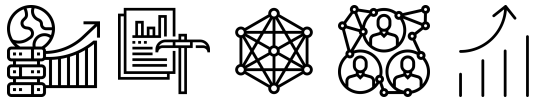


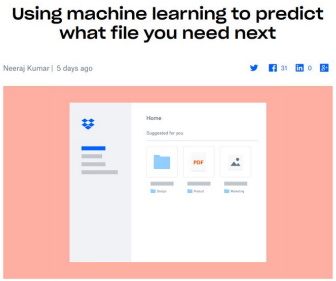
CSC 5741
Lecture 6: Introduction to Machine Learning

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia

Announcements—May 7, 2019

- Assessments
  - Class Theory Test: May 21, 2019
  - Mini Project Deliverables: May 20, 2019
    - Technical Report
    - Code Repository for Fully Functional Implementation (including interactive Jupyter Notebook)
    - Presentation Slides
  - Mini Project Presentations: May 28, 2019
    - Demonstrations [2 minutes]
    - Presentations [10 minutes]
    - Q&A [3 minutes]

Todos Showcases—May 7, 2019



Lecture Series Outline

- Part I: Introduction to Machine Learning
- Part II: Datasets

Lecture Series Outline

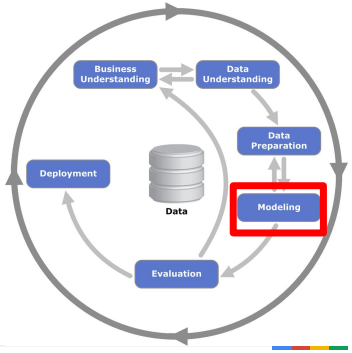
- Part I: Machine Learning
  - Introduction
  - Machine Learning
  - Data Attributes
  - Types of Learning
  - Evaluation
- Part II: Datasets

Introduction (1/2)

- The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining
  - CRISP-DM segments a data mining project into six phases with no strict order of execution
  - Surveys conducted suggest CRISP-DM is the most widely used methodology

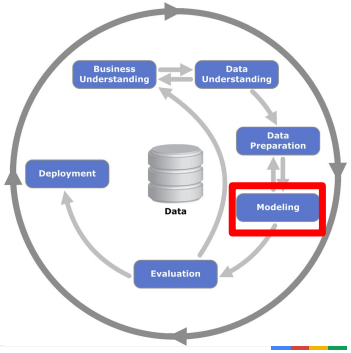
Introduction (2/2)

- Define the model components, features, how it behaves and how to interpret it
- Evaluate the various alternative techniques that can be integrated with the model
  - e.g. Evaluate different classification algorithms



Machine Learning (1/2)

- Finding patterns in data that provide insight or enable fast and accurate decision making
  - Prediction
  - Pattern recognition



Machine Learning (2/2)

- **Aspects of Machine Learning**
  - Input: x
  - Output: f(x)
  - f(x) -> y
    - f(x) -> prediction
    - f(x) -> pattern
  - Input is generally attributes of a given dataset
  - A key prerequisite to solving a machine learning problem is to have **reliable data** available

Data Representation

- **Input and output data involving machine learning algorithms need to be represented in a mathematical way using data attributes.**
  - The representation is in the form of attribute-value pairs
  - E.g. {gender: “female”, minor: “Mathematics”}
  - Attributes type: continuous, categorical and ordinal values
- **The representation is dependant on the problem domain and the goals**
  - Predicting is student will pass or fail or JCTR Basic Needs Basket
- **Remember that a predictor is a mathematical function**
  - Y = f(x)

Data Attributes (1/5)

- **Continuous/numeric attributes**
  - Integers or floats
  - Mathematical operations can be performed on values
  - Usually the case that values need normalisation to use a similar scale before applying algorithm

	A	B	C	D	E
	StudentID	CA	CA Grade	Gender	Minor
	3283f0f7cce04d7ce5cd5dfdd6d191bf	4.78	D	M	Civic
	3aa39f19efac1157cc4a2529c3f391fe	23.7	C	M	Languages
	8e5435d4f634804d12512b6caeeb4316	25.13	C+	M	Civic
	64e4a90e4dfa32a2c877c93d48354c81	17.48	D	M	History
	89407737ae371e6080f88ad29ddd07bd	19.6	D	M	Civic
	b3f4e74e8b181868a648635c0bbf376e	32.6	B	M	Civic
	4db47b46b9680eada1c7d0862e9904f1	24.43	C	M	History
	a0d3784a0e902162d676fb0ae1fb6c36	20.8	D+	M	Mathematics
	c463632dc537747e3925bc8e4038ad42	26.65	C+	M	Mathematics
	f8b7ffcf26cde897a3ad67a83c99286d	30.55	B	M	Mathematics
	a96fbf28013f6b01358acfe2e85f598d	25.2	C+	F	Languages
	cad8f195906f7a866094b5fa3d4c0633	19.6	D	F	History
	94b1c4e2f44e735b872faa1078577ac4	25.45	C+	M	Mathematics
	e8c074808b39cb38e345f4a774696590	25.28	C+	F	Civic
	5e1c7620bf6173292bb3cd1295f38f62	23.55	C	M	Mathematics
	58c274a03ebb5d172897c13e209d162b	24.13	C	M	Civic
	d73db0840f8452e20001863997cfbe5a	25.7	C+	F	Civic
	f30a59bdf6516777b67cfb5d68c0bee1	19.78	D	M	Geography
	b872461c4668ab02126d9f05c9ad95d0	28.13	C+	M	Mathematics

Data Attributes (2/5)

- **Categorical attributes**
  - Discrete set of values
  - Only one value can be held at a time
  - Categories are mutually exclusive
  - The only numeric operation performed is equality testing

	A	B	C	D	E
	StudentID	CA	CA Grade	Gender	Minor
	3283f0f7cce04d7ce5cd5dfdd6d191bf	4.78	D	M	Civic
	3aa39f19efac1157cc4a2529c3f391fe	23.7	C	M	Languages
	8e5435d4f634804d12512b6caeeb4316	25.13	C+	M	Civic
	64e4a90e4dfa32a2c877c93d48354c81	17.48	D	M	History
	89407737ae371e6080f88ad29ddd07bd	19.6	D	M	Civic
	b3f4e74e8b181868a648635c0bbf376e	32.6	B	M	Civic
	4db47b46b9680eada1c7d0862e9904f1	24.43	C	M	History
	a0d3784a0e902162d676fb0ae1fb6c36	20.8	D+	M	Mathematics
	c463632dc537747e3925bc8e4038ad42	26.65	C+	M	Mathematics
	f8b7ffcf26cde897a3ad67a83c99286d	30.55	B	M	Mathematics
	a96fbf28013f6b01358acfe2e85f598d	25.2	C+	F	Languages
	cad8f195906f7a866094b5fa3d4c0633	19.6	D	F	History
	94b1c4e2f44e735b872faa1078577ac4	25.45	C+	M	Mathematics
	e8c074808b39cb38e345f4a774696590	25.28	C+	F	Civic
	5e1c7620bf6173292bb3cd1295f38f62	23.55	C	M	Mathematics
	58c274a03ebb5d172897c13e209d162b	24.13	C	M	Civic
	d73db0840f8452e20001863997cfbe5a	25.7	C+	F	Civic
	f30a59bdf6516777b67cfb5d68c0bee1	19.78	D	M	Geography
	b872461c4668ab02126d9f05c9ad95d0	28.13	C+	M	Mathematics

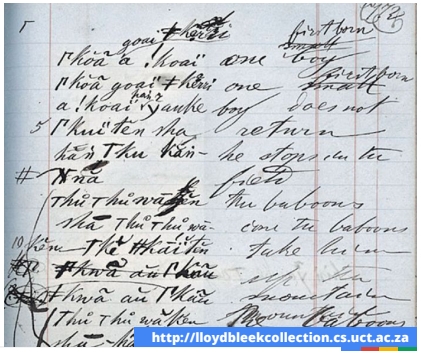
Data Attributes (3/5)

- **Ordinal attributes**
  - Similar to categorical attributes except that they exhibit a natural order
  - Likert scales
  - Can be encoded as numbers

Student ID	Experience Using Computers	Do you
c6476f43b01dd8a53e2d5ea2ae2413db	More than 5 years	Yes
eedf8db73ac1dac81aede402ae3aab81	1 to 2 years	Yes
98af9129709200207 48a468fe3b449fd	No Experience	Yes
4feeb9cab471243d43ec3296d660f538	Less than 1 year	Yes
e30aa4ad6229d9f094aa5798569f703c	Less than 1 year	No
a456f8d006faa22926986930b7522349	No Experience	Yes
a456f8d006faa22926986930b7522349	No Experience	Yes
f3723b194ea12bbc7b4040aede416be1	Less than 1 year	Yes
47f1c00713cf02769a52b433d21e413d	More than 5 years	Yes
d982aa9c301dbfb5f1b2d1d6d703ba39	No Experience	Yes
572adae2e52cf6a23f4ae9c939666abf	More than 5 years	Yes
572adae2e52cf6a23f4ae9c939666abf	More than 5 years	Yes
bed3f24be6becc81b4a6217bc054f870	No Experience	Yes
058bcf7cbe93f598c0f6031ff8f1afc5	1 to 2 years	No
8dc89b81a89aaa5372be4f4b35f76d9c	More than 5 years	No
ee3f8e4efc99e7ab9a8a384fa8eb1f0d	No Experience	Yes
8a0cc50f06c959e79f10e0d78938f34c	Less than 1 year	Yes
fdf1e8906330fc223926f199b24dfccc	No Experience	No
3975b7445569d0094df2313b91edc5ee	1 to 2 years	No
3f02716971e5569157e449171db99bd0	1 to 2 years	Yes
bb0f0941bff4b55a1a7a1055fbfb74484	More than 5 years	Yes

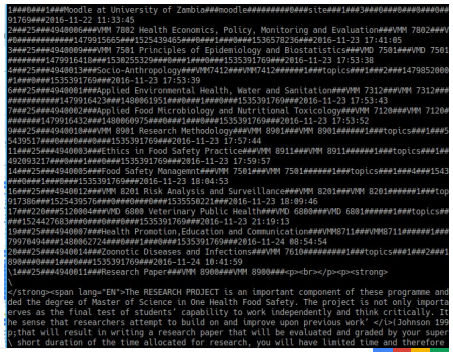
Data Attributes (4/5)

- **Data such as images can easily be represented using individual pixels**
  - Each pixel represents a distinct attribute



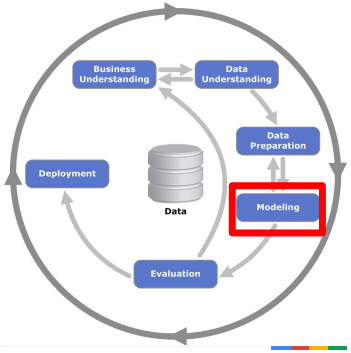
Data Attributes (5/5)

- **Textual content could easily be represented using individual terms forming the text**
  - Bag-of-words model
  - A numeric variable can be used to signal the relative importance of the word in the document



ML Techniques: Supervised Learning (1/8)

- **Supervised learning**
  - Involves the used of labeled data
  - The labels are the resulting output of the algorithm
  - Training is required to teach the algorithm
  - The goal is to predict a specific quantity
  - Accuracy can be measured directly



ML Techniques: Supervised Learning (2/8)

- **Predicting an output f(x) given an input x**
  - If f(x) is categorical, this would be a Classification problem
  - If f(x) is numerical and/or continuous, this would be a Linear Regression problem

ML Techniques: Supervised Learning (3/8)

Data-Driven Intervention-Level Prediction Modeling for Academic Performance

Mvurya Mgala
Dept of Computer Science
University of Cape Town
HPI School in ICT4D, 7701 Cape Town
mmgala@cs.uct.ac.za

Audrey Mbogho
Dept of Computer Science
University of Cape Town
HPI School in ICT4D, 7701 Cape Town
audrey.mb

Paper Reading Assignment #01

- **f(x)—High Intervention/Low Intervention**
- **x—Test scores, gender, age, student motivation, study time**
- **Labels—Manually done perhaps?**
- **Type of learning: Classification**

ML Techniques: Supervised Learning (4/8)

Document Type Classification in Online Digital Libraries

Cornelia Caragea,[1] **Jian Wu,**[2] **Sujatha Das Gollapalli,**[3] and **C. Lee Giles**[2]
[1]Department of Computer Science and Engineering, University of North Texas, Denton, TX
[2]College of Information Sciences and Technology, Pennsylvania State University, University Park, PA
[3]Institute for Infocomm Research, A*STAR, Singapore
ccaragea@unt.edu, jxw394@ist.psu.edu, gollapallis@i2r.a-star.edu

Paper Reading Assignment #02

- **f(x)—Book/Slides/Thesis/Paper/Resume**
- **x—File-specific, text-specific, section-specific, containment-specific**
- **Labels—Manually done**
- **Type of learning: Classification**

ML Techniques: Supervised Learning (5/8)

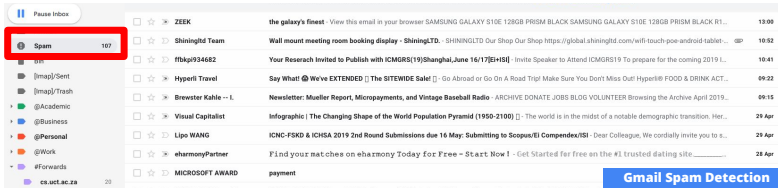
Moodle Predicta: A Data Mining Tool for Student Follow Up

Igor Moreira Félix[1], Ana Paula Ambrósio[1], Priscila Silva Neves[1],
Joyce Siqueira[1] and Jacques Duilio Brancher[2]
[1]*Instituto de Informática, Universidade Federal de Goiás, Goiânia, Brazil*
[2]*Departamento de Computação, Universidade Estadual de Lond*

Paper Reading Assignment #03

- **f(x)—Students "at risk"**
- **x—Behaviour and interaction within The Moodle (course, posts, messages, time spent on activities)**
- **Labels—Manually done perhaps?**
- **Type of learning: Classification**

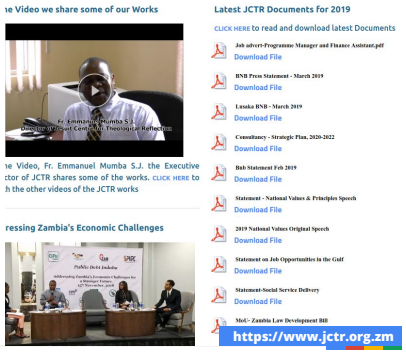
ML Techniques: Supervised Learning (6/8)



- f(x)—Spam/Not Spam
- x—Email address, subject, email content
- Labels—Mailbox user tagging; Gmail automatic detection
- Type of learning: Classification

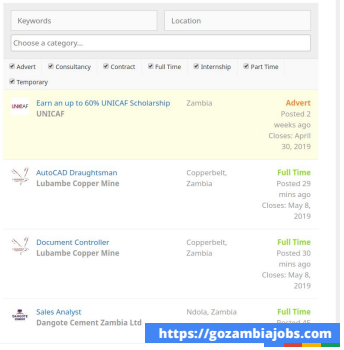
ML Techniques: Supervised Learning (7/8)

- f(x)—BnB Amount for specific month next year
- x—USD rate; Rand rate; political climate; ????
- Labels—Already available
- Type of learning: Regression



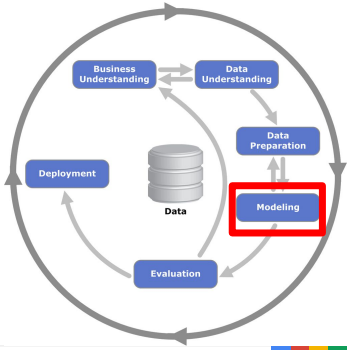
ML Techniques: Supervised Learning (8/8)

- f(x)—Expected Salary
- x—qualifications, domain, experience, references
- Labels—?????
- Regression



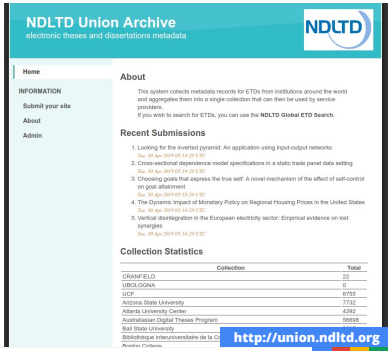
ML Techniques: Unsupervised Learning (1/2)

- Unsupervised learning
  - No specific value to be predicted
  - The goal is pattern recognition, in order to learn more about the data
  - No labelled is required
  - No training required
  - Evaluation is typically qualitative since there are no predefined labels
    - Typically subjective



ML Techniques: Unsupervised Learning (2/2)

- f(x)—WE DO NOT KNOW
- x—Institution, ETD title, ETD abstract
- Labels—No labels
- Clustering



ML Techniques: Reinforcement Learning

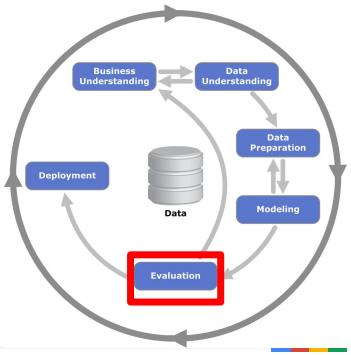
- Semi-supervised/reinforcement learning
  - Uses a combination of supervised and unsupervised learning
  - Unsupervised learning techniques are typically used to improve supervised learning algorithms
  - Example: Scenario where there is fewer labeled data

Machine Learning Techniques

- We will focus on algorithms for solving the following problems
  - Regression problems
  - Classification problems
  - Clustering problems

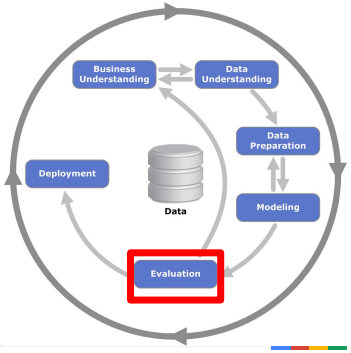
Evaluation (1/9)

- Evaluation is a systematic determination of a subject's merit, worth and significance, using a specified criteria.
  - It assess an entity to help in decision making; or to ascertain the degree of achievement or value in regard to the aim and objectives



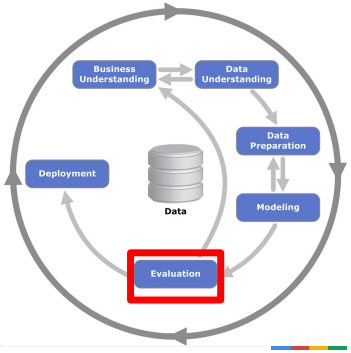
Evaluation (2/9)

- **Evaluation is systematic and rigorous.**
- **Evaluation involves critical assessment of a given set of objectives.**
  - How effective is Lighton Phiri at teaching CSC 5741?



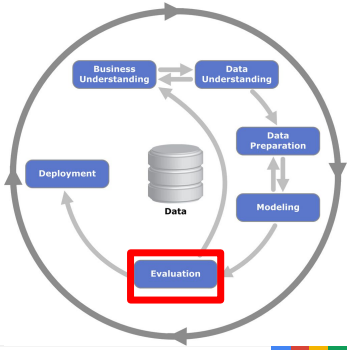
Evaluation (3/9)

- **Evaluation forms a crucial part of Machine Learning as it assess the relative effectiveness of learning**
  - How **efficient** is the learning process?
    - Computing resources are finite [...]
  - How **effective** is the learning outcome?
    - Accuracy is key [...]

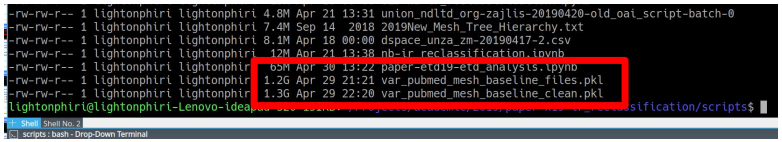


Evaluation (4/9)

- **Effectiveness—The extent towards which an ML model is successful at accomplishing its intended objectives.**
  - E.g. How accurate is it at classifying at risk students?
- **Efficiency—The relative cost implications of an ML model achieving its objectives**
  - How long does it take to present the output



Evaluation (5/9)



- **Efficiency is concerned with execution speed and relative usage of computing resources**
  - How much RAM is required to build the models?
  - How much processing power is required to build the models?
  - How long does it take to build the model?
  - How much storage space is required?
  - Cost implications?

Evaluation (5/9)

High Performance Computing @ ZAMREN

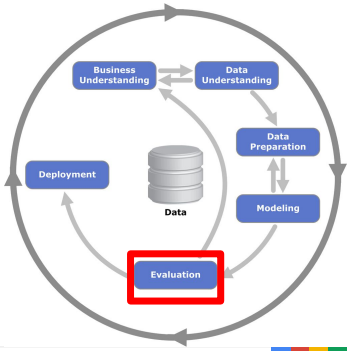
High Performance Computing(HPC) is an aggregation or clustering of computing power in a way that yields greater performance than could be obtained from a typical workstation or server.

This is done to solve large problems in science, engineering, or business. HPC enables computation and analysis of vast data, for example, in areas such as Water, Energy and Environment, Materials Engineering, Nuclear Physics, Genetics, Neurology, Astrophysics, Bio-informatics, Geosciences, Visualization and Imaging, among the numerous types science and mathematical analyses. They are also intensely used in product development, redesigns and process optimization.

This service at ZAMREN provides an opportunity for our researchers and students to undertake science driven research and be part of the global research communities in their respective research categories and indeed create innovative solutions for social and economic development. **http://hpc.zamren.zm**

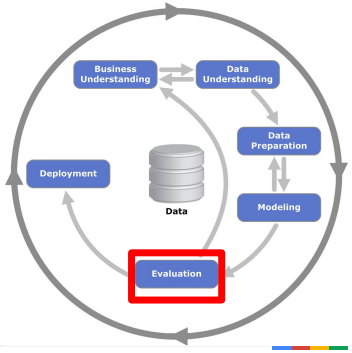
Evaluation (6/9)

- **Effectiveness of ML models involves measurement of accuracy in order to determine the relevance of the results**
  - Recall—The total number of relevant results returned.
    - Recall = relevant retrieved / total relevant



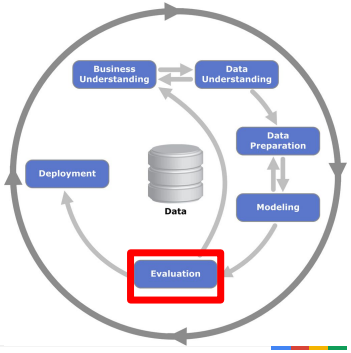
Evaluation (6/9)

- **Recall**
  - E.g. If the UNZA institutional repository has a total of 100 documents relevant to a query on “Information Retrieval” and only 70 are retrieved when a query related to IR is issued, then the recall is 70%



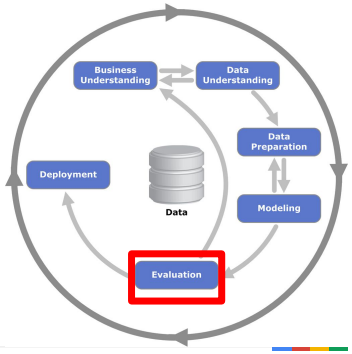
Evaluation (7/9)

- **Effectiveness of ML models involves measurement of accuracy in order to determine the relevance of the results**
  - Precision—The number of returned results that are relevant.
    - Precision = relevant retrieved / total retrieved

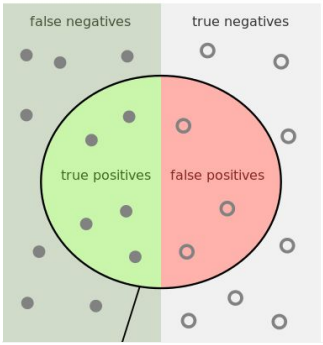


Evaluation (7/9)

- **Precision**
  - E.g. If a search of "Information Retrieval" in the UNZA institutional repository retrieves 100 documents and 40 of those documents are relevant, the precision is 40%



Evaluation (7/9)



How many selected items are relevant?

Precision =

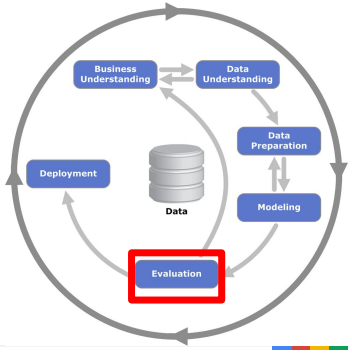
How many relevant items are selected?

Recall =

https://www.wikipedia.org

Evaluation (8/9)

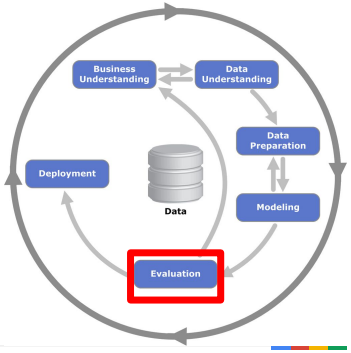
- **Effectiveness of ML models involves measurement of accuracy in order to determine the relevance of the results**
  - F1-score provides a comprehensive measure of a test's accuracy.
    - It considers both the precision p and the recall r



Evaluation (9/9)

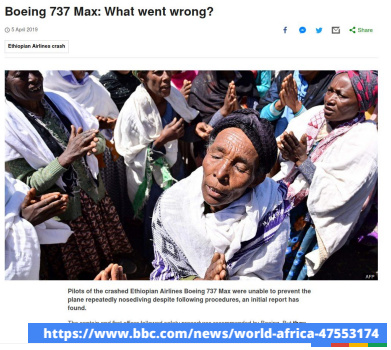
- **The F1 score conveys the balance between the precision and the recall.**

F1 Score = 2*((precision*recall)/(precision +recall))



Implications of Evaluation (1/3)

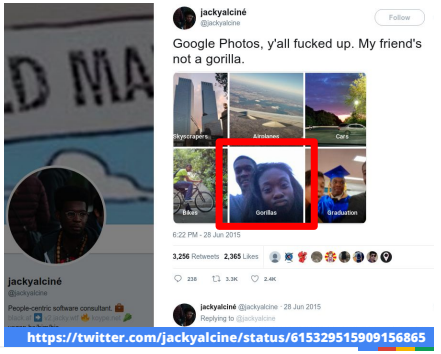
- Careful attention needs to be place on evaluation of learning algorithms, especially for sensitive domains
  - Domains with safety concerns



https://www.bbc.com/news/world-africa-47553174

Implications of Evaluation (2/3)

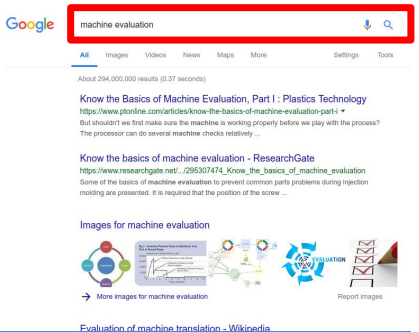
- Careful attention needs to be place on evaluation of learning algorithms, especially for sensitive domains
  - Domains where socially accepted norms are compromised



https://twitter.com/jackyalcine/status/615329515909156865

Implications of Evaluation (3/3)

- Some application domains are flexible
  - E.g. wrong search results are generally acceptable



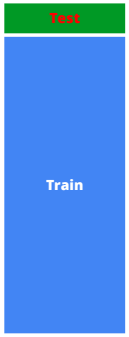
https://twitter.com/jackyalcine/status/615329515909156865

Model Evaluation Process

- A common evaluation technique for supervised learning involves the use of labeled data, split into training and testing datasets
  - Model uses training dataset to train estimators
  - Model uses testing dataset to determine effectiveness of predictions
  - Model evaluation involves determining the proportion of accurate predictions relative to the training set

Model Testing Techniques (1/3)

- **The holdout method reserves a representative proportion of the dataset as testing data**
  - While there is not prescribed training/testing ration, 80/20, 90/10 and 70/30 ratios are common
  - The samples might not be representative enough and so stratification might be necessary



Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (2/3)

- **K Fold cross-validation is a more viable alternative to the holdout method**
  - Dataset is randomly split up into "k" equal groups
  - One of the groups (fold) is used as the test set and the rest are used as the training set
  - The model is trained on the training set and scored on the test set.
    - Then the process is repeated until each unique fold as been used as the test set.
  - In 10-fold cross-validation, the dataset is split into 10 groups

Model Testing Techniques (3/3)

- **Leave-one-out is an extreme version of K Fold cross-validation**
  - It is essentially n-fold cross-validation where n = number of data points
    - Each instance is predicted, while training on the remaining (n-1) instances
  - Very comprehensive
  - Computationally intensive
  - The balance of training and testing sets is compromised

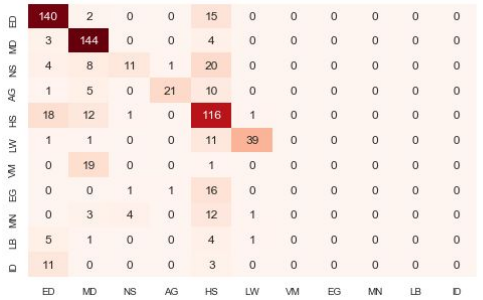
Model Evaluation Techniques (1/3)

- **The confusion matrix is commonly used during classification**
  - Actual labels are compared against predictions to determine the number of True Positives, False Positives, True Negatives and False Negatives

		Actual Classes	
		Pass	Fail
Predicted Classes	Pass	80	5
	Fail	3	10

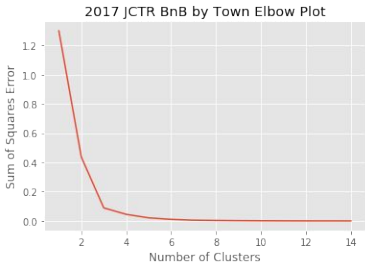
Model Evaluation Techniques (2/3)

- **Complex non-binary classification problems are sometimes easily interpreted using the confusion matrix**

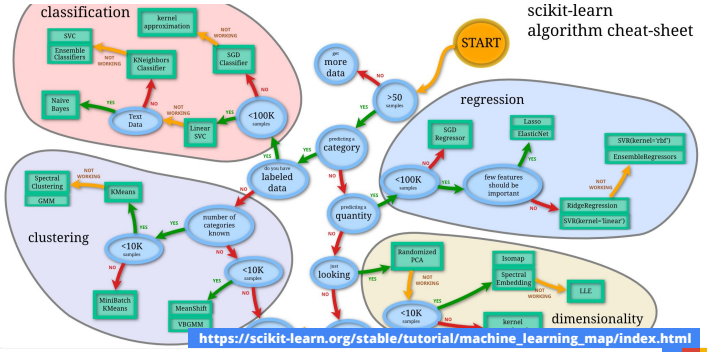


Model Evaluation Techniques (3/3)

- **Unsupervised learning approaches use specific evaluation techniques**
  - For instance with K Means clustering, the elbow method is commonly employed to determine the appropriate number of clusters
  - Remember that no training is involved during unsupervised learning



Summary (1/2)



Summary (2/2)

- **We will attempt to answer the following questions:**
  - What does the approach/technique/estimator/ algorithm do?
  - What are the inputs: x?
  - What are the outputs: f(x)?
  - How do we evaluate the approach/technique/estimator/ algorithm?

Q & A Session

- **Comments, concerns and complaints?**

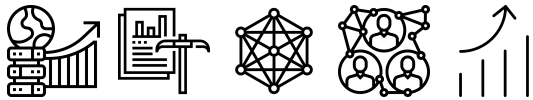
Lecture Series Outline

- **Part I: Machine Learning**
- **Part II: Datasets**
  - Scikit-learn Standard dataset
  - JCTR Basic Needs Basket
  - University of Zambia Electronic Theses and Dissertations
  - B.ICTs Ed. ICT 1110 Continuous Assessment Scores
  - Jupyter Notebook Walkthrough

Bibliography

[1] **Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017) Data Mining: Practical Machine Learning Tools and Techniques. Chapter 2** https://www.cs.waikato.ac.nz/ml/weka/book.html

[2] **An introduction to machine learning with scikit-learn** https://scikit-learn.org/stable/tutorial/basic/tutorial.html



CSC 5741
Lecture 6: Introduction to Machine Learning

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia