



TEXAS TECH UNIVERSITY

Predictive Analytics Process Using Machine Learning for Student's Retention

2021 TAIR Annual Conference (Virtual)

Luciano Boas

Rohan Patil

Abhishek Kumar

February 26th, 2021



01 Overview

- Introduction
- Machine Learning Basics

02 Exploratory Analysis

- Dataset
- Exploratory Analysis

03 Machine Learning

- Logistic Regression
- SVM
- KNN
- Comparison of Algorithms
- Conclusion and Future Directions

A PRELIMINARY STUDY ON RETENTION



“...eight years later, one-third of those individuals are no longer enrolled and have not earned any formal credentials.”
—National Student Clearinghouse.

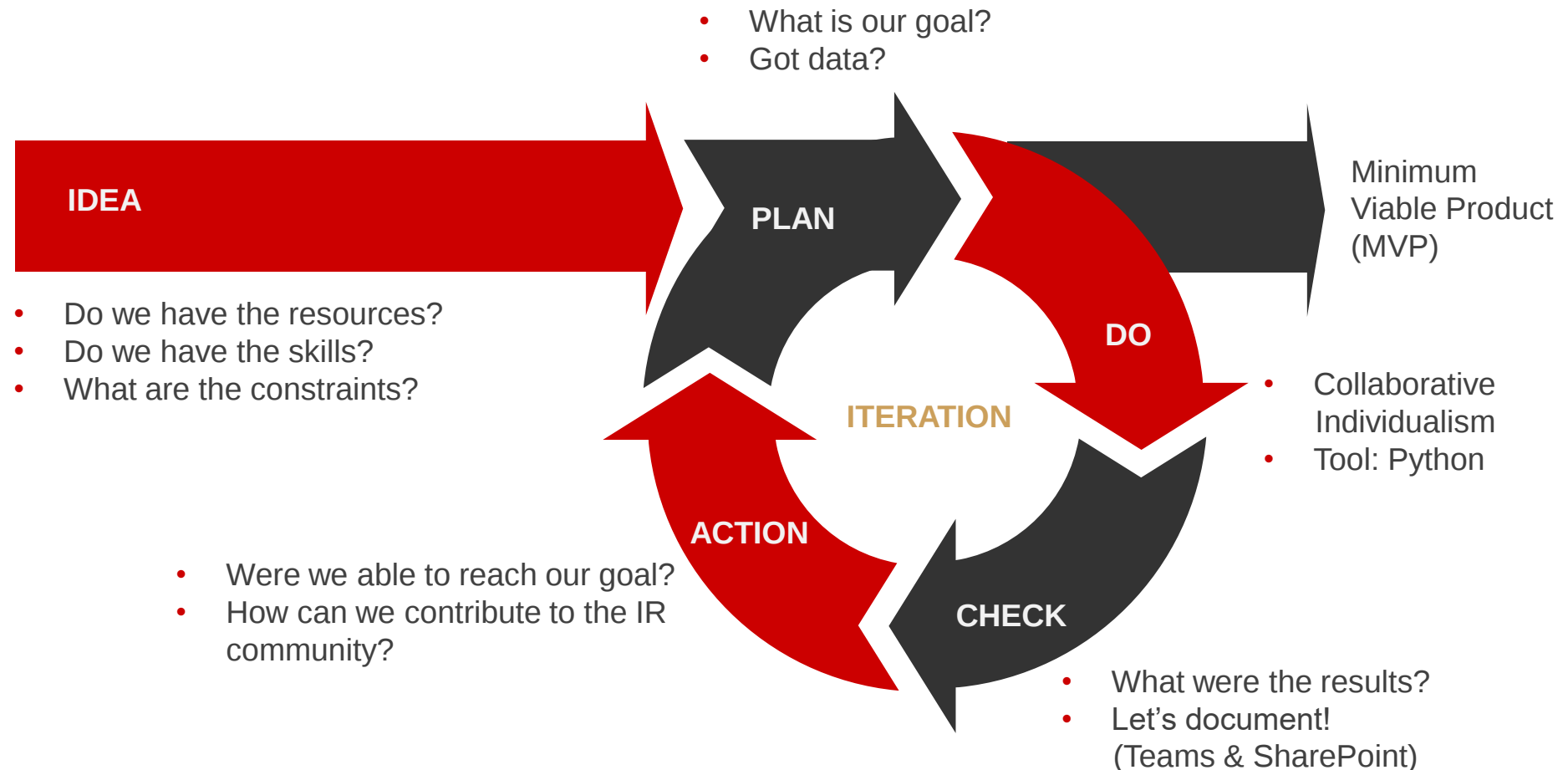




PROJECT MANAGEMENT OVERVIEW

“First comes thought; then organization of that thought, into ideas and plans; then transformation of those plans into reality.”

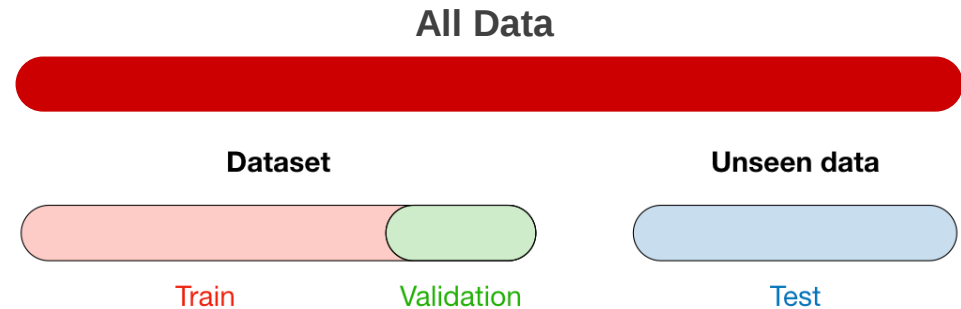
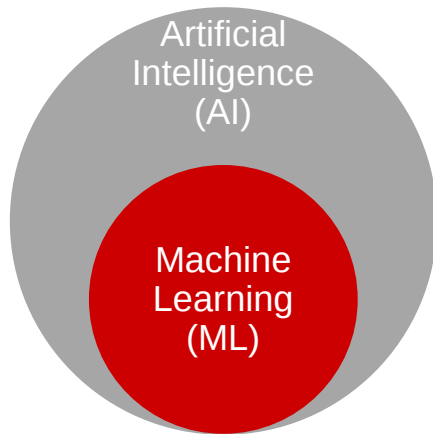
- Napoleon Hill



MACHINE LEARNING (ML) APPROACHES/BEST PRACTICES



“Machine Learning learns patterns from historic data (training) and tries to generalize that in unseen (new) data.”
Mathangi Sri

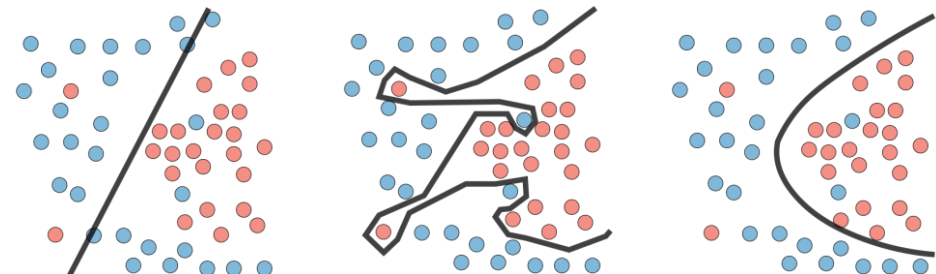


Source: <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks>

Best Practices (not limited to):

- Check for missing data
- Check for collinearity
- Avoid data leakage
- Manage imbalanced classifier
- Don't reinvent the wheel
- Focus on model explainability
- **Quick and dirty | Fail fast**
- **Minimum Viable Product (MVP)**

Common problems in the Training dataset (ML modeling):



Source: <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks>

Underfitting

Overfitting

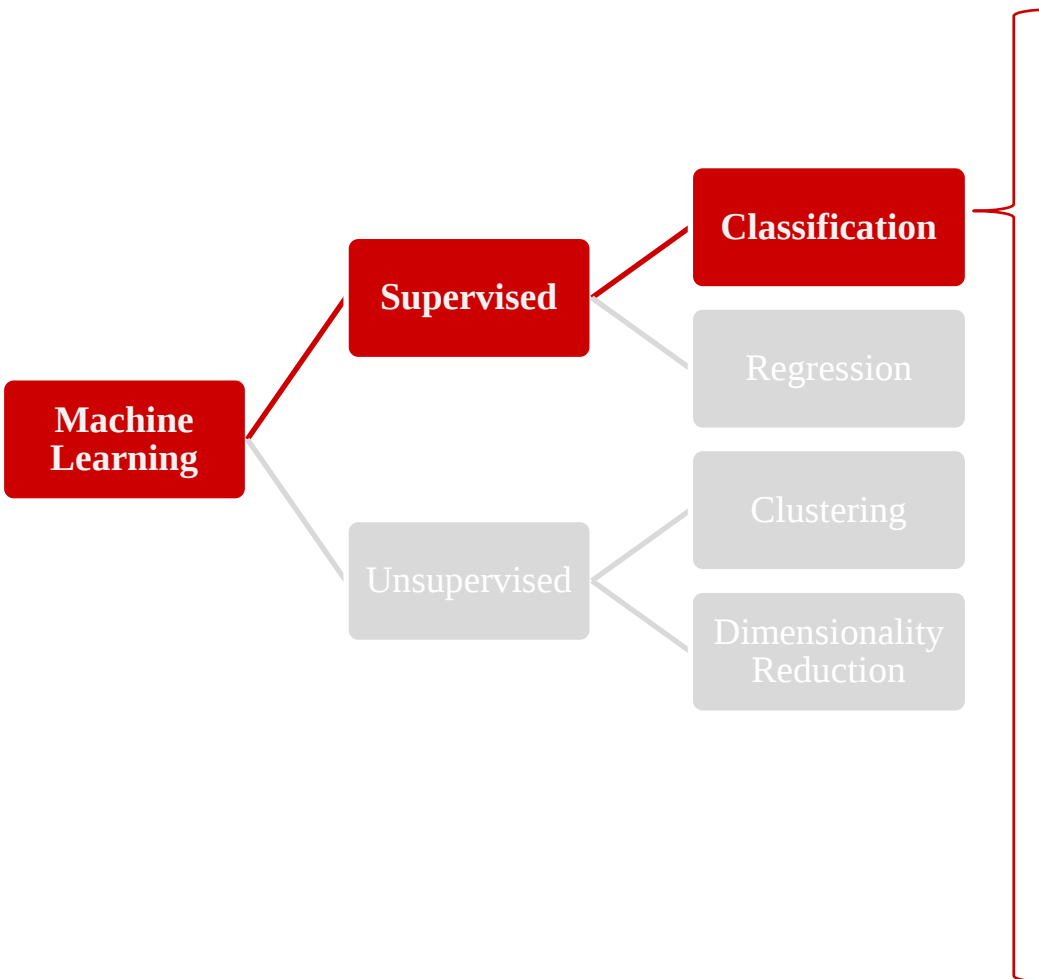
Good fit ✓

Fail to generalize

MACHINE LEARNING – TWO MOST COMMON TYPES OF LEARNING



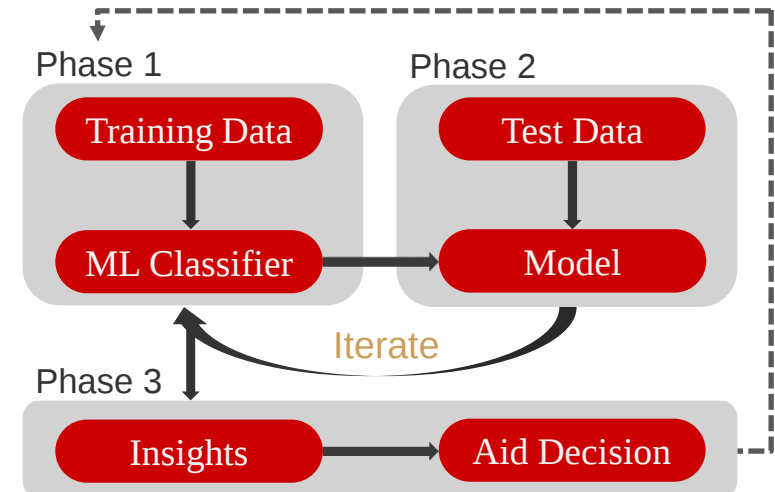
Machine Learning allows testing how different algorithms perform when doing the same task.



Multiple Algorithms Benchmarking (example):

- Logistic Regression
- Decision Trees
- Support Vector Machine

ML Classification Framework:





Machine Learning & Data Science Project Management

- [Structuring Machine Learning Projects](#) by DeepLearning.AI – Coursera.
- [The Machine Learning Canvas](#) by Louis Dorard.
- [The Practical Guide to Managing Data Science at Scale](#) by Domino.
- [Team Data Science Process](#) by Microsoft.
- [TAMIDS Data Science Webinars](#) by Jian Tao, Texas A&M Institute of Data Science (TAMIDS).



01 Overview

- Introduction
- Machine Learning Basics

02 Exploratory Analysis

- Dataset
- Exploratory Analysis

03 Machine Learning

- Logistic Regression
- SVM
- KNN
- Comparison of Algorithms
- Conclusion and Future Directions



Dataset Attributes – Fall 2012 to Fall 2014

- No. of students retained = 9442
- No. of students not retained = 1919
- Total number of students = 11361
- Dependent Variable : RET1FF
(Retention Fall to Fall)

We Coded categorical data
into numerical values

Columns	Column Desc
ACAD_STAND	Academic Standing
AGE	Age
CITIZEN	Citizen
CRED_ATMP	Credits Attempted
EOT_GPA	End of Term GPA
ETHNICITY	Ethnicity
GENDER	Gender
GPA_CREDITS	GPA Credits
HOME	Home
HONORS	Honors
HS_GPA	High School GPA
HS_PRCTL	High School Percentile
HS_RANK	High School Rank
HS_SIZE	High School Size
PART_FULL	Part Full
QL PTS	Quality Points
RET1_FS	Retention Fall to Spring
Sat_super_Score	SAT Super Score
STUD_CLASS	Student Classification

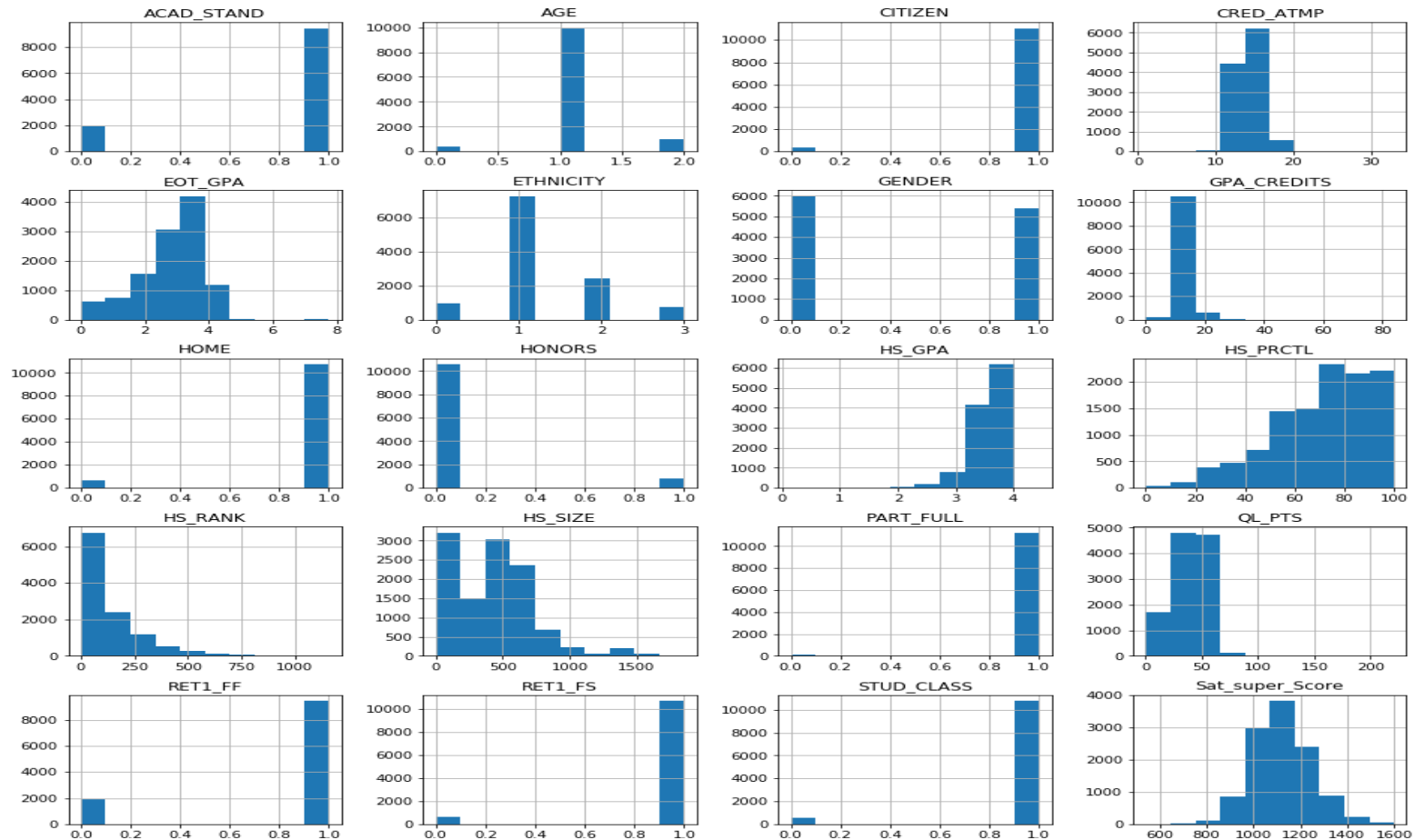
Columns	Coded Values
Part Full	F = 1 , P = 0
Student Classification	FR = 1 , Others = 0
Ethnicity	B = 0 , WH = 1 , HI = 2, A = 3, All Others = 4
Gender	M = 1 , F = 0
Home	Instate = 1, Others 0
Academic Standing	GS = 1 , SP = 2 , Else 0
Honors	Y = 1 Else 0
AGE	LTE 17 = 0, EQ 18 = 1, GTE 19 = 2

*Null Values were removed from
the dataset

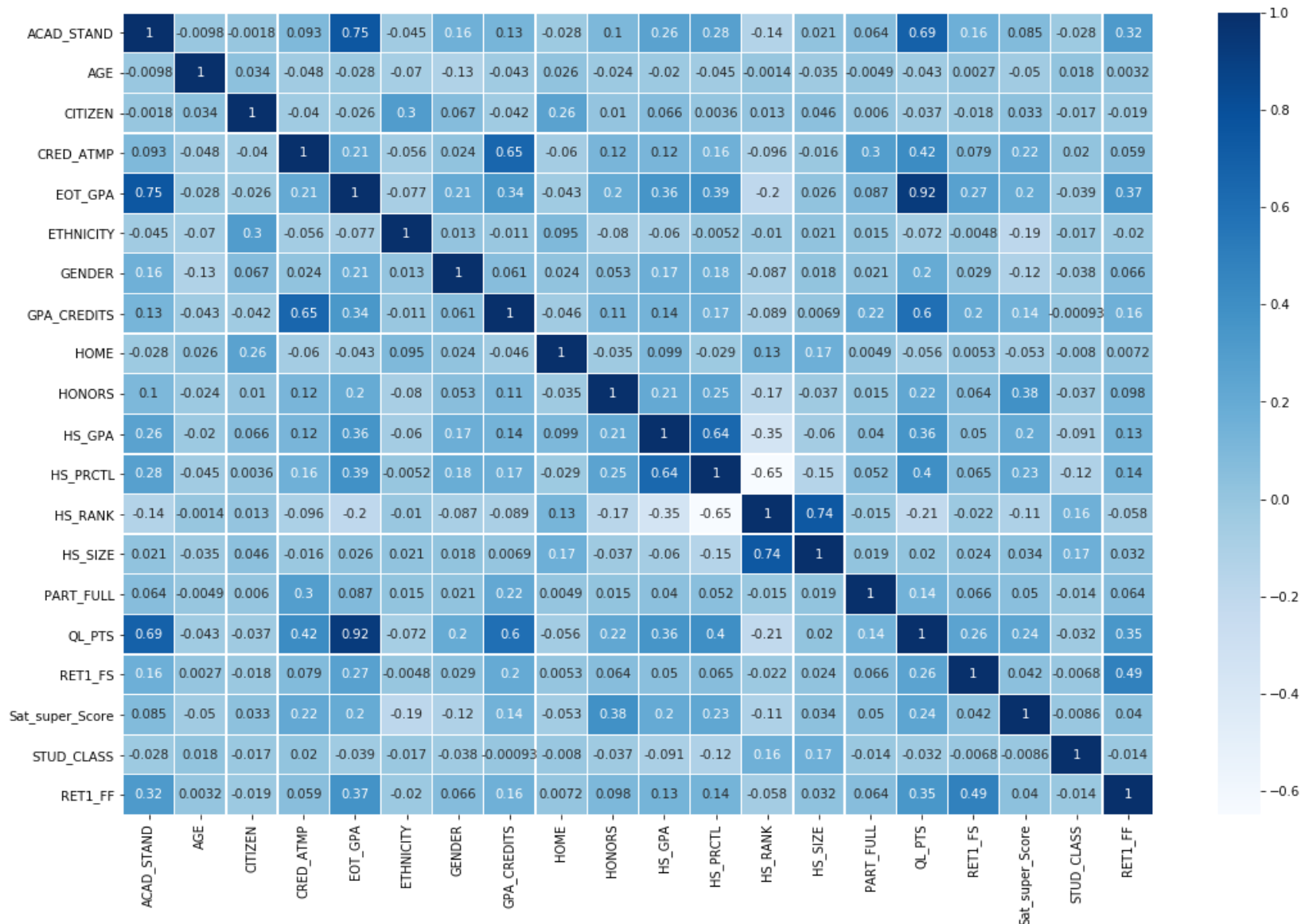
Dataset Distribution



“Histogram is a plot that lets you show the underlying frequency distribution or the **probability distribution** of a **continuous numerical variable**.”



Correlation Matrix





Feature Importance Test – Extra Trees Classifier

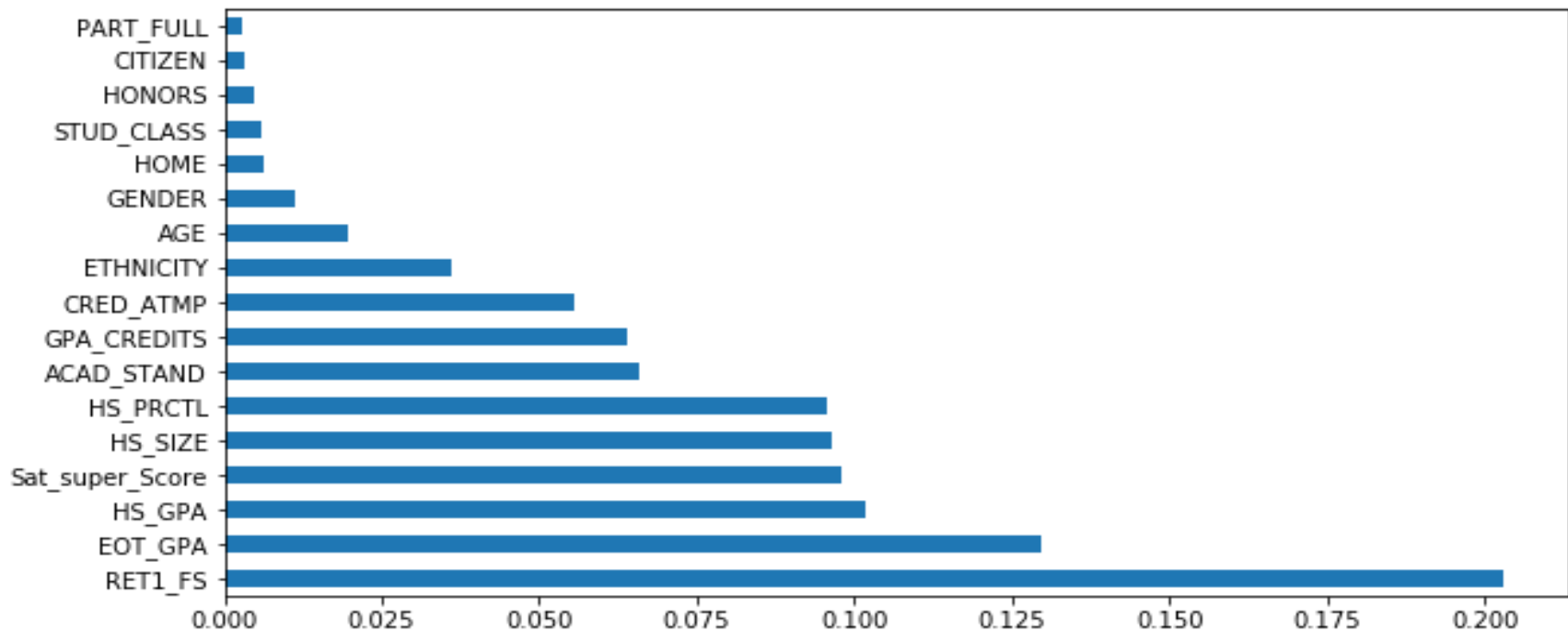
Top reasons to use feature selection are:

"It enables the machine learning algorithm to train faster."

"It reduces the complexity of a model and makes it easier to interpret."

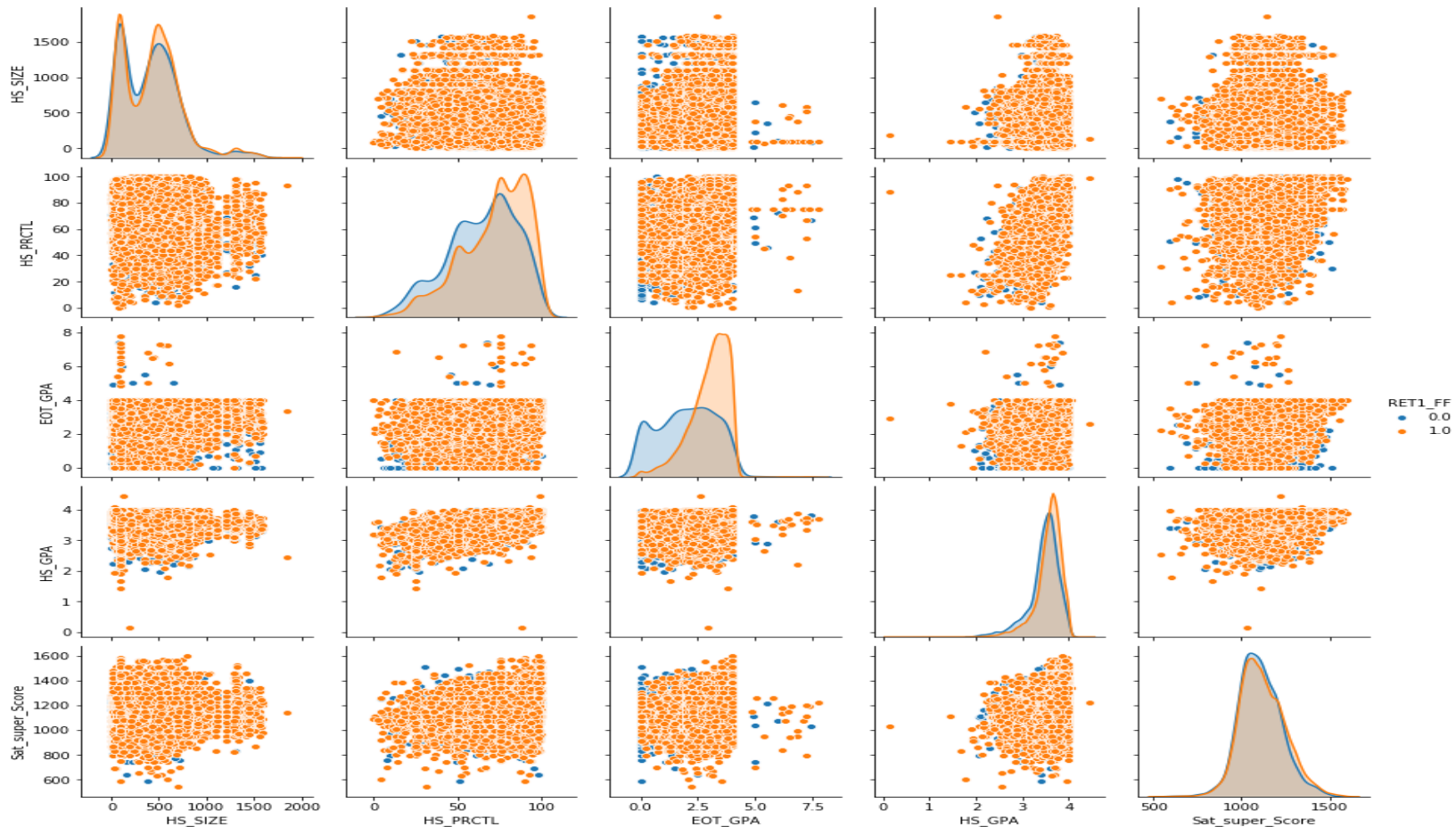
"It improves the accuracy of a model if the right subset is chosen."

"It reduces overfitting."



Important Features with Ret1FF

“Pairplot shows the relationship for (n,2) combination of variable in a DataFrame as a matrix of plots and the diagonal plots are the univariate plots.”



Confusion Matrix



True Positive:

“Interpretation: You predicted positive and it’s true.
You predicted that the student will be retained, and is retained”

True Negative:

“Interpretation: You predicted negative and it’s true.
You predicted that the student will not be retained and is not retained”

False Positive: (Type 1 Error)

“Interpretation: You predicted positive and it’s false.
You predicted that the student will be retained when is not retained

False Negative: (Type 2 Error)

“Interpretation: You predicted negative and it’s false.
You predicted that the student will not be retained when is retained.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN



Classification Report - Part 1

Dependent Variable	Precision	Recall	F1 - Score	Support
0 (Not Retained)	$TN / (TN + FN)$	$TN / (TN + FP)$	AVG(Precision, Recall)	No of Obs
1 (Retained)	$TP / (TP + FP)$	$TP / (TP + FN)$	AVG(Precision, Recall)	No of Obs

Precision:

“The precision returns the proportion of true positives among all the values predicted as positive.”

Recall:

“The recall returns the proportion of positive values correctly predicted.”

Source: <https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>

Source: <https://medium.com/swlh/confusion-matrix-and-classification-report-88105288d48f>

**F1-score:**

“The f1-score is the harmonic mean of precision and recall. It is often used to compare classifiers.”

Support: “ number of observations for each class.”



01 Overview

- Introduction
- Machine Learning Basics

02 Exploratory Analysis

- Dataset
- Exploratory Analysis

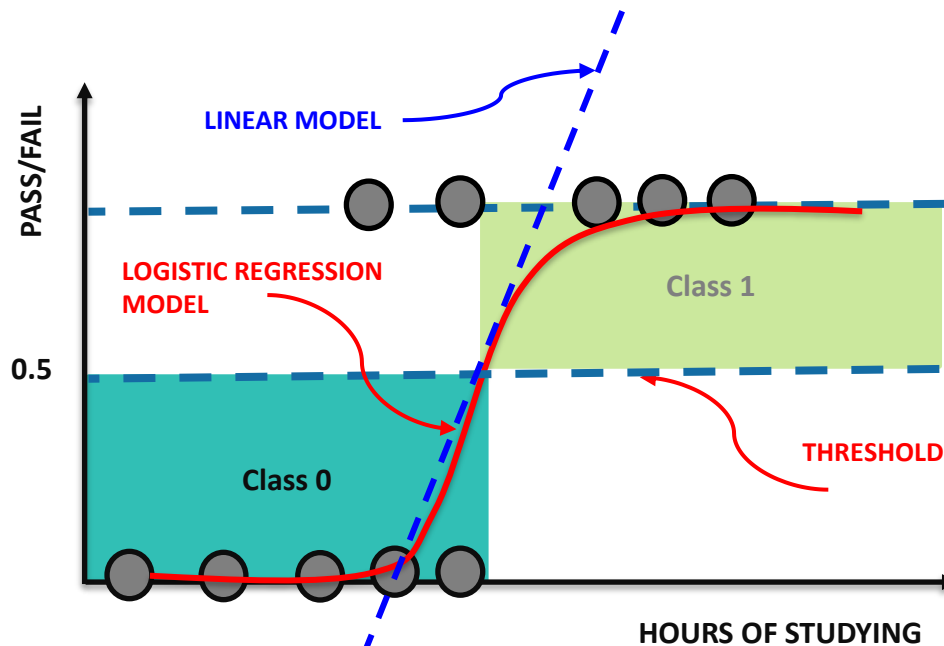
03 Machine Learning

- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- Comparison of Algorithms
- Conclusion and Future Directions

LOGISTIC REGRESSION

- **Linear regression:** used to predict outputs on a continuous spectrum
- **Logistic regression:** used to predict binary outputs with two possible values labeled “0” or “1”
- Logistic model can be one of the two classes: pass/fail, win/loss, retained/not retained

Hours Studying	Pass/Fail
1	0
1.5	0
2	0
3	1
3.25	0
4	1
5	1
6	1

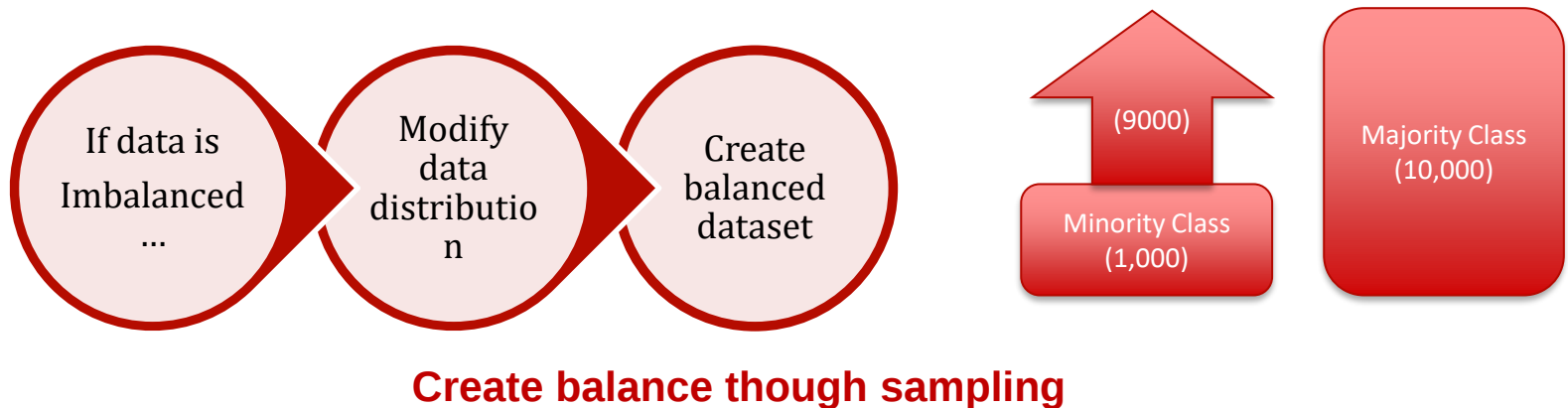


R. Ahmed, M. Bouchard ML Classification Bootcamp in Python

- Linear equation:
 - $y = b_0 + b_1 * x$
- Apply Sigmoid function:
 - $P(x) = \text{sigmoid}(y)$
 - $P(x) = \frac{1}{1+e^{-y}}$
 - $P(x) = \frac{1}{1+e^{-(b_0+b_1*x)}}$

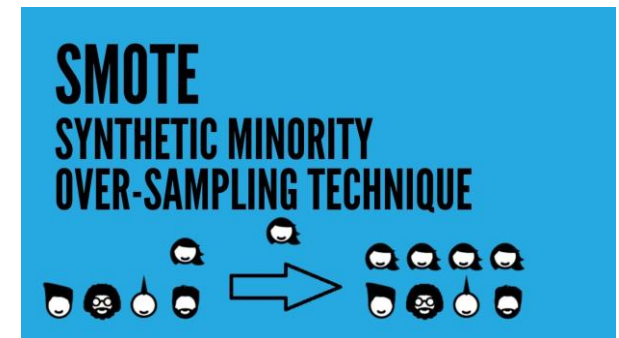
THE CLASS IMBALANCE PROBLEM

- A common problem in Machine Learning
- Almost all the instances belong to one major class and the rest to minor class
-



Re-sampling the data:

- Over-sampling increases the number of minority class members
- Under-sampling aims to reduce the number of majority samples
- We used SMOTE for our preliminary research





LOGISTIC REGRESSION – MODEL TRAINING & TESTING

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.37	-
Retained	-	-	0.92	-

Without sampling and without feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.50	-
Retained	-	-	0.93	-

Without sampling and with feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.52	-
Retained	-	-	0.93	-

With oversampling and without feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.41	-
Retained	-	-	0.77	-

With undersampling and without feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.53	-
Retained	-	-	0.88	-

With oversampling and with feature selection

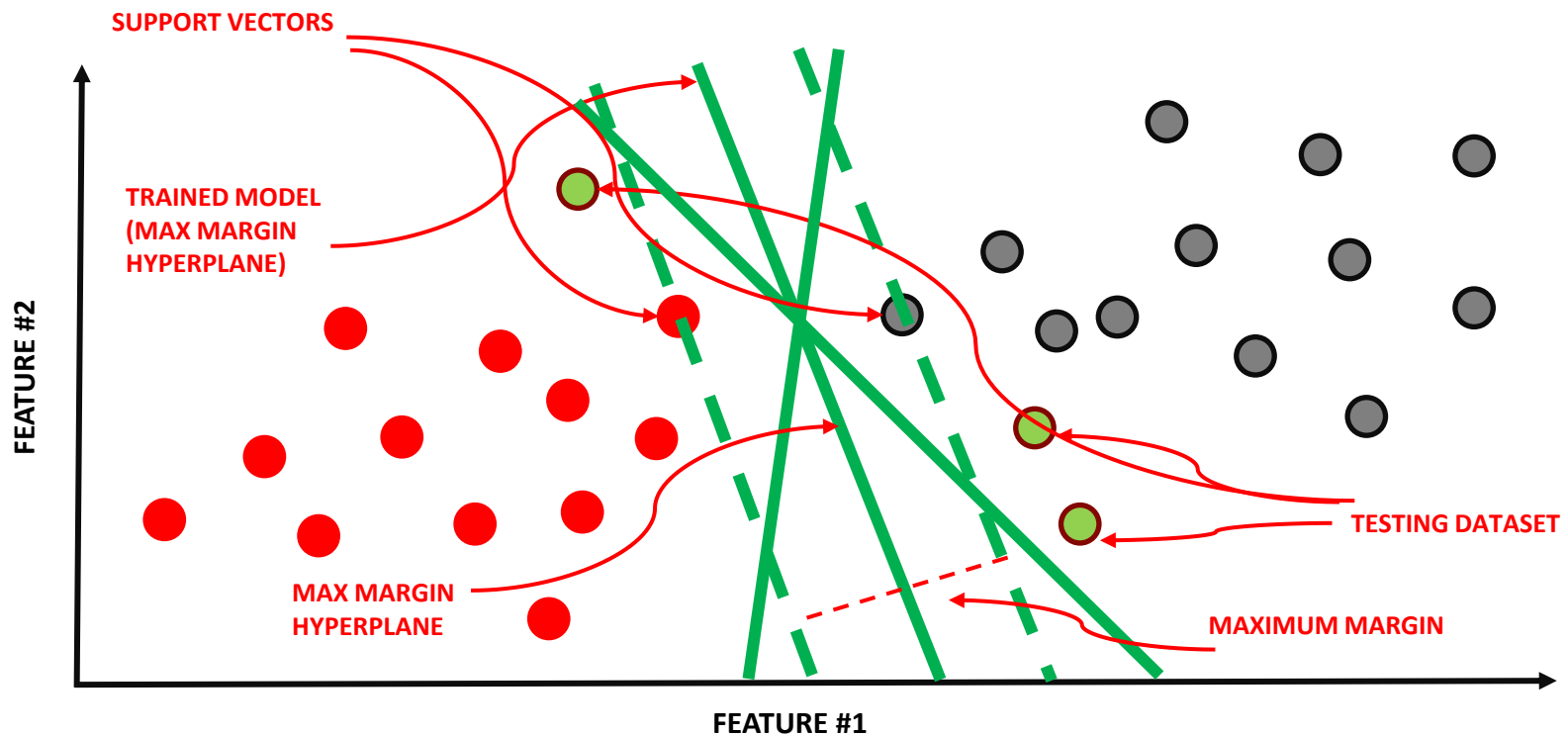
	Precision	Recall	F1-score	Support
Not Retained	-	-	0.44	-
Retained	-	-	0.77	-

With undersampling and with feature selection

SUPPORT VECTOR MACHINES (SVM): MODEL EVALUATION

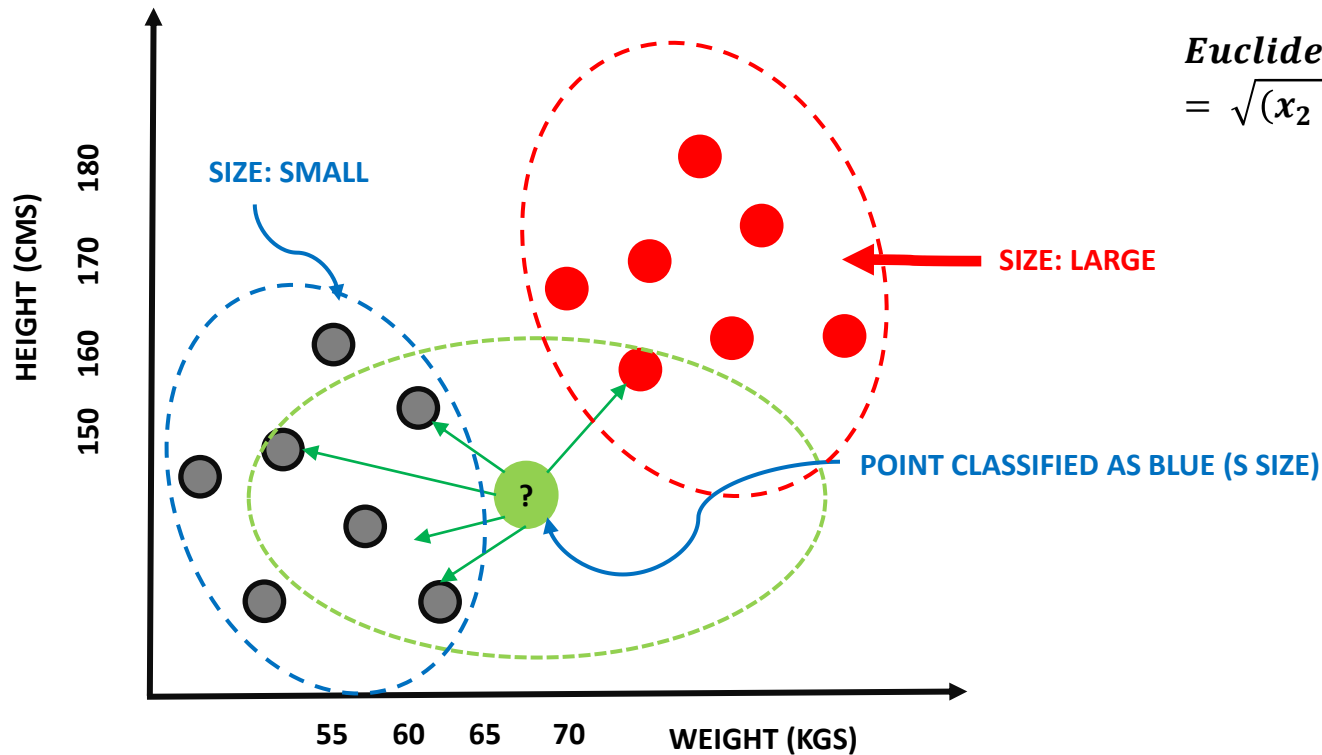


- SVMs are based on the idea of finding a hyperplane that divides a dataset into two classes
- The line that maximizes the minimum margin is a good bet.
- This maximum-margin separator is determined by a subset of the datapoints.



K NEAREST NEIGHBORS (KNN): MODEL EVALUATION

- k-nearest neighbors algorithm (KNN) is a classification algorithm
- KNN works by finding the **most similar** data points in the training data, and attempt to make an **educated guess** based on their classifications



$$\text{Euclidean Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

SVM and KNN – MODEL TRAINING & TESTING

- **train_test_split** is a function in Sklearn model selection for splitting data arrays into two subsets
- **train_test_split:** 70:30

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.39	-
Retained	-	-	0.92	-

SVM - Without sampling and without feature selection

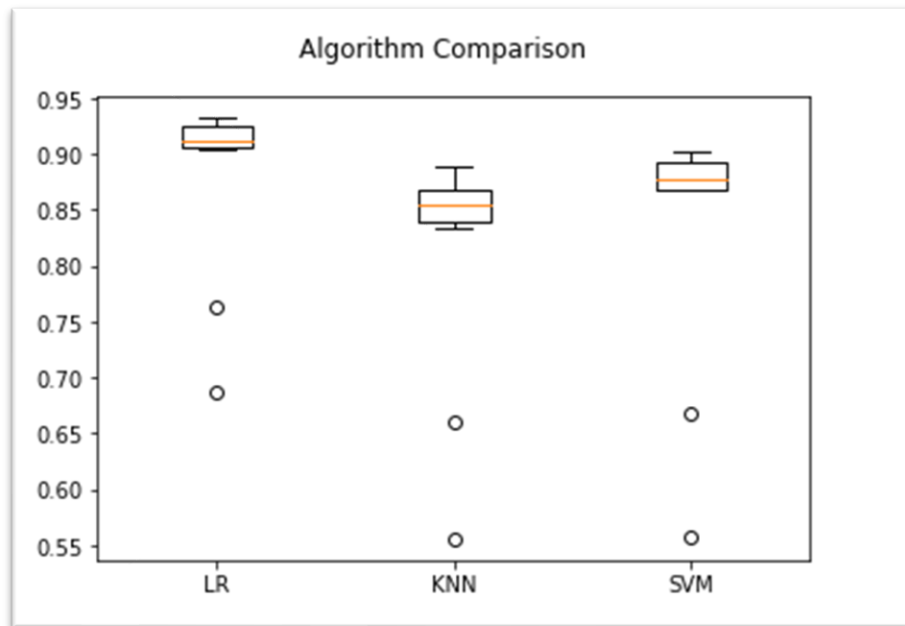
	Precision	Recall	F1-score	Support
Not Retained	-	-	0.44	-
Retained	-	-	0.89	-

KNN - Without sampling and without feature selection

COMPARISON OF ALGORITHMS - BENCHMARKING

Each model will have different performance characteristics

- **Prepare Dataset.** Load the libraries and dataset ready to train the models
- **Train Models.** Train standard machine learning models on the dataset ready for evaluation
- **Compare Models.** Compare the trained models using k-fold cross validation procedure



ML Models	Results
Logistic Reasoning	87.93% (0.079328)
Support Vector Machine	82.98% (0.112209)
K Nearest Neighbor	81.05% (0.104777)



PREDICTION USING THE MODEL & FUTURE WORK

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.45	-
Retained	-	-	0.95	-

Classification Report for Fall 2015

Conclusion and Future Directions

- Logistic regression with oversampling of our data and feature selection produced best results
- Update our preliminary research with more salient parameters
- Use of other techniques and methods to handle imbalance datasets
- Use different machine learning models to predict various possibilities with HigherEd data



TEXAS TECH UNIVERSITY®

From here, it's possible.