U-4 → Predictive Big Data Analytics in Python.

- Python
  - → high-level scripting language.
  - → true object-oriented language
  - → Guido van Rossum (developed by)
  - → 2008
  - → no need to end with special characters.

```
print ("Hello World")
Hello World
```

* Features -
  - high level programming
  - interactive
  - object-oriented scripting.
  - simple & easy
  - portable
  - free & open source.
  - perform complex tasks
  - run equally on different platforms.
  - vast range of libraries.

Adv → . ease
. min time
. modular & object-oriented
. Large community of users.
. large standard & user-contributed library.

Disadv → . interpreted & therefore slower than compiled language.
. decentralized with packages.

Page No.
Date

* Python Libraries -

1) Tensorflow -

- high performance numerical computations. with around 35,000 comments
- across various scientific fields.
- defining & running computation. that involves tensors, ~~produce~~ produce a value

Features → better computational graph visualization.
→ reduce error by 50 to 60% in neural ML.
→ Libraries backed by Google.
→ Quicker updates & frequent new releases.
→ applⁿ → speech & image recognition.
→ text-based applications.
→ time-series analysis
→ video detection.

2) Scipy -

- free & open-source Python library for data science.
- high level computations.
- 19,000 comments → Github.
- scientific & technical computations.
- extends Numpy & provides many user-friendly & efficient for scientific calculations.

Features → • collection of algorithms
• High level commands
• Data manipulation & visualization
• multidimensional image processing
• Inbuilt-functions → solving differential equations

Applications → • multiple dimensional image operations
• diff equations & fourier
• optimization algo
• Linear algebra.

3) Numpy-

- fundamental package for numerical computation.
- n-~~array~~ dimensional array object :: (powerful)
- 18,000 comments on Github.
- general-purpose array processing package.
- address the slowness problem partly by processing multidimensional arrays.

Features → fast
→ array oriented → better efficiency
→ object-oriented approach.
→ compact & fast.

Appln → data analysis
→ N-dimensional array
→ form base for other libraries
→ replacement of MATLAB

4) ~~Attin~~ Pandas

- python data analysis
- python lib → data science Numpy → matplotlib.
- 17,000 comments → Github
- fast, flexible, data structures & data frames EDS.
- structured data easily.

Features → Syntax & functionalities.
→ create own functions.
→ High-level abstraction
→ high level data structures & manipulation tools.

Applications → Data wrangling & data cleaning.
→ ETL & loading CSV files
→ variety of academic & commercial areas.
→ Time series-specific functionality

5) Matplotlib-

- powerful & beautiful visualizations.
- github community → 26,000
- graphs & plots → extensively used for data visualizats
- provides an object-oriented API.

Features → MATLAB replacement → free & open source.
→ supports dozens of backends & output types.
→ Pandas → used as wrappers → MATLAB API
→ low memory consumption & better runtime.

Applications → correlation analyses of variables
→ 95% confidence intervals of models.
→ outliers detection.
→ visualize distribution of data.

6) Keras.

- used for deep learning & neural network modules
- supports Tensorflow & Theano backends.

Features → provide vast prelabeled datasets.
→ various implemented layers & parameters contains

Applications-

- deep learning models → pretrained weights.
- models → predict or extract features.

7) Scikit-Learn
- ML algo (provided)
- interpolated into Numpy & SciPy.

Application → clustering → classification → regression
regression → model selection → dimensionality reduction.

8) PyTorch -
- Python based scientific package → power of graphics.
- deep learning research platform → flexibility & speed.

Appln → High-level features → deep neural network → tape-based system.
Appln → tensor computations → GPU acceleration

9) Scrapy -
- Popular, Fast, open-source web crawling framework.
- extract data from web page → XPath.

Appln → Building crawling programs (spider bots)
gather data from APIs.
Don't Repeat Yourself Principle.

10) Beautiful Soup -
- web crawling & data scraping.
- collect data → without proper CSV or API
- help collect & required format (arrange).

Page No.
Date

* Data Preprocessing -

1) Data Cleaning → missing values, noisy data or inconsistency
2) Data Integration → diff representation put together & conflicts with data are resolved.
3) Data transformation → normalized, aggregated & generalized.
4) Data Reduction → reduction of number of values of a continuous attribute.
5) Data Discretization → dividing range of attributes.
   - dealing with continous values.
   - mapping
   - (numerical data)

- Handle Missing Values -
  - Ignoring the Tuple
  - Manually filling in missing value.
  - Using global constant to fill missing value.

Analytics Type (Data Processing)

Page No.
Date

Predictive

→ predict with confidence
- smarter decisions
- improve business outcomes.

→ likelihood different samples. (find)
→ calculates live transactions.
→ utilize variety of variable data
→ variability → component data
→ big basket data
→ validate findings.
→ monitored → desired results.
→ examples → social media analysis, weather, retail, health care & fraud detection.

Descriptive

→ gathering, organizing tabulating & depicting data.
→ relation b/w product / service.
→ model → organize a customer by their personal preferances.
→ Business intelligence → sense of data. communication data → this solves provides historical insights about company
→ examples → reports that
→ data visulization
↳ easier communication.

Prescriptive

→ suggests course of action.
→ finding optimal solution to problem.
→ what-might-happen Analysis
→ determine best course of action
→ examples → Traffic applications, Product optimization & Operational Research

Page No.
Date

* Association Rule -

- unsupervised learning method.
- descriptive method → discover rel^n b/w hidden in large dataset.
- used to mining transactions in databases.
- FREQUENTLY BROUGHT TOGETHER.
- goal → discover interesting relationships among items.

- Support → popularity of product of all product transactions.

$$\text{support}(A) = \frac{\text{No. of transaction in which A appears}}{\text{Total no. of transactions.}}$$

- confidence → likelihood of purchasing both products A & B.

$$\text{confidence}(A \to B) = \frac{\text{No. of transaction includes both A \& B}}{\text{No. of transaction include only product A}}$$

- Itemset - set of one or more items.
- Frequent Itemset - an itemset whose support is greater than or equal to minsup threshold.

Page No.
Date

* Market Basket Analysis -

- determine what products ~~are~~ customers purchase together.
- name → idea of customers throwing all purchases into shopping (market basket) during grocery shopping. cart.
- Association Analysis & Frequent itemset mining.
- creates If-Then scenario rules

If item A is purchased then item B is likely to be purchased.

Rule → If {A} Then {B}

- Algorithm → Association Rule & Apriori

Application → • Retail • Telecommunications • Banks
• Insurance • Medical

* Apriori Algorithm -

Improve efficiency → • Partioning • Reducing • Sampling • Hashing • Dynamic counting.

- classic ML algorithm.
- designed to work on databases covering transactions.
- aimed to find subsets which are common to at least a min. number.

Steps → 1) computing support for each individual item.
2) Deciding on support threshold
3) selecting frequent items
4) Finding support of frequent itemsets
5) Repeats for larger sets
6) Generation Association Rules & compute confidence
7) compute lift

$$Lift = \frac{P(X \cap Y)}{P(X) * P(Y)}$$

Adv → easy
↳ Join & Prune (easy on large itemsets)

Disadv → requires high computation
↳ entire database has to be scanned.

Appln → education • Medical Field • Forestry

Page No.
Date

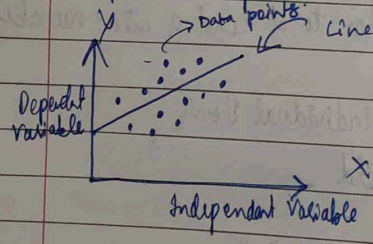
* Regression.

→ data mining function → predicts a number.
→ profit, sale, house values, temperature or distance can be predicted.
→ dataset in which target values are known.
→ good choice → all predictor variables are continously valued.
→ for input x → output continous → regression problem.

- Dependant Variable
- Independant Variable
- Outliers
- Multicollineasity
- Underfitting & Overfitting

1) Linear Regression -

- ML algorithm → supervised learning.
- predict dependent variable based on independant



Disadv → • outliers affect
• over simplifies real-world problems.

- regression line is best fit line for model.
- X → independant variable.
  Y → output.
- $Y = \beta_0 + B_1X_1 + B_2X_2 + \beta_3 X_3 + \cdots \beta_n X_n + \epsilon$

Page No.
Date

2) Simple Linear Regression (SLR) -

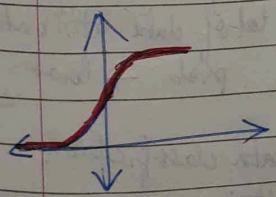
- one independent (or input) variable.
- no. of litres of petrol & kilometers driven. (Univariate Regression).

3) Multiple Linear Regression (MLR) -

- more than one independent variables.
- no. of litres of petrol, age of vehicle, speed & kilometers driven.
- Multivariate Regression.
- multiple inputs & multiple possible outputs.

3) Logistic Regression -

- used to model & estimate probability of an event.
- estimate probability of an event occuring → value (independent)
- is a curve.
- sigmoid curve or S-curve in short
- outcome data points → 0 or 1
- 0 → totally uncertain & 1 → certain
- solve classification issues.
- predictive analytics.
- predict likelihood of categorical dependent variable



1) Binomial
2) Multinomial
3) ordinal

* Naive Bayes:

- uses relation between probabilities of events for classification
- supervised learning algorithm
- solves classification problems
- Based on Bayes Theorem.
- Used for Text classification - (High dimensional training dataset)
- simple, effective, quick predictions.
- probalistic classifier.
- spam filtering, sentiment analysis of classifying articles

* Bayes Theorem -

determine probability of hypothesis with prior knowledge (conditional probability).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- $P(A|B)$ → Posterior prob. → prob of hypothesis A on event B
- $P(B|A)$ → likelihood prob → hypothesis → true (prob)
- $P(A)$ → Prior Probability → before evidence
- $P(B)$ → Marginal Probability → prob. of evidence.

Adv -

- handle missing data attribute values.
- Handle irrelevant input variables.
- simple
- fast
- better → categorical variables.

Disadv -

- assume → data are independent
- not reliable for prob. estimate
- only with categorical variables
- lot of data attributes → prob → loses near zero

Appl^{n} → • Credit Scoring • Medical data classification
• Sentiment analysis • Spam filtering

Page No.
Date

Decision Trees:-

- uses concept of trees to structure the given info in sequence of decisions & sequences.
- ML algorithms → use a hierachical structure of decisions to model rel^n input features & target variables.

- tree where node → feature (attribute)
  each link (branch) → decision (rule) & leaf represent an outcome (categorical or continuous)

- consists of
  - Nodes → Test for value of certain attribute
  - Edges → corresponds to outcome of test.
  - Leaves → predict outcomes.

Steps → • splitting
• Pruning
• Tree selection

Adv →
- simple & easy.
- both nominal & numerical attributes.
- handling → errors
- handling → missing values
- non parametric method.
- self-explanatory.

Disadv →
- target attribute → discrete values.
- difficult → solve XOR
- less appropriate → estimation
- prone to errors.
  ↳ classification.

Page No.
Date

* More algos-

1) ID3 Algo → Iterative Dichotomiser 3
   → divide into 2 parts, classes or groups
   → entropy → measure of impurity → decision at each
   → selects features with high info → gain to split data & grows tree reaches a stopping criteria.
   → no backtracking

2) C45 → gain ratio to address bias towards attribute → large no of val
   → discrete & continous attributes & support missing values.

3) CART → (Classification & Regression Trees)
   → splitting data based on GINI impurity or min squared errs.

4) Random Forest → ensemble method → combines multiple trees
   → random subsets of data & of features
   → prediction of individual trees.

5) XGBoost → finds optimal splits at each other.

6) Gradient boosting → another ensemble learning method that combines multiple decision trees.