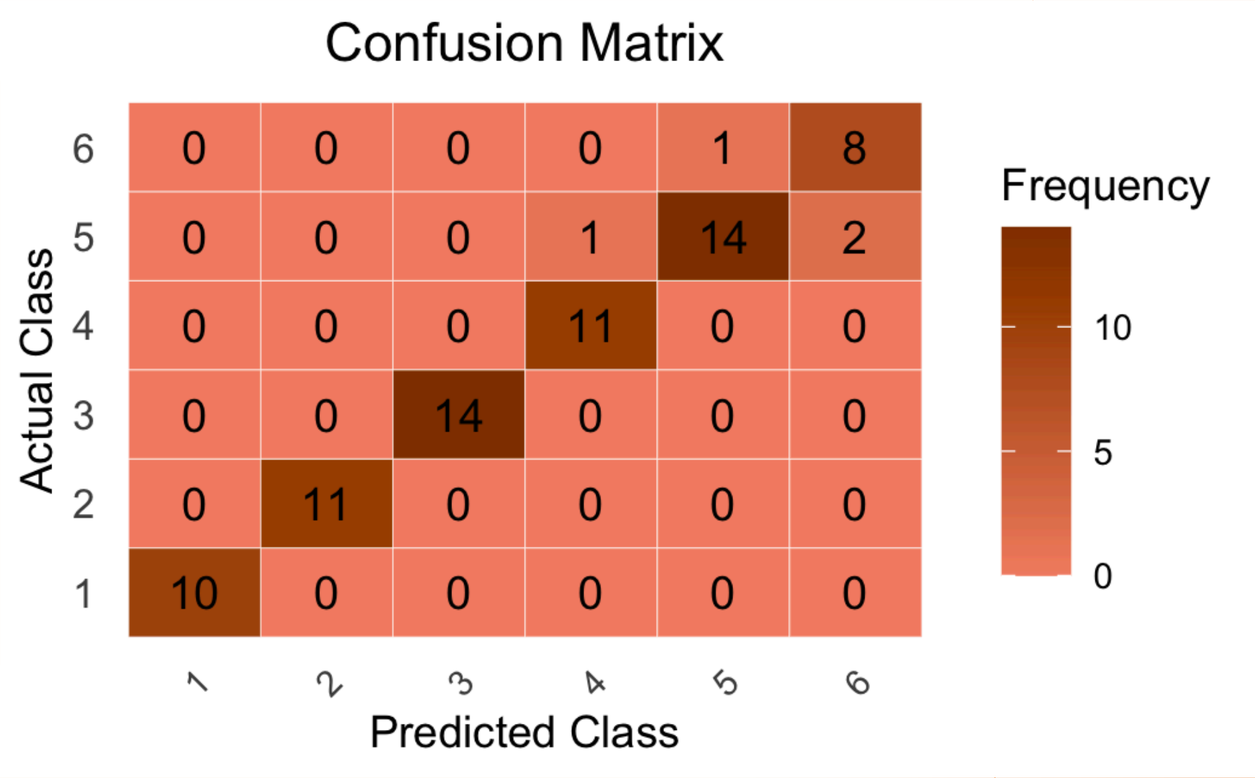
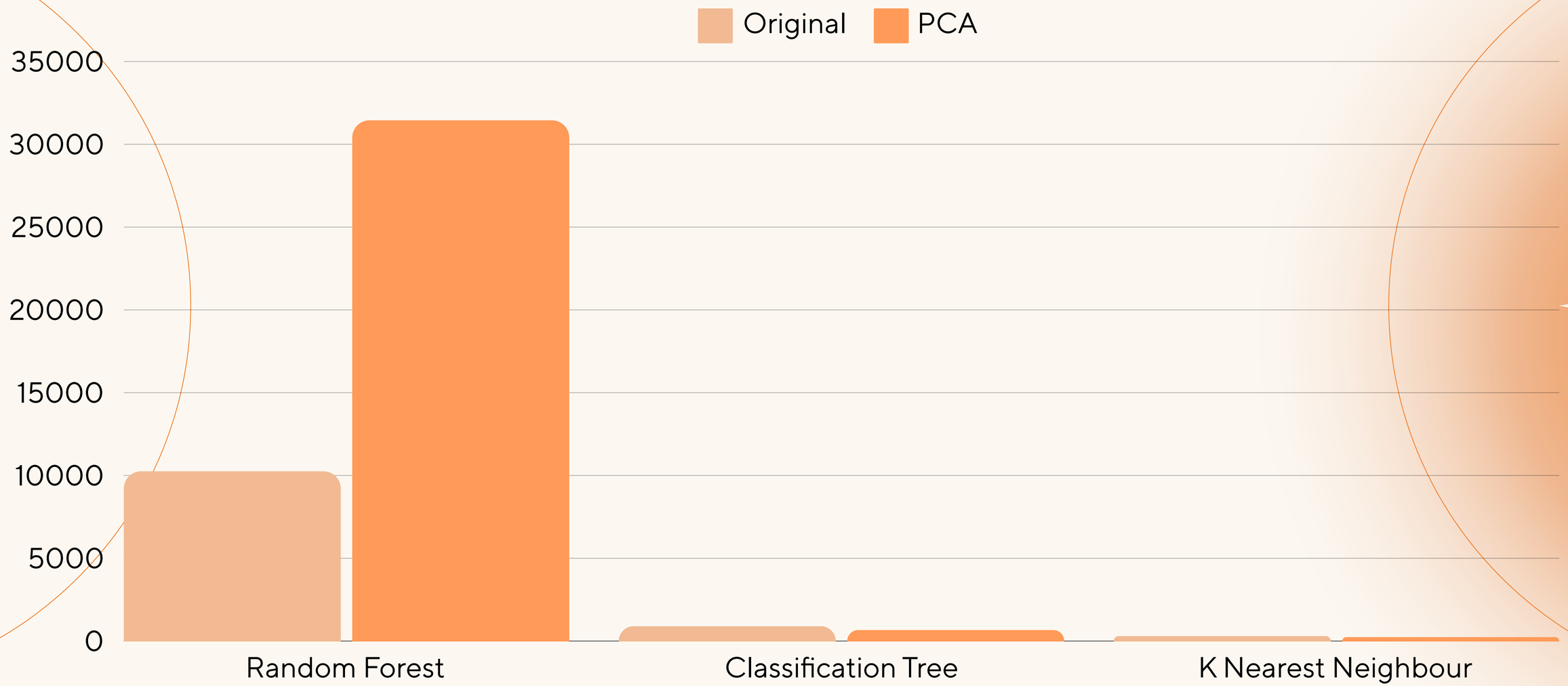


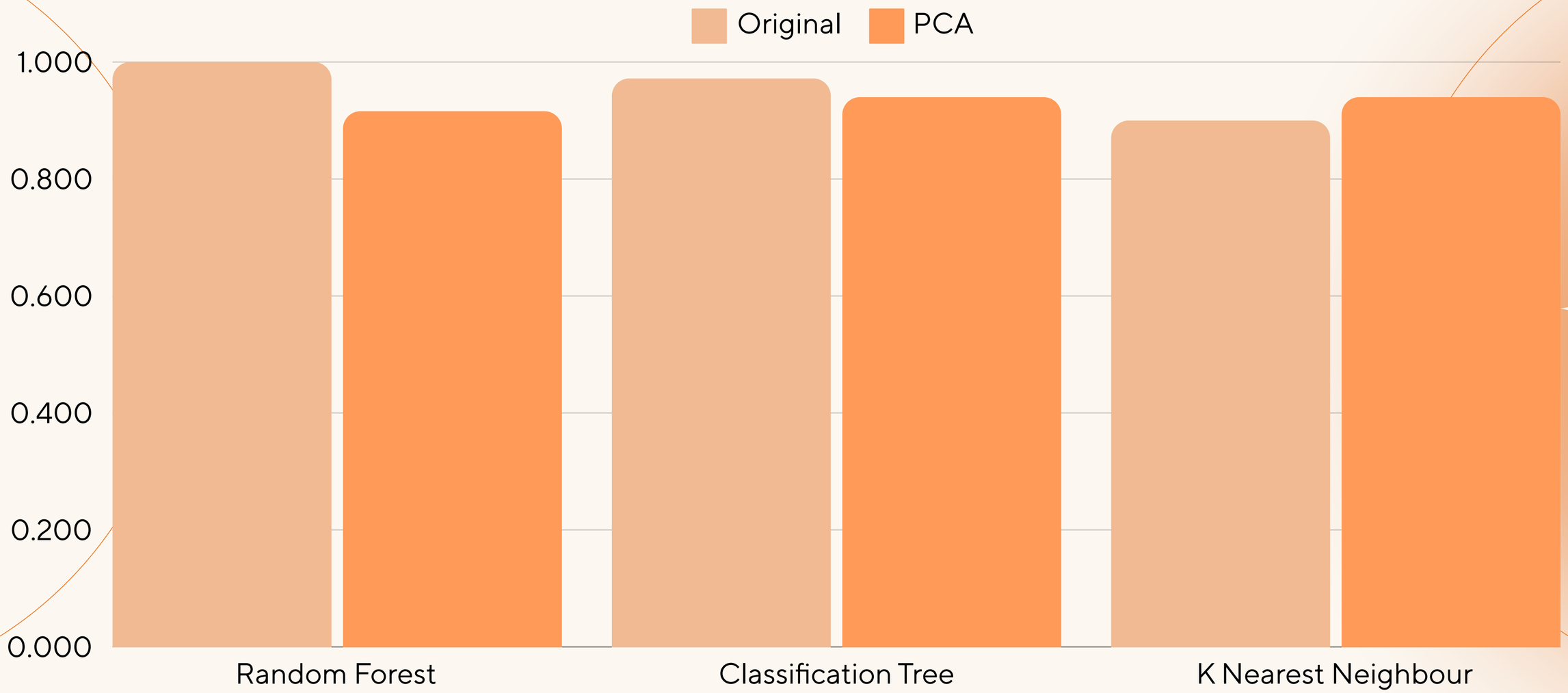
TABLE OF CONTENTS

Topics Covered

TIME COMPARISON



ACCURACY COMPARISON



CONCLUSIONS

Random Forest achieves the greatest performance among all the implemented models across both the datasets.

Comparing time and performance Classification tree seems like the best choice.

For KNN the PCA performed better than Numerical Covariates. This can be because of reduced noise or mitigation of overfitting.

Numerical Covariates VS PCA seems to do not much difference in time and accuracy in this case.

If the main aim of the research institution is to get the highest number of correctly classified gestures, then using random forest will for sure achieve this goal.