MIDDLE TENNESSEE
STATE UNIVERSITY

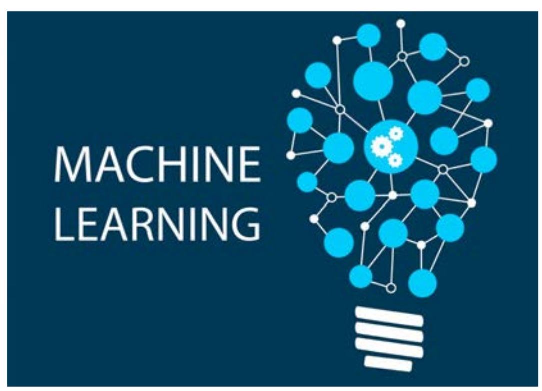
JONES COLLEGE OF BUSINESS

# Python for Data Science
# What is Machine Learning?

MIDDLE TENNESSEE
STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

## What is machine Learning?

- **Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

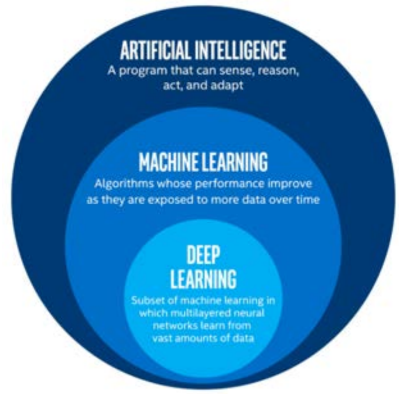



**Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves.

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# What is machine Learning?

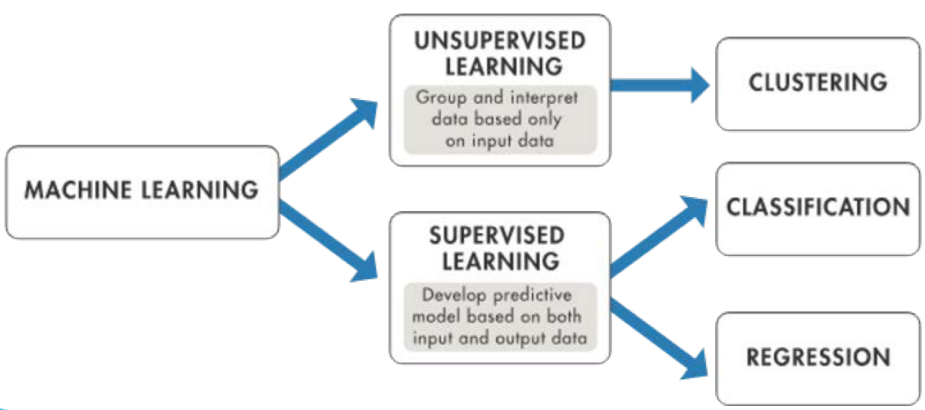
- **Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.




**Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves.

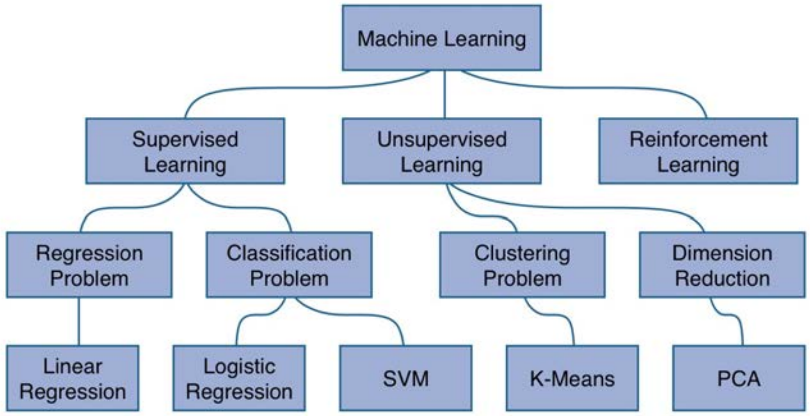
MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Different types of ML



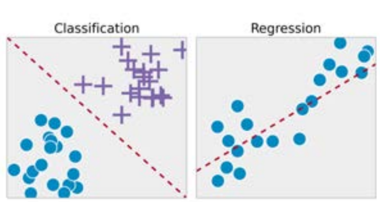
MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Different types



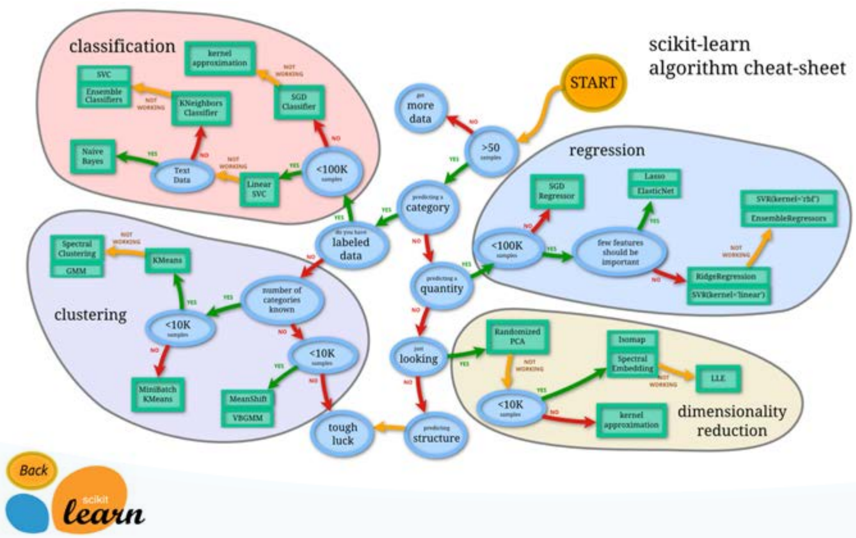

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Machine Learning Algorithms Mind-Map




MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Scikit-Learn for Machine Learning Algorithm Selection

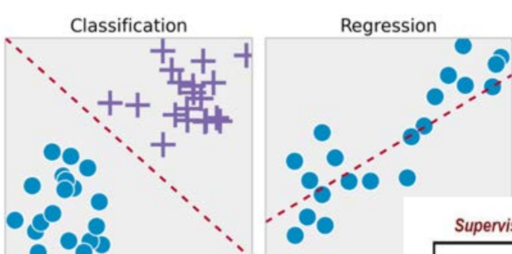


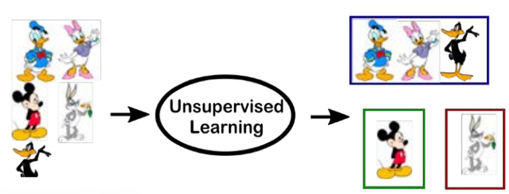

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Supervised vs. Unsupervised

Supervised

Unsupervised



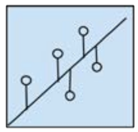

| | Supervised Learning | Unsupervised Learning |
|---|---|---|
| Discrete | classification or categorization | clustering |
| Continuous | regression | dimensionality reduction |

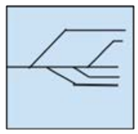
https://www.kdnuggets.com/2018/04/supervised-vs-unsupervised-learning.html

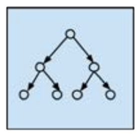
MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Types of Algorithms

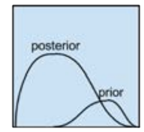
- Regression Algorithms
  - Ordinary Least Squares Regression (OLSR)
  - **Linear Regression**
  - **Logistic Regression**
  - Stepwise Regression
- Regularization Algorithms
  - **Ridge Regression**
  - **Least Absolute Shrinkage and Selection Operator (LASSO)**
  - Elastic Net
- Decision Tree Algorithms
  - **Classification and Regression Tree (CART)**
  - Iterative Dichotomiser 3 (ID3)
  - Chi-squared Automatic Interaction Detection (CHAID)


Regression Algorithms

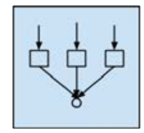

Regularization Algorithms

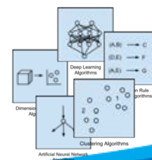

Decision Tree Algorithms

- Bayesian Algorithms
  - Naive Bayes
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
- Ensemble Algorithms
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Gradient Boosting Machines (GBM)
  - **Random Forest**
- Clustering Algorithms
- Association Rule Learning Algorithms
- Dimensionality Reduction Algorithms
- Deep Learning Algorithms
- Artificial Neural Network Algorithms
- …and many others



Bayesian Algorithms


Ensemble Algorithms

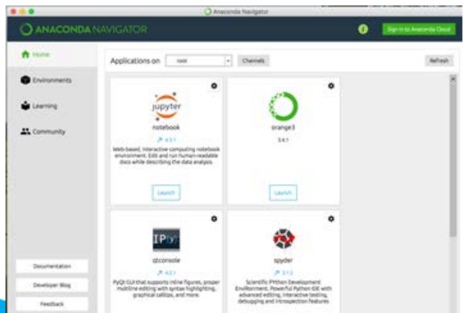


MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

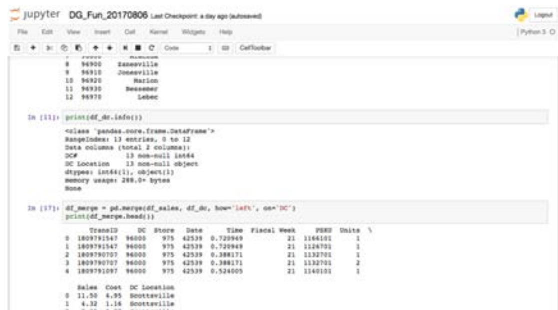
MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# What will we use?

Anaconda
https://anaconda.org/




Jupyter Notebooks
http://jupyter.org/

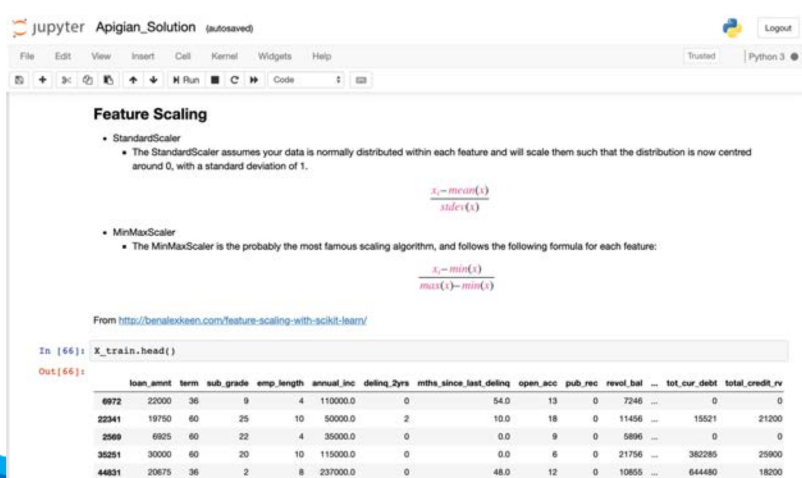



mtsu.edu/dsi

*mt_dsi['using data for good'].max()*

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Jupyter Notebooks




MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

## Use of Python Libraries


http://chris35wills.github.io/courses/pydata_stack/

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

## How do you build a model?

1.
2.
3.
4.
5.
6.

Modeling

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Model Building Process

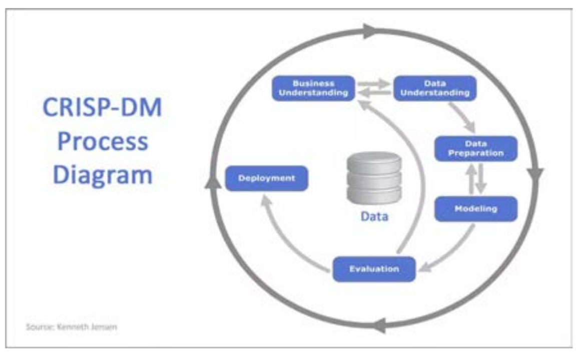
1. Select a model
2. Identify and select the data that fits that model
3. Transform the data
4. Identify a business problem
5. Train/Test Split
6. Run Model

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Data Analysis Process

1. Business Understanding
    a. Frame the problem and the REAL pain point
    b. Available resources, problems, goals
2. Data Understanding
    a. What data do you have available to you?
    b. Setup your workspace with tools or applications
        - Programming – Jupyter notebooks for Python or R Studio for R
        - BI/spreadsheets – Excel – PowerPivot - Tableau
    c. Import or download the data
    d. View, explore, and summarize the data
3. Data Preparation
    a. Clean up null values, outliers, mistakes
    b. Construct new data, transform or feature engineering
    c. Integrate and merge data
    d. Format data (strings, integers, floats, etc.)
    e. Create you X and y
4. **Modeling**
    a. Split your data (Train/Test Split)
    b. Setup models for machine learning/AI processes
    c. Can include visuals, dashboards or reports




5. Evaluation
    a. Fine tune your model
    b. Create a report of the findings
6. Deployment of models

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

## Appleton Lending Co

- **Operations**
  - Over 80% of the loans provided by Appleton are personal. These loans are mostly made by borrowers in order to consolidate debt or pay off credit cards, but they may be provided for numerous reasons such as weddings, vacations, and for small businesses.
- **Strategy**
  - Over the past two years, Appleton has provided over 3 billion dollars in loans. The company provides personal loans for amounts between $1,000 and $40,000 that can be repaid over time periods of 3 or 5 years. Appleton approves loans based on credit history, credit score, debt to income ratio (dti), and the amount of the loan applied for. Appleton is highly selective with the loans it accepts, with over an 80% denial rate over the past four years. This ensures that Appleton provides high quality opportunities for itself and for lenders.

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

## Appleton Lending Co

- After some negative publicity at the board level, Appleton is looking to refocus its efforts on providing high quality loans. They are wanting to better understand their customers and most importantly, the difference between good loans and bad loans.
- After understanding the type of customers that they serve, they would like to improve the company's ability to predict borrowers who will default on loans. Additionally, Appleton is interested in predicting how much a borrower would be able to pay back, regardless of how large of a loan they have applied for.

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

## Business Understanding

1. **Business Understanding**
   - Available resources, problems, goals
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment of models

- What are the available resources?
  – What are the key performance indicators (variables)?
- What are Appleton's expressed problems?
- What are Appleton's expressed and underlying goals?

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

- What data do you have available?

Why is this data too much?

| Feature | Description |
|---|---|
| | |
| member_id | A unique Appleton assigned Id for the borrower member. |
| loan_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| funded_amnt | The total amount committed to that loan at that point in time. |
| funded_amnt_inv | The total amount committed by investors for that loan at that point in time. |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| int_rate | Interest Rate on the loan |
| installment | The monthly payment owed by the borrower if the loan originates. |
| grade | Appleton assigned loan grade: A, B, C, D, etc. with A being the best |
| sub_grade | Appleton assigned loan subgrade: A1, A2, A3, etc. with A1 being the best |
| emp_title | The job title supplied by the Borrower when applying for the loan. |
| emp_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| home_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| annual_inc | The self-reported annual income provided by the borrower during registration. |
| verification_status | Indicates if income was verified by Appleton, not verified, or if the income source was verified |
| issue_d | The month which the loan was funded |
| loan_status | Current status of the loan |
| pymnt_plan | Indicates if a payment plan has been put in place for the loan |
| desc | Loan description provided by the borrower |
| purpose | A category provided by the borrower for the loan request. |
| title | The loan title provided by the borrower |
| zip_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| addr_state | The state provided by the borrower in the loan application |
| dti | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested Appleton loan, divided by the borrower's self-reported monthly income. |
| dti | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested Appleton loan, divided by the borrower's self-reported monthly income. |
| delinq_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| earliest_cr_line | The month the borrower's earliest reported credit line was opened |
| inq_last_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| mths_since_last_delinq | The number of months since the borrower's last delinquency. |
| mths_since_last_record | The number of months since the last public record. |
| open_acc | The number of open credit lines in the borrower's credit file. |
| pub_rec | Number of derogatory public records |
| revol_bal | Total credit revolving balance |
| revol_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| total_acc | The total number of credit lines currently in the borrower's credit file |
| initial_list_status | The initial listing status of the loan. Possible values are – W, F |
| out_prncp | Remaining outstanding principal for total amount funded |
| out_prncp_inv | Remaining outstanding principal for portion of total amount funded by investors |
| total_pymnt | Payments received to date for total amount funded |
| total_pymnt_inv | Payments received to date for portion of total amount funded by investors |
| total_rec_prncp | Principal received to date |
| total_rec_int | Interest received to date |
| total_rec_late_fee | Late fees received to date |
| recoveries | post charge off gross recovery |
| collection_recovery_fee | post charge off collection fee |
| last_pymnt_d | Last month payment was received |
| last_pymnt_amnt | Last total payment amount received |
| next_pymnt_d | Next scheduled payment date |
| last_credit_pull_d | The most recent month Appleton pulled credit for this loan |
| collections_12_mths_ex_me | Number of collections in 12 months excluding medical collections |
| mths_since_last_major_derc | Months since most recent 90-day or worse rating |
| policy_code | publicly available policy_code=1<br>new products not publicly available policy_code=2 |
| application_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| acc_now_delinq | The number of accounts on which the borrower is now delinquent. |
| tot_coll_amt | Total collection amounts ever owed |
| tot_cur_bal | Total current balance of all accounts |
| total_credit_rv | Total revolving high credit/credit limit |
| loan_is_bad | True if Borrower defaulted on loan. False if loan was good. |

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

- What is the RIGHT available data?

```
df_loandata = pd.read_csv('data/Loan_Data.csv', index_col = 0, header = 0)
df_loandata.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 49870 entries, 149512 to 4076727
Data columns (total 20 columns):
loan_amnt                      49870 non-null int64
term                           49870 non-null int64
sub_grade                      49870 non-null object
emp_length                     49870 non-null int64
home_ownership                 49870 non-null object
annual_inc                     49870 non-null float64
purpose                        49870 non-null object
delinq_2yrs                    49870 non-null int64
mths_since_last_delinq         21790 non-null float64
open_acc                       49870 non-null int64
pub_rec                        49870 non-null int64
revol_bal                      49870 non-null int64
total_acc                      49865 non-null float64
collections_12_mths_ex_med     49870 non-null int64
mths_since_last_major_derog    49870 non-null int64
acc_now_delinq                 49870 non-null int64
tot_coll_amt                   49870 non-null int64
tot_cur_debt                   49870 non-null int64
total_credit_rv                49870 non-null int64
loan_status                    49870 non-null object
dtypes: float64(3), int64(13), object(4)
memory usage: 8.0+ MB
```

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Supervised Learning Models

- Which tests will we conduct?
- Is it a bad loan?
  - Logistic Regression (prob.)
  - Decision Tree
  - Random Forest (ensemble)
- How much to loan?
  - Regression
  - Ridge Regression (predict)
  - Lasso Regression (sig. features)

# Confusion Matrix

| | | Predictions | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| Actual | 0 | 11354 | 1278 | 12632 |
| | 1 | 1798 | 527 | 2325 |
| | | 13152 | 1805 | |

| | | Predictions | | |
|---|---|---|---|---|
| | | No | Affair | |
| Actual | Not a Bad Loan | TP | FN | 12632 |
| | Bad Loan | FP | TN | 2325 |
| | | 13152 | 1805 | |

# Recall

| | | Predictions | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| Actual | 0 | 11354 | 1278 | 12632 |
| | 1 | 1798 | 527 | 2325 |
| | | 13152 | 1805 | |

| | | Predictions | | |
|---|---|---|---|---|
| | | No | Affair | |
| Actual | Not a Bad Loan | **TP** | **FN** | 12632 |
| | Bad Loan | FP | TN | 2325 |
| | | 13152 | 1805 | |

The recall is the ratio

- tp / (tp + fn)
- where tp is the number of true positives
- fn the number of false negatives.
- The recall is intuitively the ability of the classifier to find all the positive samples.
- 11354 / 12632 = 0.90
- 1278 / 12632 = 0.23

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# Precision

| | | Predictions | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| Actual | 0 | 11354 | 1278 | 12632 |
| | 1 | 1798 | 527 | 2325 |
| | | 13152 | 1805 | |

| | | Predictions | | |
|---|---|---|---|---|
| | | No | Affair | |
| Actual | Not a Bad Loan | TP | FN | 12632 |
| | Bad Loan | FP | TN | 2325 |
| | | 13152 | 1805 | |

The precision is the ratio:

- tp / (tp + fp)
- tp is the number of true positives
- fp the number of false positives.
- The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

11354 / 13152 = 0.86

1798 / 13152 = 0.29

MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

# F1

sklearn.metrics. **f1_score** (*y_true, y_pred, labels=None, pos_label=1, average='binary', sample_weight=None*)
[source]

Compute the F1 score, also known as balanced F-score or F-measure

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

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

- F1 = 2 * (0.86 * 0.90) / (0.86 + 0.90)
- F1 = 2 * (0.7452) / (1.76)
- F1 = 0.86