YANTING LIU, KAIYU SONG

THE RATING OF CHOCOLATE BAR

INTRODUCTION TO PROBLEM

DATA TYPE

- Company (String): Name of the company make the bar.
- Specific Bean Origin or Bar Name (String): The of the bar.
- REF (Numeric): A value linked to when the review was entered in the database. Higher = more recent.
- Review Date (Numeric): Date of publication of the review.
- Cocoa Percent (String): Cocoa percentage (darkness) of the chocolate bar being reviewed.
- Company Location (String): The location of company.
- Bean Type (String): Cocoa percentage (darkness) of the chocolate bar being reviewed.
- Broad Bean Origin (String): The broad geo-region of origin for the bean.
- **Rating (Numeric): The broad geo-region of origin for the bean.**

EXAMPLE OF SEVERAL ROW

Company	Bar Name	REF	Review Date	Cocoa Percent	Company Location	Bean Type	Broad Bean	**Rating**
A. Morin	Agua Grande	1876	2016	63%	France		Sao Tome	**3.75**
A. Morin	Carenero	1315	2014	70%	France	Criollo	Venezuela	**2.75**
Akesson's (Pralus)	Bali (west), Sukrama Family, Melaya area	636	2011	75%	Switzerland	Trinitario	Indonesia	**3.75**
Alexandre	La Dalia, Matagalpa	1944	2017	70%	Netherlands	Criollo, Trinitario	Nicaragua	**3.5**

ANALYZE AND TRANSFORM

- Turn string type data into binary int. (Dummy Variables)
- Drop unless data (REF/Review data)
- Split data into train (80%), test (10%) and valid (10%) data set.

TOOLS TO USE

- python3
- numpy
- sklearn
- pandas

METHOD TO USE

- Decision Tree / Random Forest:
  - No many leaves, few features.
- K-NN:
  - Dataset could be turned into numbers.
  - There are multiple labels, they aren’t linear separable.

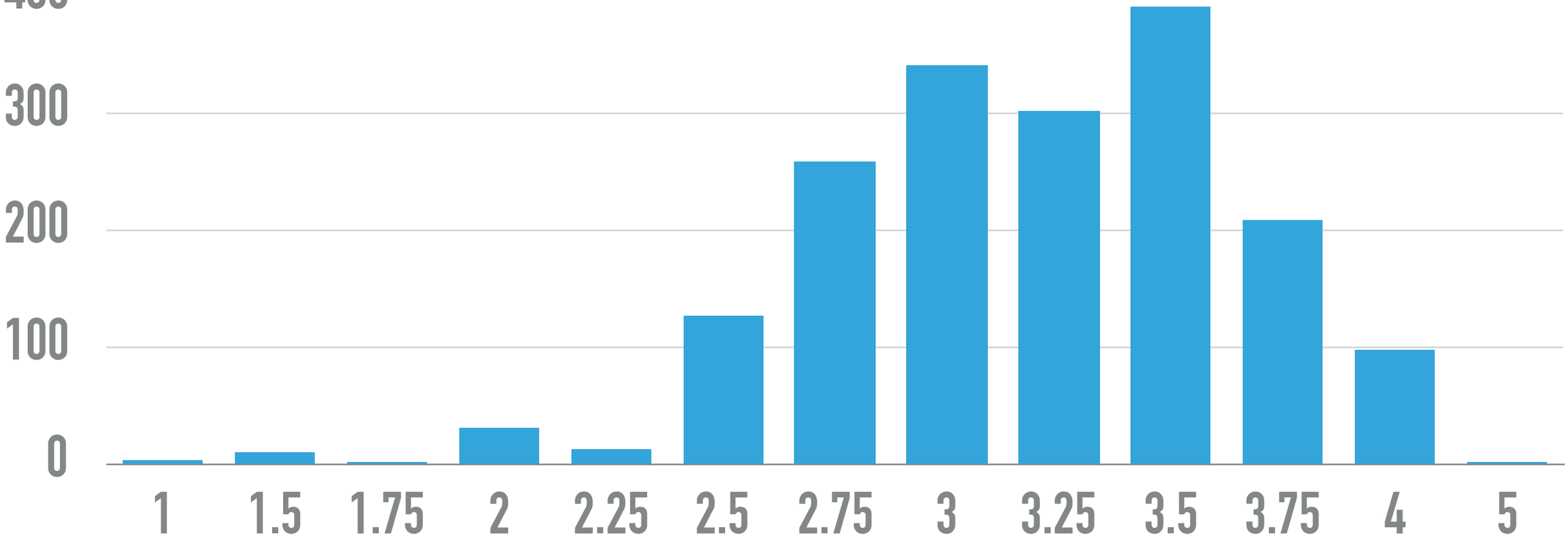
ACCURACY OF EACH METHOD

- Decision Tree: About 66% ~ 68%
- Random Forest: About 71% ~ 73%
- K-NN: About 69% ~ 70%

ACCURACY FOCUS ON RATING 2.75 ~ 3.75.

Problem We Found In Solving Problem

NO ENOUGH DATA



THANKS