Automated Spectral Classification of Two Spectral Libraries

Group Members

Akant Vats

Ruchi Pandey

Shrish

Project Supervisors

Prof. Ranjan Gupta

Prof. Harinder P. Singh

August 24, 2019

Introduction

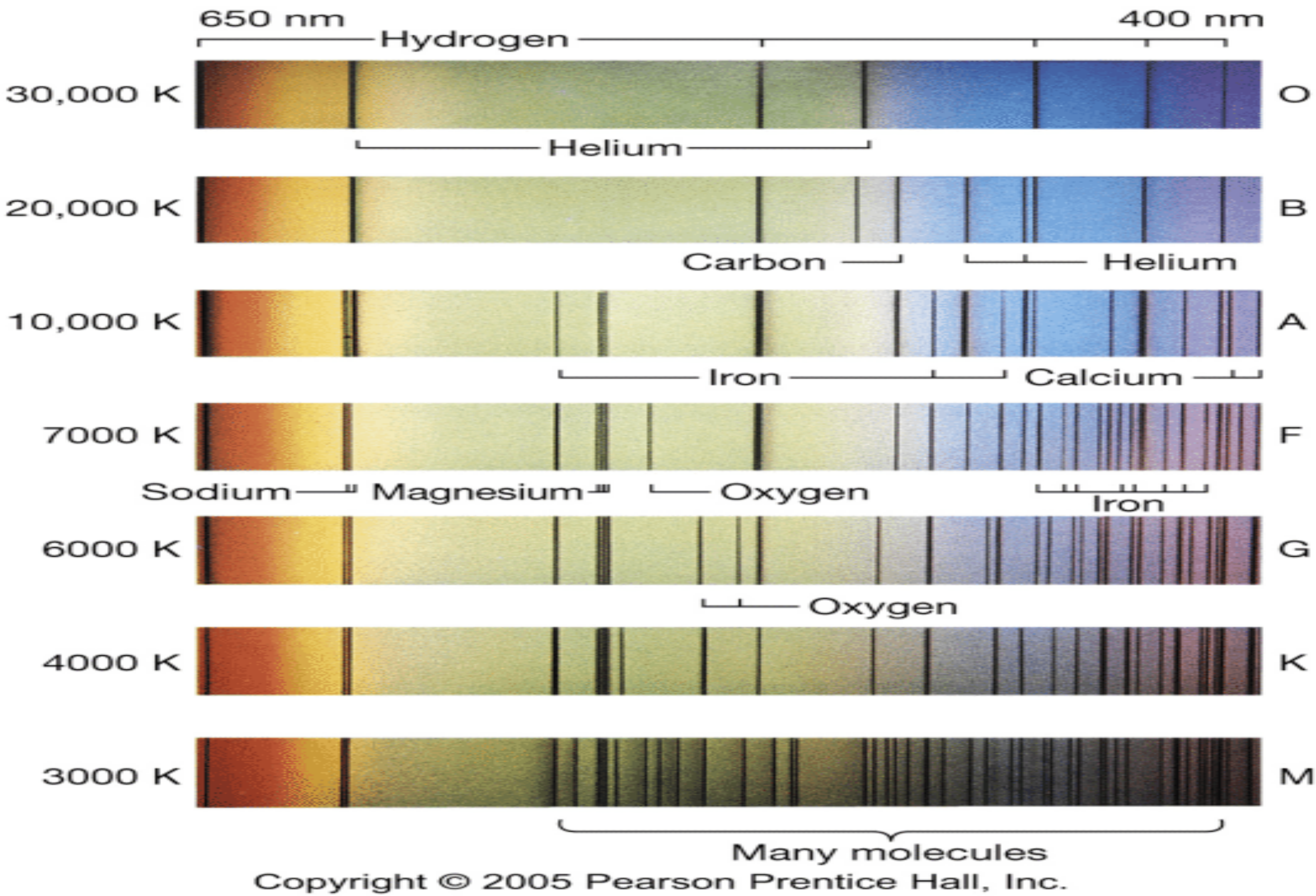
Stellar classification is the classification of stars based on their spectral characteristics. The absorption features present in stellar spectra allow us to divide stars into several spectral types depending on the temperature of the star.

Type	Color	Approximate Surface Temperature	Main Characteristics	Examples
O	Blue	> 25,000 K	Singly ionized helium lines either in emission or absorption. Strong ultraviolet continuum.	10 Lacertra
B	Blue	11,000 - 25,000	Neutral helium lines in absorption.	Rigel Spica
A	Blue	7,500 - 11,000	Hydrogen lines at maximum strength for A0 stars, decreasing thereafter.	Sirius Vega
F	Blue to White	6,000 - 7,500	Metallic lines become noticeable.	Canopus Procyon
G	White to Yellow	5,000 - 6,000	Solar-type spectra. Absorption lines of neutral metallic atoms and ions (e.g. once-ionized calcium) grow in strength.	Sun Capella
K	Orange to Red	3,500 - 5,000	Metallic lines dominate. Weak blue continuum.	Arcturus Aldebaran
M	Red	< 3,500	Molecular bands of titanium oxide noticeable.	Betelgeuse Antares

Within each **spectral type** there are significant variations in the strengths of the absorption lines, and each type has been subdivided into 10 sub-classes numbered 0 to 9.

Stars of a particular spectral type can differ widely in **luminosity** and must also be assigned a **luminosity class**. Luminosity classes are labeled with Roman numerals from I to V;
I are supergiant stars,
II are bright giants,
III are ordinary giants,
IV are subgiants, and
V are ordinary main sequence stars.

The complete spectral classification for a star is then given by specifying both the spectral class and the luminosity class.


650 nm
400 nm
Hydrogen
30,000 K
O
Helium
20,000 K
B
Carbon
Helium
10,000 K
A
Iron
Calcium
7000 K
F
Sodium
Magnesium
Oxygen
Iron
6000 K
G
Oxygen
4000 K
K
3000 K
M
Many molecules
Copyright © 2005 Pearson Prentice Hall, Inc.

A **spectral library** is a spectrophotometric library of the stars which covers the HR diagram and an important tool of astronomy to model stellar atmosphere and spectral classification.

Code Number = 1000.0 X A1 + 100.0 X A2 + (1.5 + 2 X A3), **(Gulati et al 1994)**

Where,

A1 is the main spectral type of the star (i.e., O to M as 1 to 7),

A2 is the sub-spectral type of the star (from 0.0 to 9.5),

A3 is the luminosity class of the star (i.e., I to V as 0 to 4)

Example: G9.5 V would be coded as 5959.5

Data Preprocessing

(Training Set)

Jacoby Library
(3510-7427 Ang)
(Resolution 4.5 Ang)

(Wavelength in the Steps of 1.4 Å ng)

↓

Interpolation (3600 - 7400 Ang)

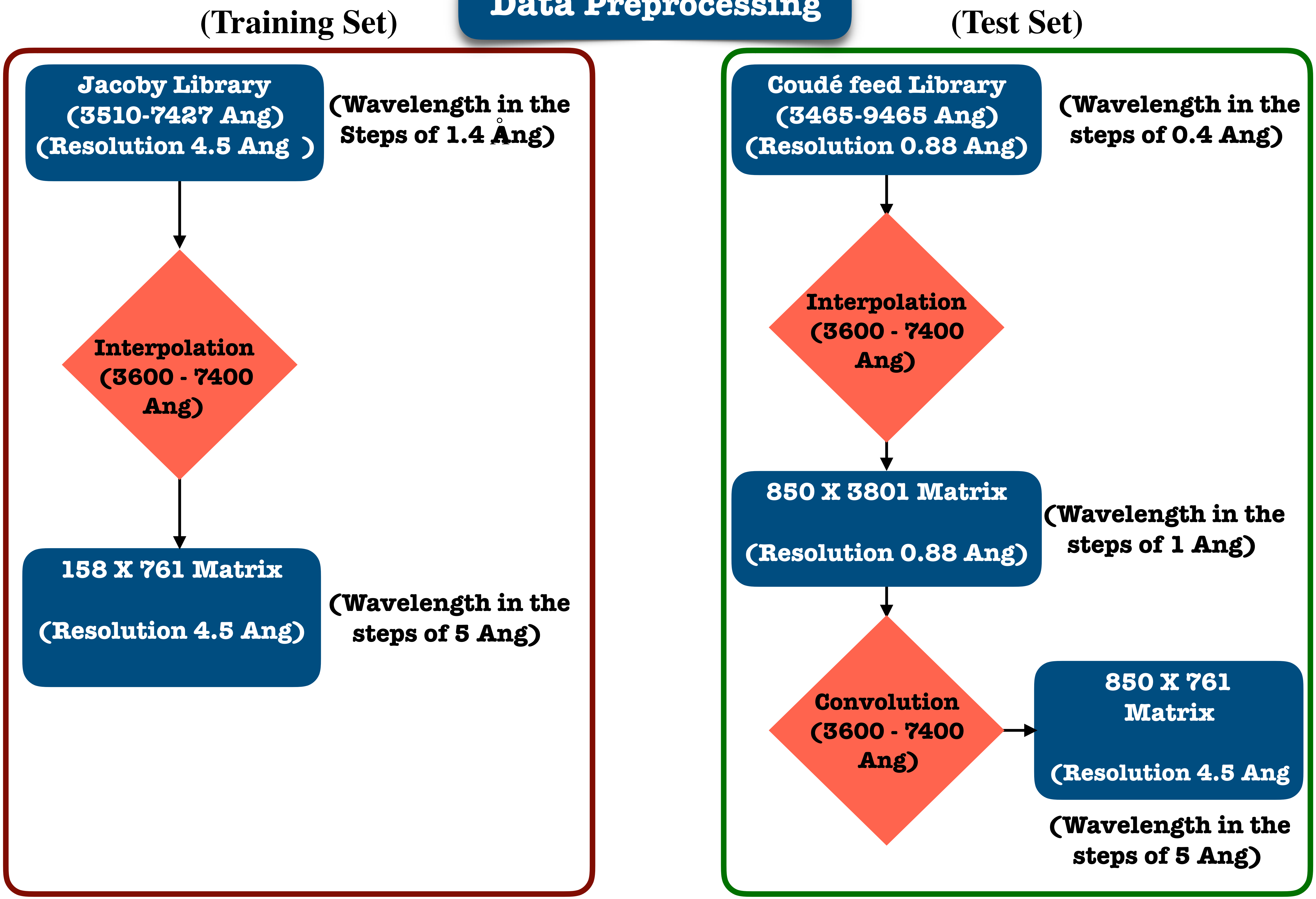
↓

158 X 761 Matrix

(Resolution 4.5 Ang)

(Wavelength in the steps of 5 Ang)

Data Preprocessing

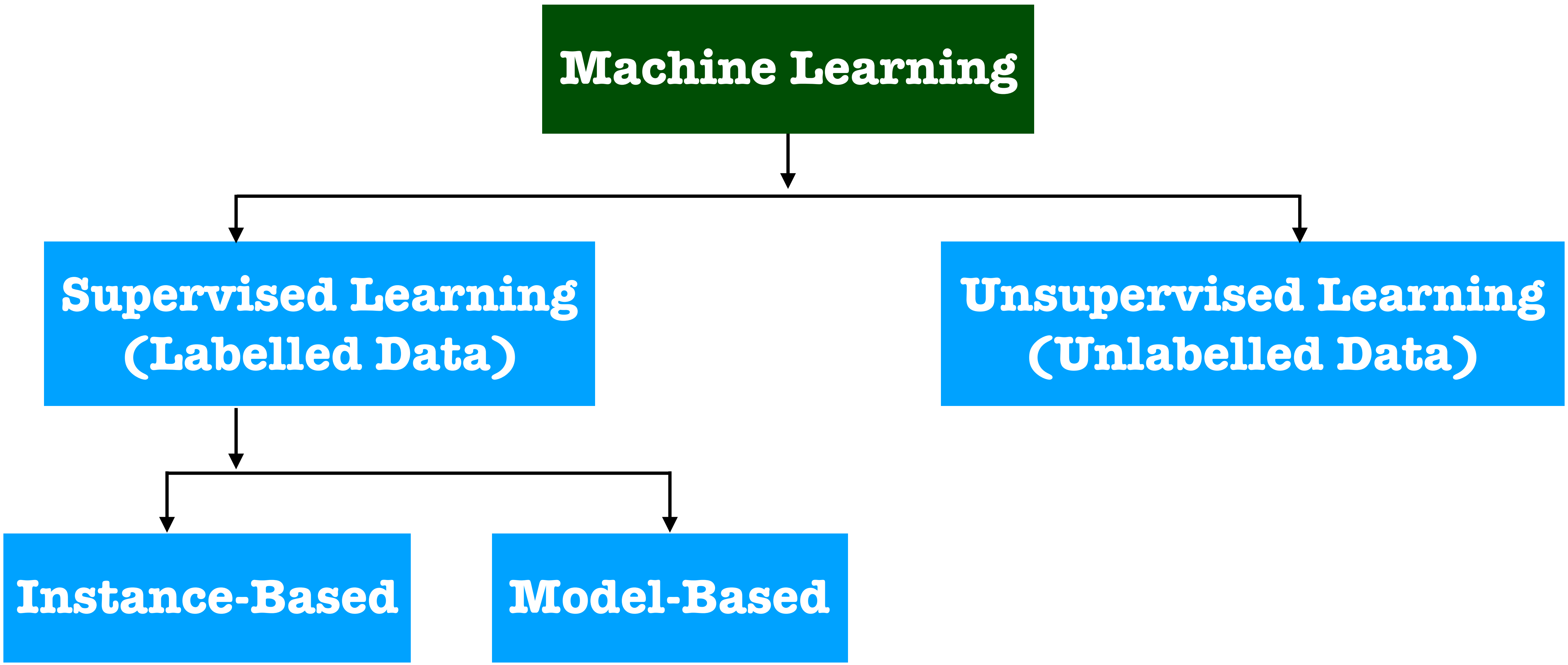


Methods

Machine Learning

- **Supervised Learning (Labelled Data)**
- **Unsupervised Learning (Unlabelled Data)**

Methods


Machine Learning
Supervised Learning (Labelled Data)
Unsupervised Learning (Unlabelled Data)
Instance-Based
Model-Based

Methods

Machine Learning

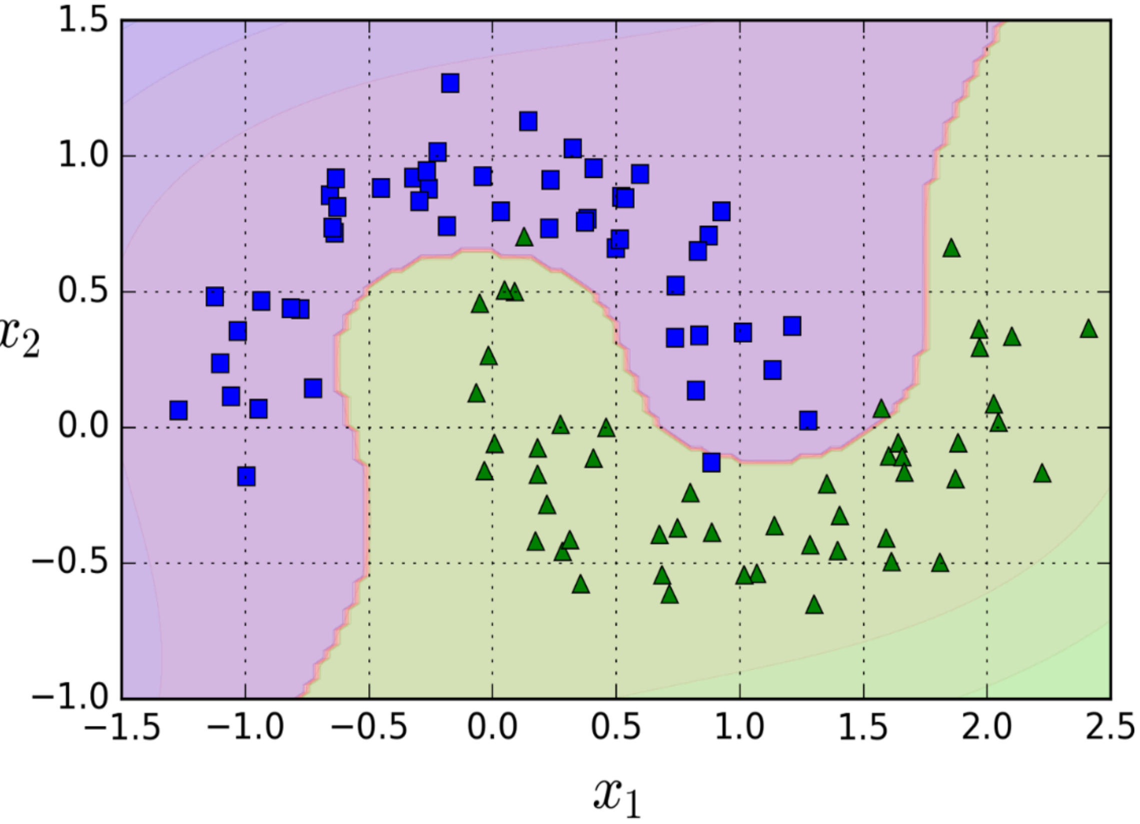
- Supervised Learning (Labelled Data)
  - Instance-Based
    - K Neighbor
  - Model-Based
    - Decision Tree
- Unsupervised Learning (Unlabelled Data)

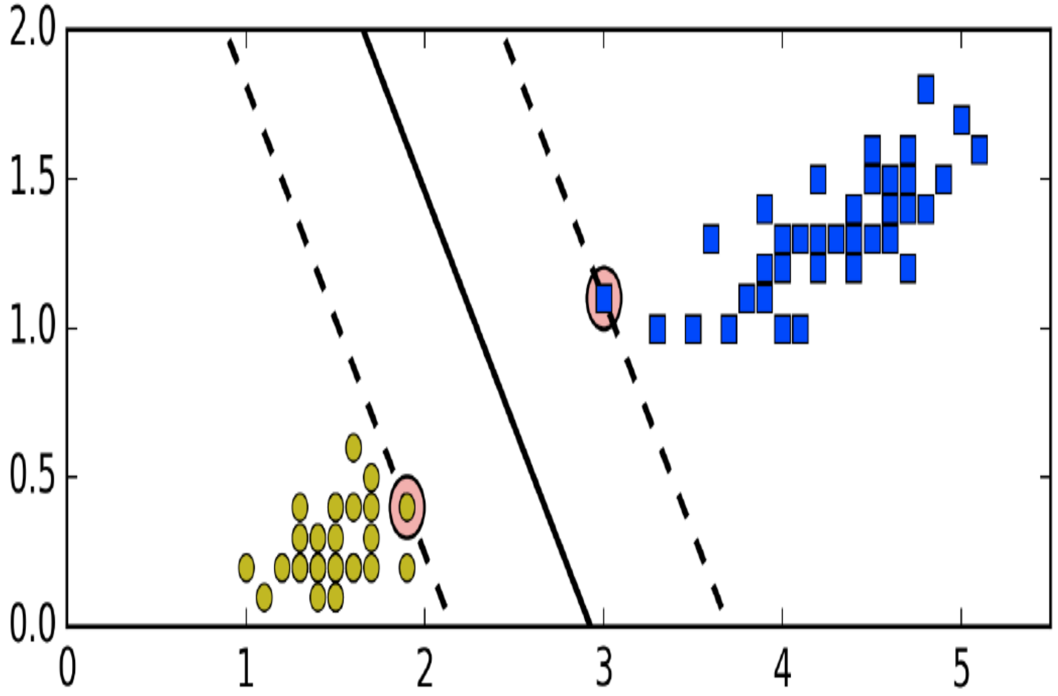
Methods

Machine Learning

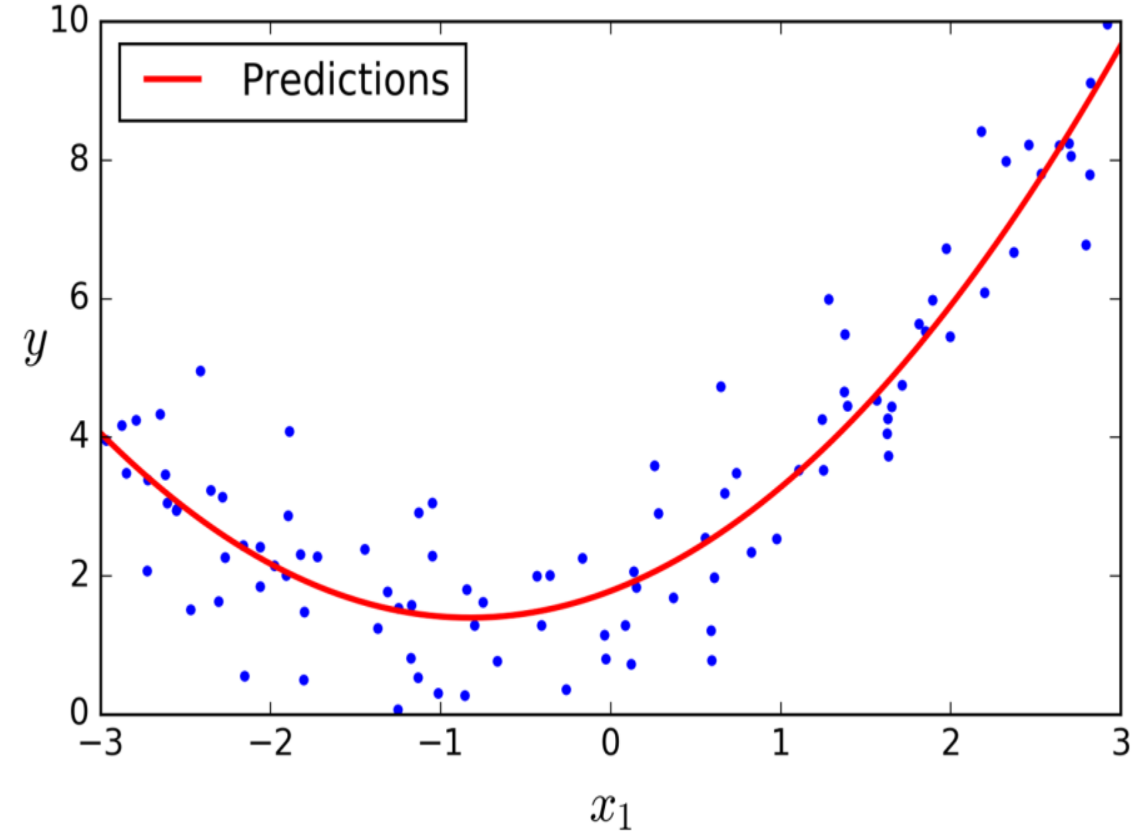
- **Supervised Learning (Labelled Data)**
  - **Instance-Based**
    - K Neighbor
  - **Model-Based**
    - Decision Tree
- **Unsupervised Learning (Unlabelled Data)**

Classification



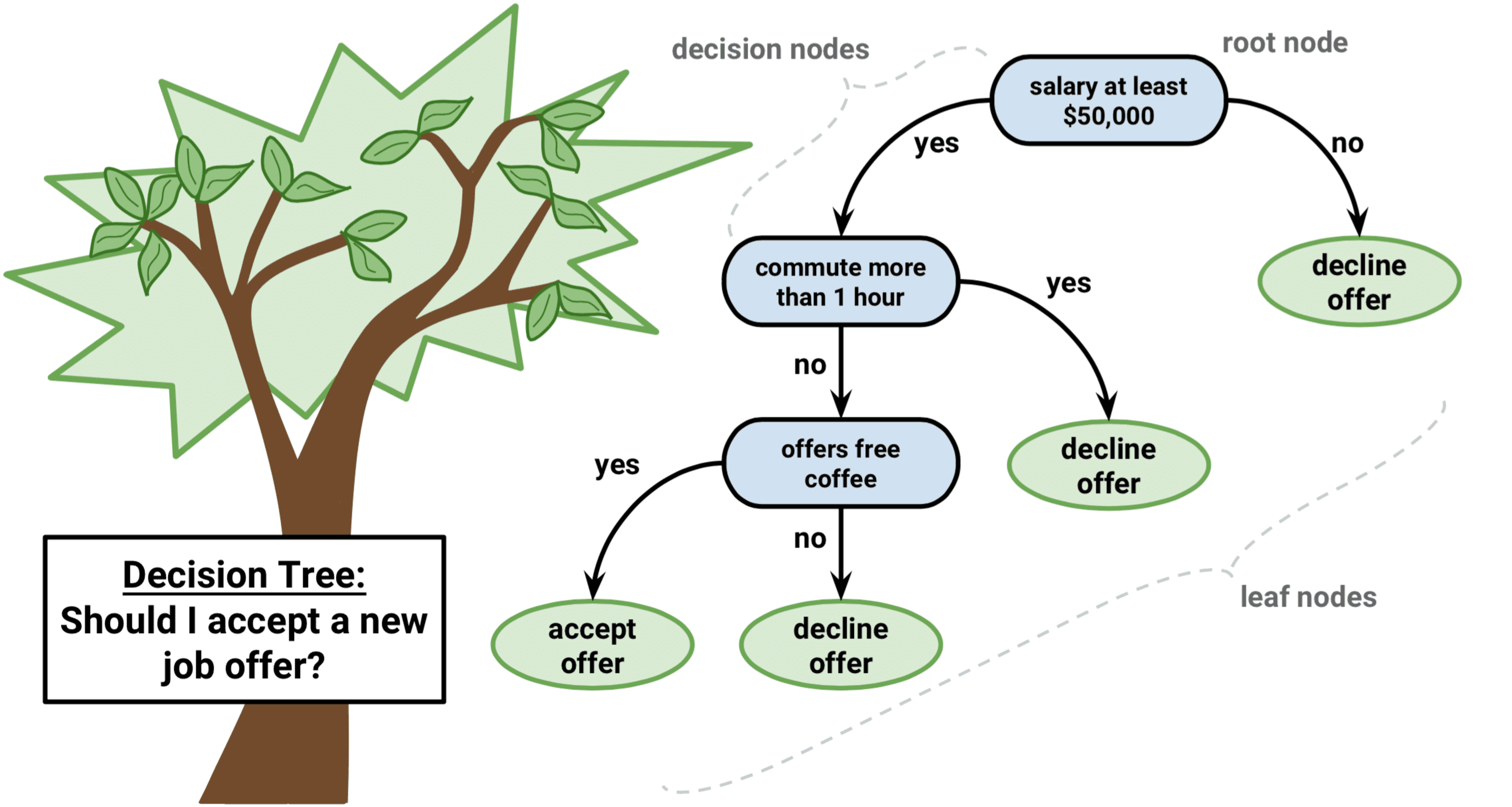


Regression



2-D	n-D
Lines	Hyper Plane
Curves	Hyper Surfaces

Classification



Cost functions

- **SGD Regression**

$$\mathrm{MSE}\left(\mathbf{X}, h_\theta\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T \cdot \mathbf{x}^{(i)} - y^{(i)}\right)^2$$

- **Logistic Regression**

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} log\left(\hat{p}^{(i)}\right) + \left(1 - y^{(i)}\right) log\left(1 - \hat{p}^{(i)}\right)\right]$$

- **Decision Tree Classification**

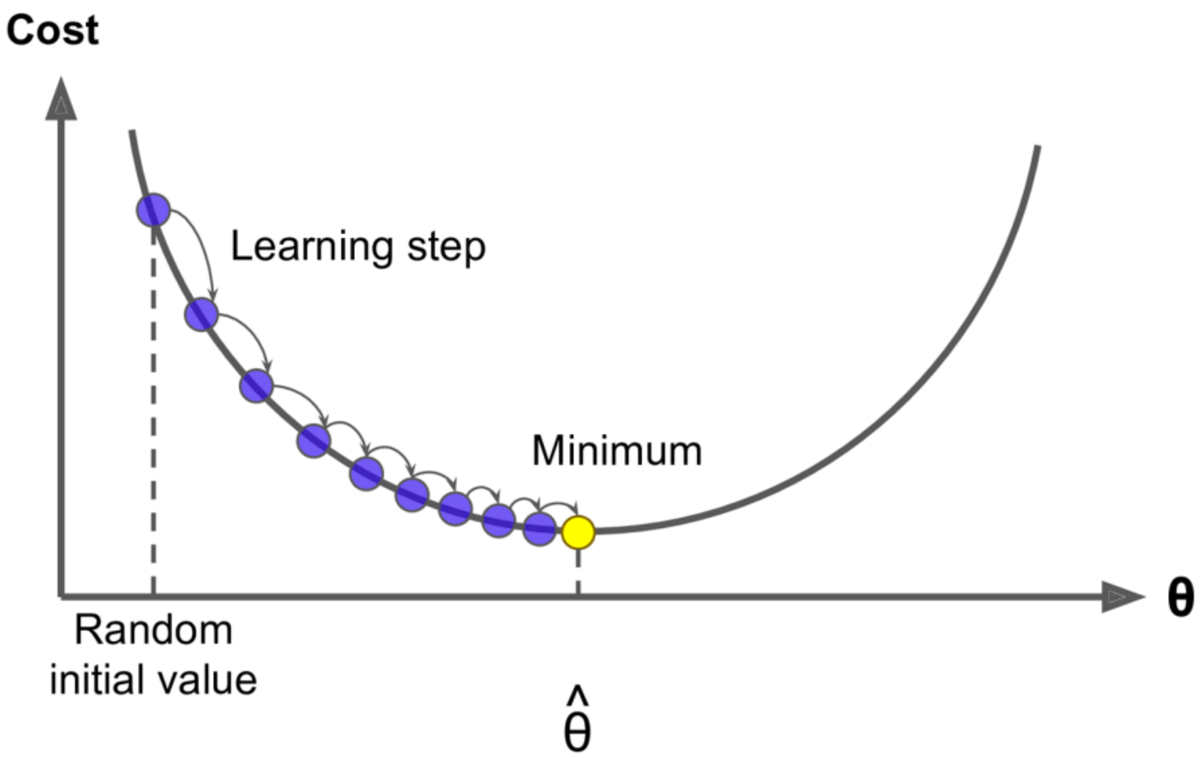
$$J\left(k, t_k\right) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

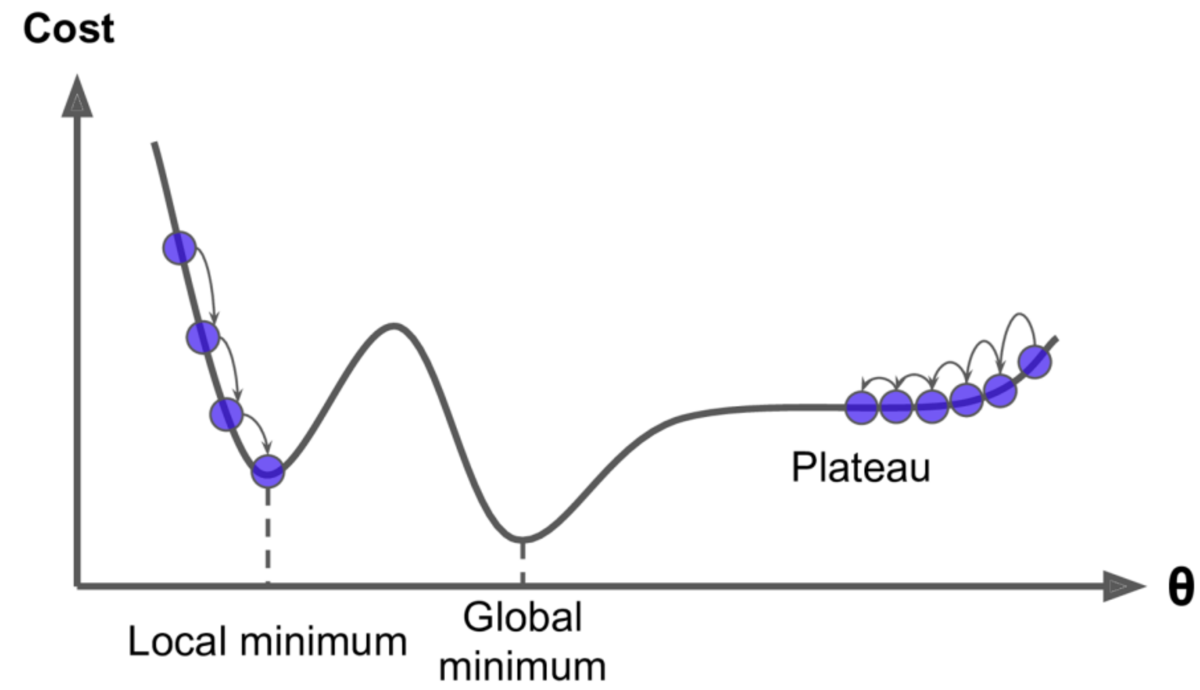
$$\text{where} \begin{cases} G_{\text{left/right}} \text{ measures the impurity of the left/right subset,} \\ m_{\text{left/right}} \text{ is the number of instances in the left/right subset.} \end{cases}$$

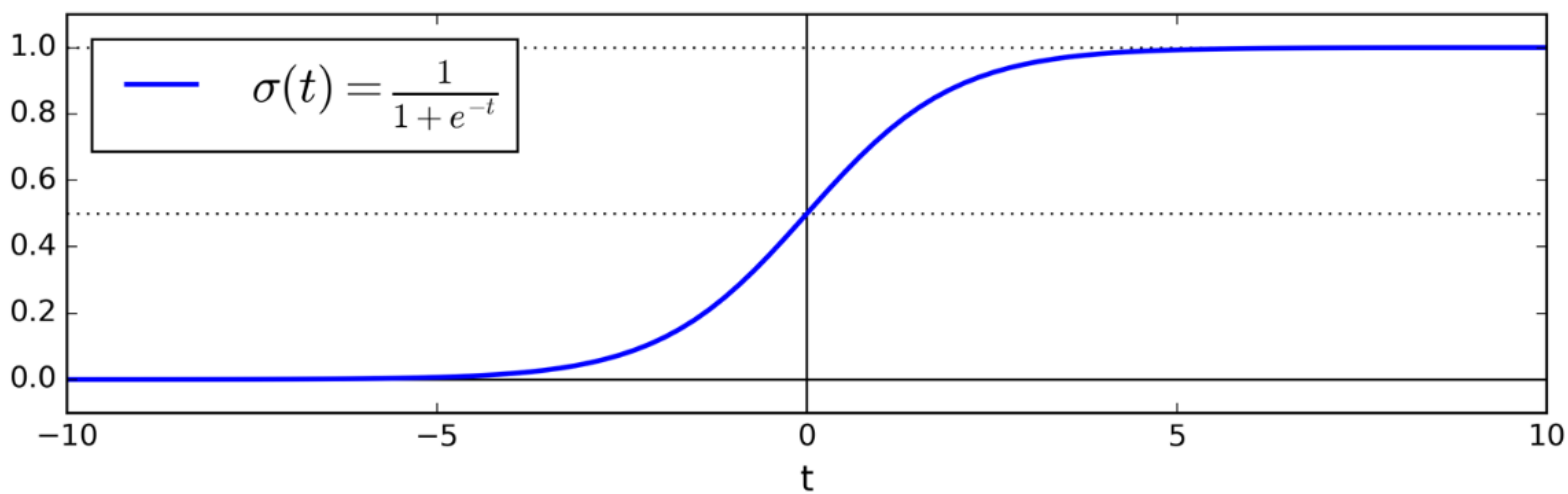
- **Decision Tree Regression**

$$J\left(k, t_k\right) = \frac{m_{\text{left}}}{m} \mathrm{MSE}_{\text{left}} + \frac{m_{\text{right}}}{m} \mathrm{MSE}_{\text{right}} \quad \text{where} \begin{cases} \mathrm{MSE}_{\text{node}} = \sum_{i \in \text{node}} \left(\hat{y}_{\text{node}} - y^{(i)}\right)^2 \\ \hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)} \end{cases}$$

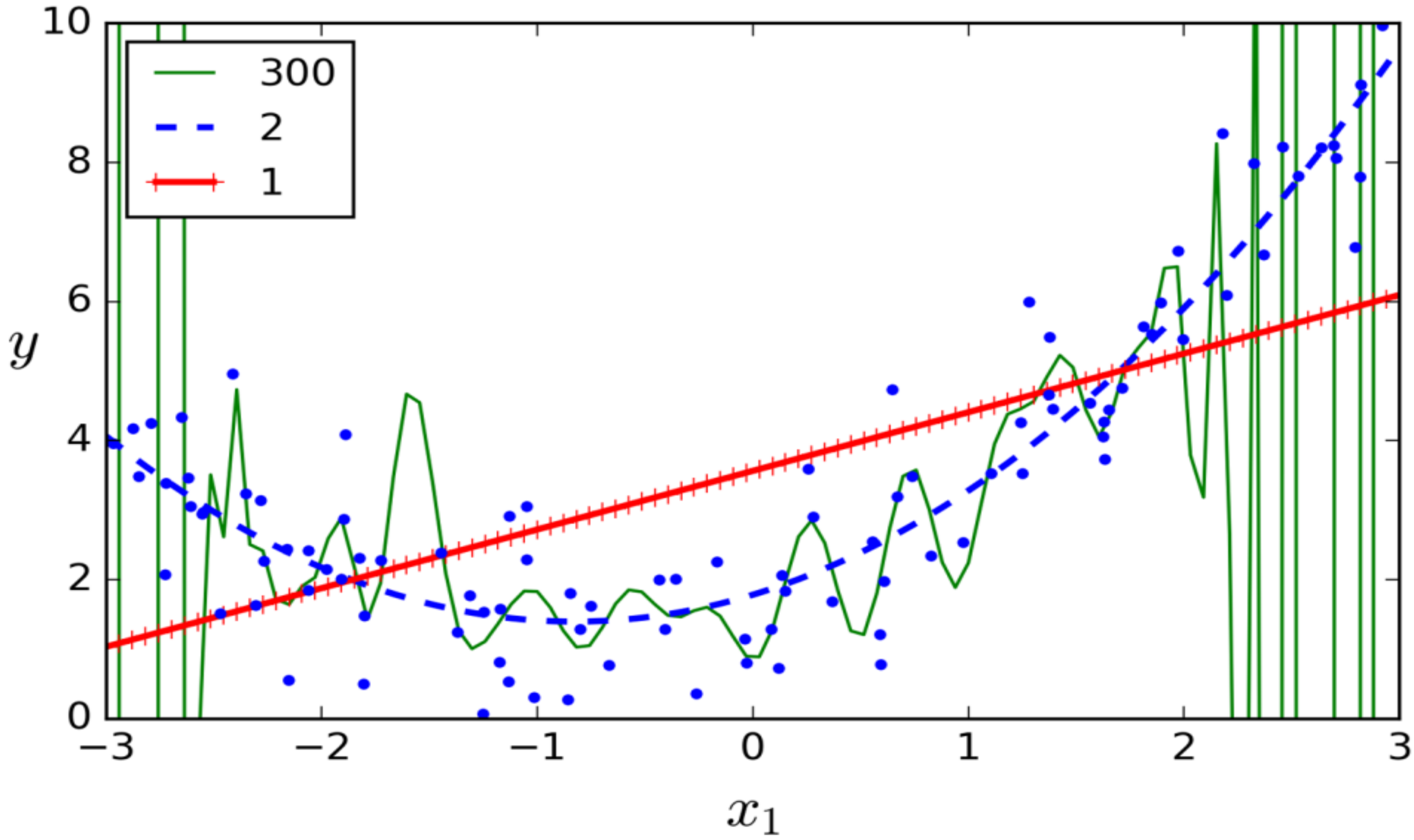
Optimization of Cost Functions







Non-linear Regression



Machine Learning Tools

1. SGD Classifier
2. SGD Regressor
3. Decision Tree Classifier
4. Decision Tree Regressor
5. Random Forest Classifier
6. Random Forest Regressor
7. Support Vector Machine
8. Gaussian Naive Bayes
9. KN Classifier
10. KN Regressor
11. ANN Classifier
12. ANN Regressor

Results

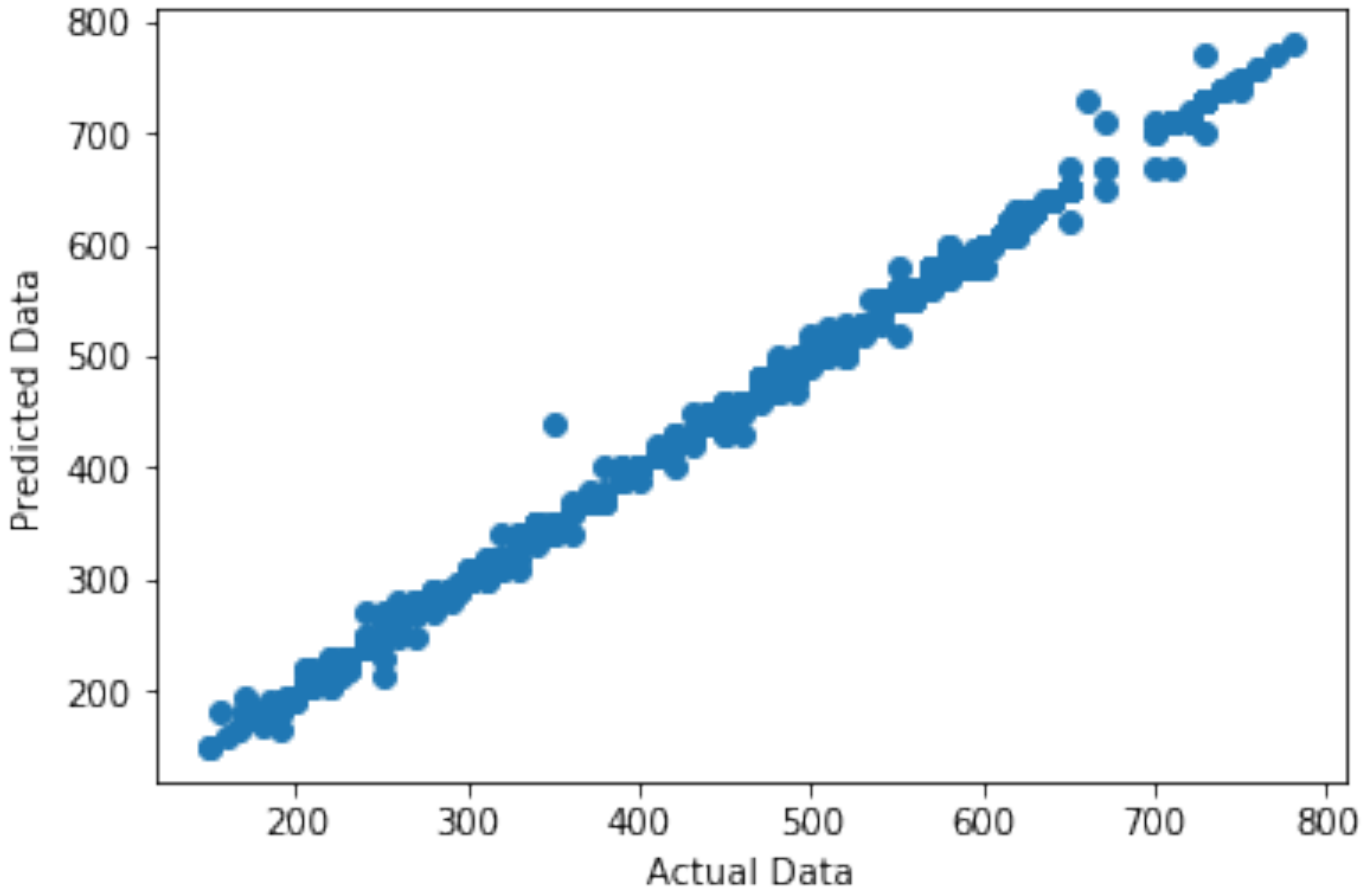


Figure showing scatter plot of test_predicted vs test_actual for ANN classification.

Spectral Class, Train data = 158, Test data = 850

	Slope		Intercept		Subclass Error		R	
SGD Classifier	0.57	0.47	122	214	13.2	11.4	0.66	0.57
SGD Regresser	0.96	0.90	9.44	40	3.3	8.1	0.98	0.83
Random Forest Classifier	0.98	0.91	4.5	40.3	0.7	3.9	0.99	0.95
Random Forest Regresser	0.98	0.89	5.8	46.0	0.9	3.0	0.99	0.97
KN Classifier	0.97	0.90	0.16	37.6	2.1	3.0	0.99	0.97
KN Regresser	0.99	0.93	2.33	32.8	1.5	2.4	0.99	0.98
Decission Tree Classifier	1.00	0.88	-0.6	62.2	2.7	4.0	0.98	0.95
Decission Tree Regresser	1.00	0.95	0.0	32.2	0.0	24.7	1.00	0.98
ANN Classifier	0.99	0.94	0.8	27.9	0.2	1.6	0.99	0.98
ANN Regressor	0.99	0.92	1.2	44.2	1.0	1.5	0.99	0.96
Naive Bayes	0.99	0.90	1.4	49.3	1.0	1.7	0.99	0.97
SVM	0.99	0.49	3.72	275	0.5	10	0.99	0.65

■ **Train Data** ■ **Test Data**

Spectral Class, Train data = 158, Test data = 850

	Slope		Intercept		Subclass Error		R	
SGD Classifier	0.57	0.47	122	214	13.2	11.4	0.66	0.57
SGD Regresser	0.96	0.90	9.44	40	3.3	8.1	0.98	0.83
Random Forest Classifier	0.98	0.91	4.5	40.3	0.7	3.9	0.99	0.95
Random Forest Regresser	0.98	0.89	5.8	46.0	0.9	3.0	0.99	0.97
KN Classifier	0.97	0.90	0.16	37.6	2.1	[illegible]	0.99	0.97
KN Regresser	0.99	0.93	2.33	32.8	[illegible]	[illegible]	0.99	0.98
Decission Tree Classifier	1.00	0.88	-0.6	62.2	[illegible]	[illegible]	0.98	0.95
Decission Tree Regresser	1.00	0.95	0.0	32.2	0.0	24.7	1.00	0.98
ANN Classifier	0.99	0.94	0.8	27.9	0.2	1.6	0.99	0.98
ANN Regressor	0.99	0.92	1.2	44.2	1.0	1.5	0.99	0.96
Naive Bayes	0.99	0.90	1.4	49.3	1.0	1.7	0.99	0.97
SVM	0.99	0.49	3.72	275	0.5	10	0.99	0.65

■ Train Data ■ Test Data

Best Fit

Spectral Class, Train data = 158, Test data = 850

	Slope		Intercept		Subclass Error		R	
SGD Classifier	0.57	0.47	122	214	13.2	11.4	0.66	0.57
SGD Regresser	0.96	0.90	9.44	40	3.3	8.1	0.98	0.83
Random Forest Classifier	0.98	0.91	4.5	40.3	0.7	3.9	0.99	0.95
Random Forest Regresser	0.98	0.89	5.8	46.0	0.9	[illegible]	0.99	0.97
KN Classifier	0.97	0.90	0.16	[illegible]	[illegible]	[illegible]	0.99	0.97
KN Regresser	0.99	0.93	2.33	[illegible]	[illegible]	[illegible]	0.99	0.98
Decission Tree Classifier	1.00	0.88	-0.6	62.2	2.7	4.0	0.98	0.95
Decision Tree Regresser	1.00	0.95	0.0	32.2	0.0	24.7	1.00	0.98
ANN Classifier	0.99	0.94	0.8	27.9	0.2	1.6	0.99	0.98
ANN Regressor	0.99	0.92	1.2	44.2	1.0	1.5	0.99	0.96
Naive Bayes	0.99	0.90	1.4	49.3	1.0	1.7	0.99	0.97
SVM	0.99	0.49	3.72	275	0.5	10	0.99	0.65

Over Fit

■ Train Data ■ Test Data

Randomly Mixed Data, Train data = 80%, Test data = 20%

	Slope		Intercept		Subclass Error		R	
SGD Classifier	0.80	0.78	106	116	5.5	5.9	0.92	0.90
SGD Regresser	0.97	0.96	7.5	11	2.6	2.7	0.98	0.98
Random Forest Classifier	0.99	1.00	1.1	-1.7	0.4	1.6	0.99	0.99
Random Forest Regresser	0.99	0.99	3.1	3.7	0.8	1.4	0.99	0.99
KN Classifier	0.99	1.00	0.6	-1.5	1.6	1.7	0.99	0.99
KN Regresser	0.99	0.99	1.7	0.4	1.3	1.4	0.99	0.99
Decission Tree Classifier	0.97	0.97	11	16	3.0	4.1	0.97	0.95
Decision Tree Regresser	0.99	0.99	0.0	1.9	0.0	2.0	0.99	0.99
ANN Classifier	0.99	0.99	1.1	2.1	0.5	1.1	0.99	0.99
ANN Regressor	0.99	1.00	4.4	0.7	1.4	1.6	0.99	0.99
Naive Bayes	0.99	0.96	1.4	15	2.0	3.1	0.98	0.97
SVM	0.99	0.88	3.0	50	0.6	4.7	0.99	0.94

■ Train Data ■ Test Data

Spectral Type

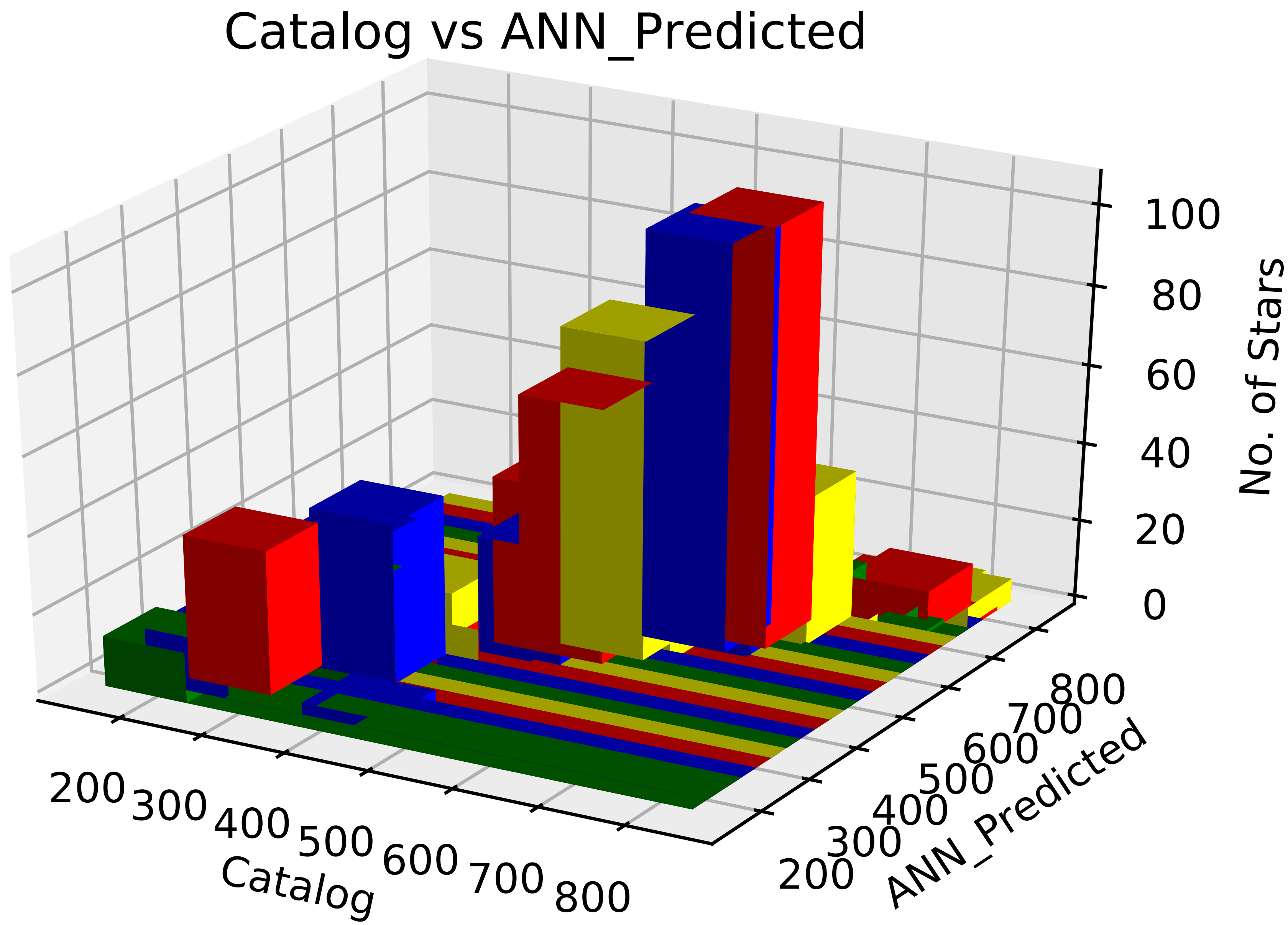


Figure showing 3D plot of test_predicted vs test_actual with their frequencies

Luminosity Type

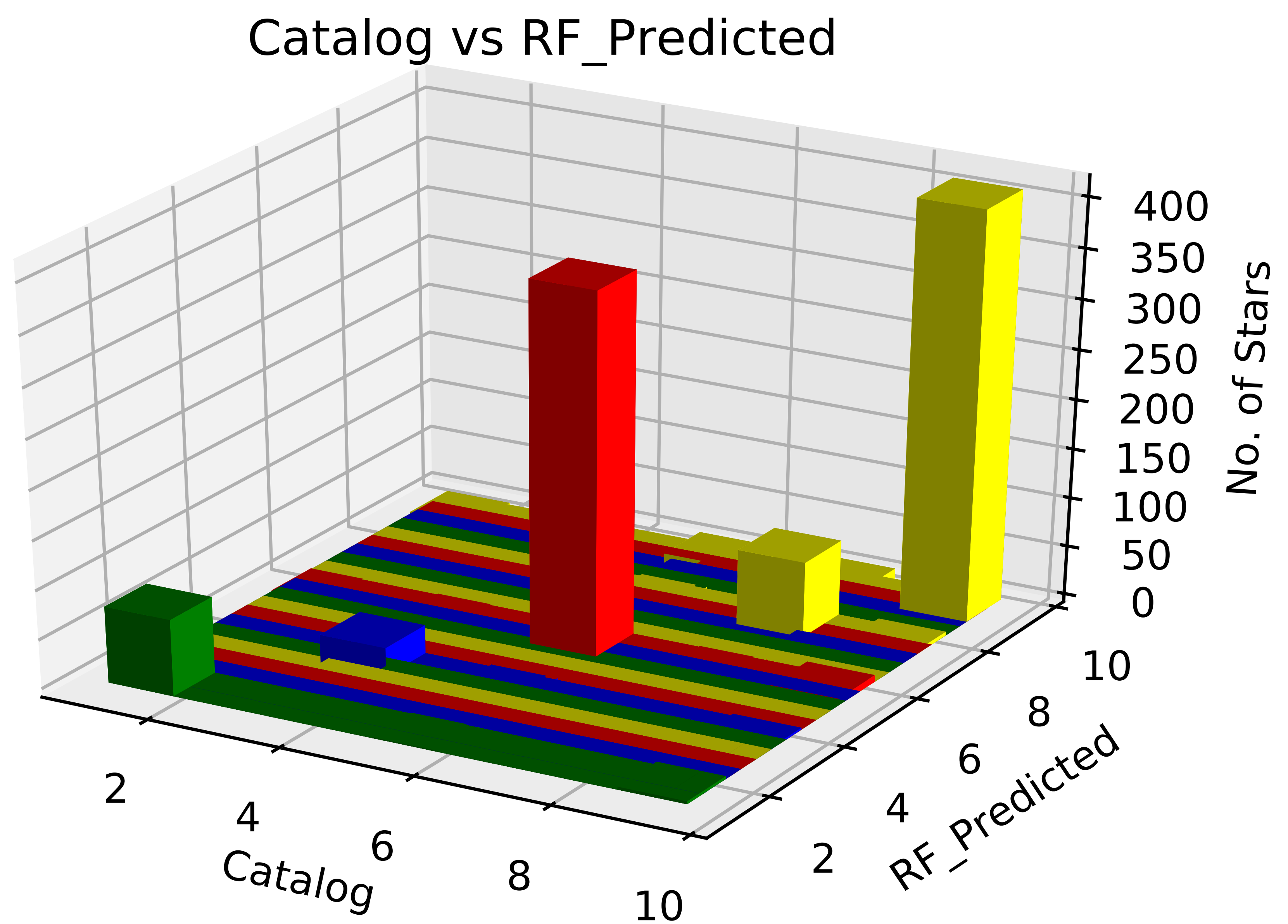
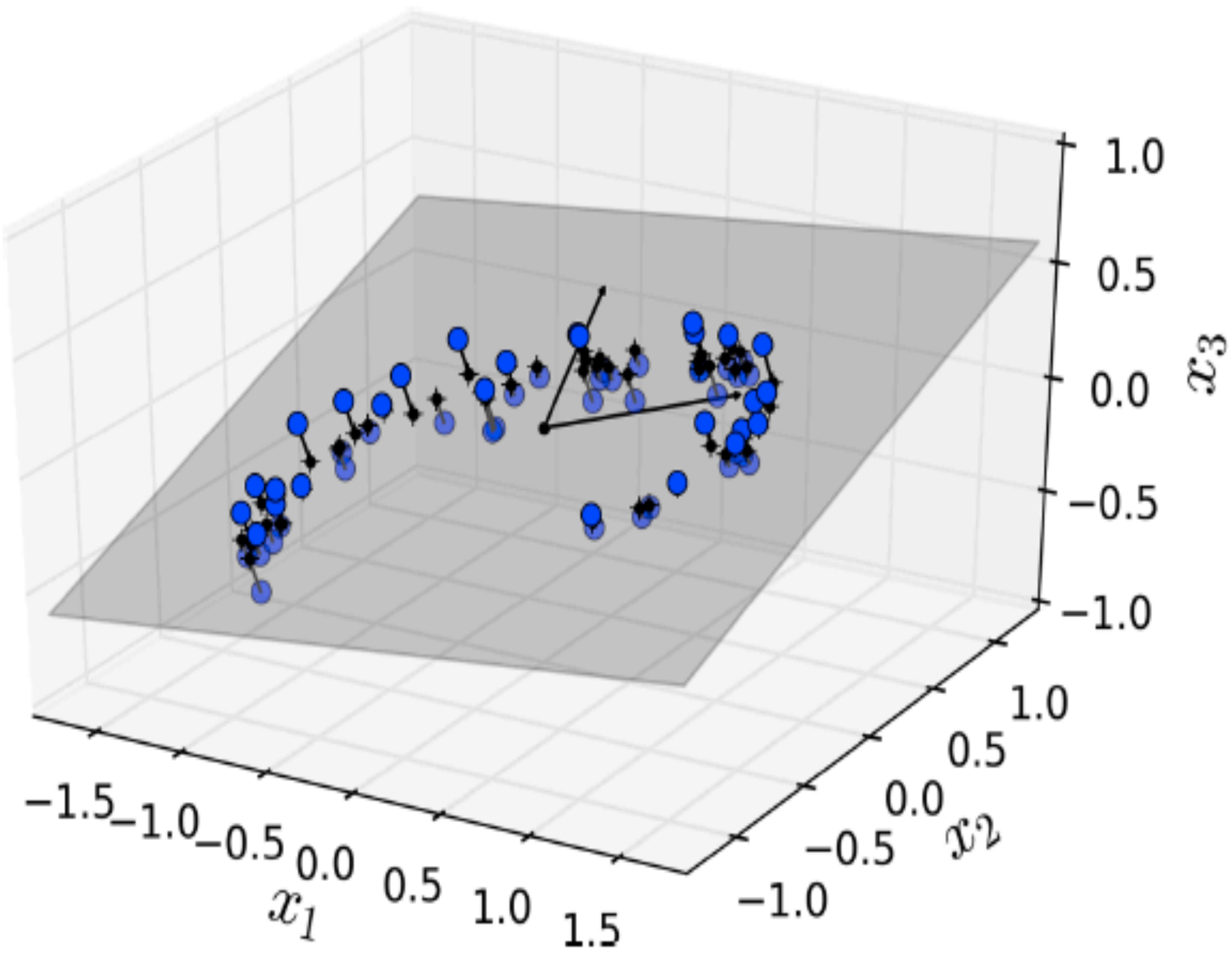


Figure showing 3D plot of test_predicted vs test_actual with their frequencies

Dimensionality Reduction

Principal Component Analysis (PCA)



A 3D dataset lying close to a 2D subspace

99% of variance conserved

ANN	761 Attributes	5 Attributes
Slope	0.99	0.99
Intercept	1.1	1.2
Subclass error	0.7	0.8

Future Possibilities

Machine learning algorithms can be trained better with more number of samples per class.

For example, LAMOST dataset (7.7 million good dataset).

References

- Hands-on machine learning with scikit learn and Tensor Flow by Aurélien Géron, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
- Scikit-learn: Machine Learning in Python, Pedregosa *et al*., JMLR 12, pp. 2825-2830, 2011.
- Jacoby, et al., ApJ, 56: 257-281, 1984.
- Gulati, et al., ApJ, 426: 340-344, 1994.

Future Possibilities

Machine learning algorithms can be trained better with more number of samples per class.

For example, LAMOST dataset (7.7 million good dataset).

References

- Hands-on machine learning with scikit learn and Tensor Flow by Aurélien Géron, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
- Scikit-learn: Machine Learning in Python, Pedregosa *et al*., JMLR 12, pp. 2825-2830, 2011.
- Jacoby, et al., ApJ, 56: 257-281, 1984.
- Gulati, et al., ApJ, 426: 340-344, 1994.

https://github.com/Shrishml/Classification-of-Spectral-Library


THANK
YOU!