**BIO542: Machine Learning for Biomedical Applications (MLBA)**

**(7th October 2024, Mid-sem exam)**

**Maximum Marks: 60 Duration: 90 Minutes**

**Instructions:** This question paper has two sections, A and B. Attempt any 14 questions from section A, each question carries 2 marks (A total of 28 marks). Attempt any 8 questions from section B, each question carries 4 marks (Total 32 marks). Write all answers in the answer sheet only.

**Section A – (2 marks each)**

1. **Name any two unicellular organisms.**

Answer 1: Paramecium, Amoeba, Yeast, Euglena, Bacteria

1. **Write the number of chromosomes in a) Humans and b) coronavirus.**

Answer 2: a) Humans: 46 chromosomes or 23 pairs; b) coronavirus: 1 chromosome or no chromosome (both correct)

1. **Write the name of a restriction enzyme and its recognition sequence**

Answer 3: Enzyme - EcoR1; Recognition Sequence - GAATTC

Or

Enzyme - HIND III; Recognition Sequence - AAGCTT

Or

Enzyme - BAMH1; Recognition Sequence – GGATCC

1. **Name any two algorithms used for unsupervised learning.**

Answer 4: K-means clustering, Mean shift clustering, DBSCAN clustering, Agglomerative Hierarchical clustering, Gaussian mixture

1. **What is the range for a TanH function?**

Answer 5: (-1, 1)

1. **Which two shortcomings of Matplotlib have been fixed in Seaborn?**

Answer 6:

Graphical user interface, text, application, email

Description automatically generated

1. **List two operations of List in Python.**

Answer 7: Append, insert, index, count, sort, reverse, remove, pop, extend, slicing, concatenation, repetition, membership test, lengths

1. **Write two example DNA sequences in FASTA format.**

Answer 8: >Sequence\_1

ATGCGTACGTAGCTAGCTAGCTAGCTAGCTAGC

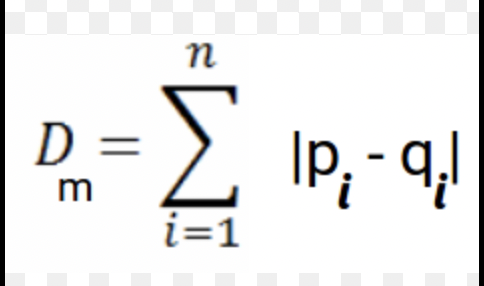
>Sequence\_2

TGCATGCTAGCTAGCTAGCGTACGTAGC

1. **Write the name of a web server used to predict ATP-interacting residues.**

Answer 9: ATPint

1. **Write the formula for computing Manhattan distance.**

Answer 10:

where, P & Q are two data points , n = number of data points, i = ith data point

1. **List the names of any two web servers used for exogenous antigen processing.**

Answer 11: Propred: Promiscuous MHC-II binders

MHCBN: Database of MHC

Pcleavage: Proteome cleavage sites

TAPpred: for predicting TAP binders

CTLpred: Prediction of CTL epitopes

1. **Which cytokine or interleukin is overexpressed in Covid-19 patients?**

Answer 12: Interleukin 6 (IL 6)

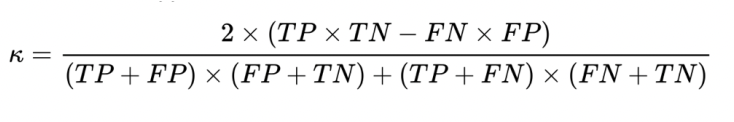
1. **Which database was used to create datasets for web servers ESLPred and PSLpred?**

Answer 13: ESLPred = SWISSPROT database release 33.0

PSLpred = SWISSPROT database release 40.29

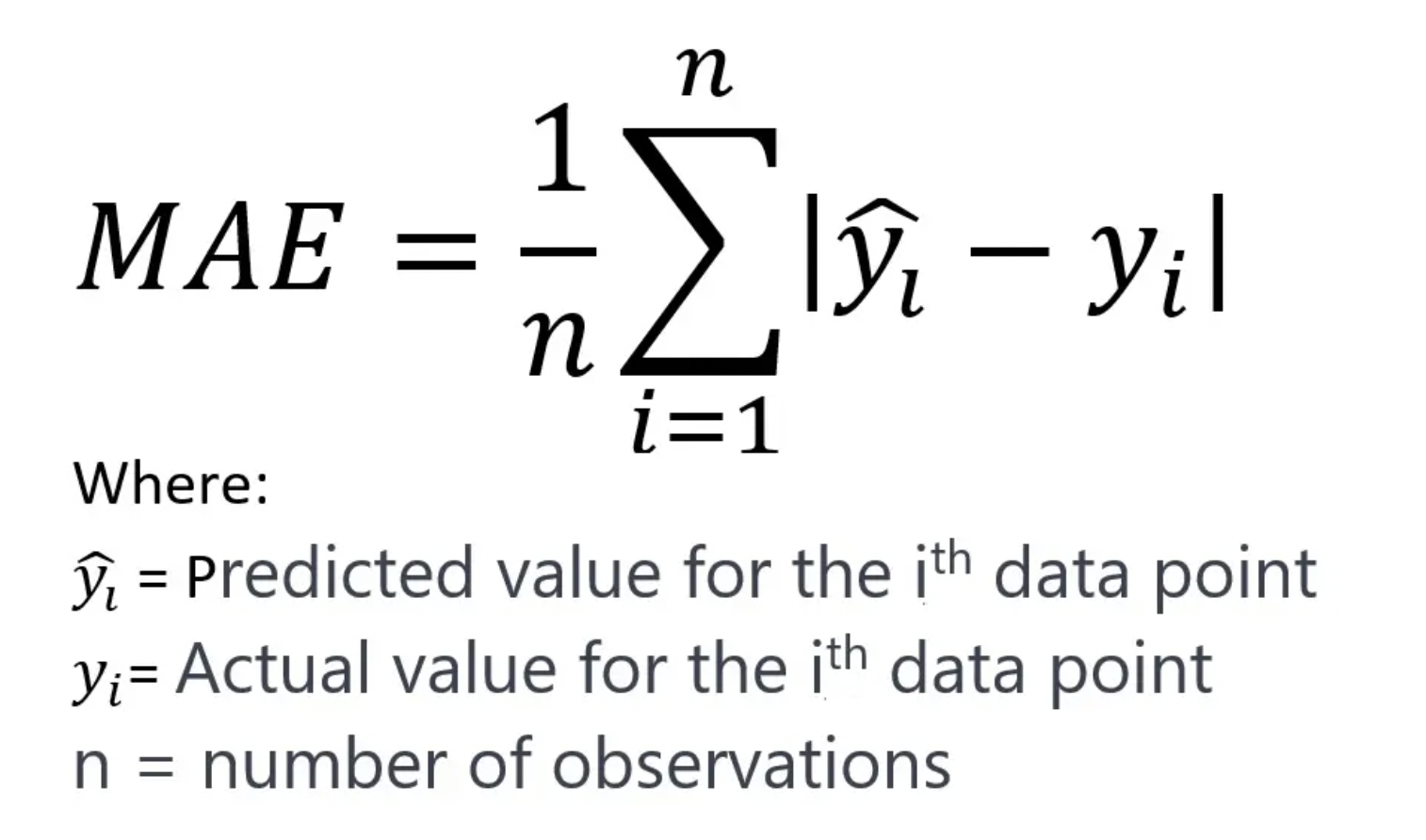
1. **Write the formula for computing Cohen’s Kappa.**

Answer 14:



1. **Write the formula for calculating mean absolute error (MAE) between predicted and actual values.**

Answer 15: MAE/AAE is mean of absolute errors within actual and predicted value.



1. **What does ADMET & QSAR stand for?**

Answer 16: ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity

QSAR: Quantitative Structure-Activity Relationship

1. **Write one major application of the web server ChAlPred.**

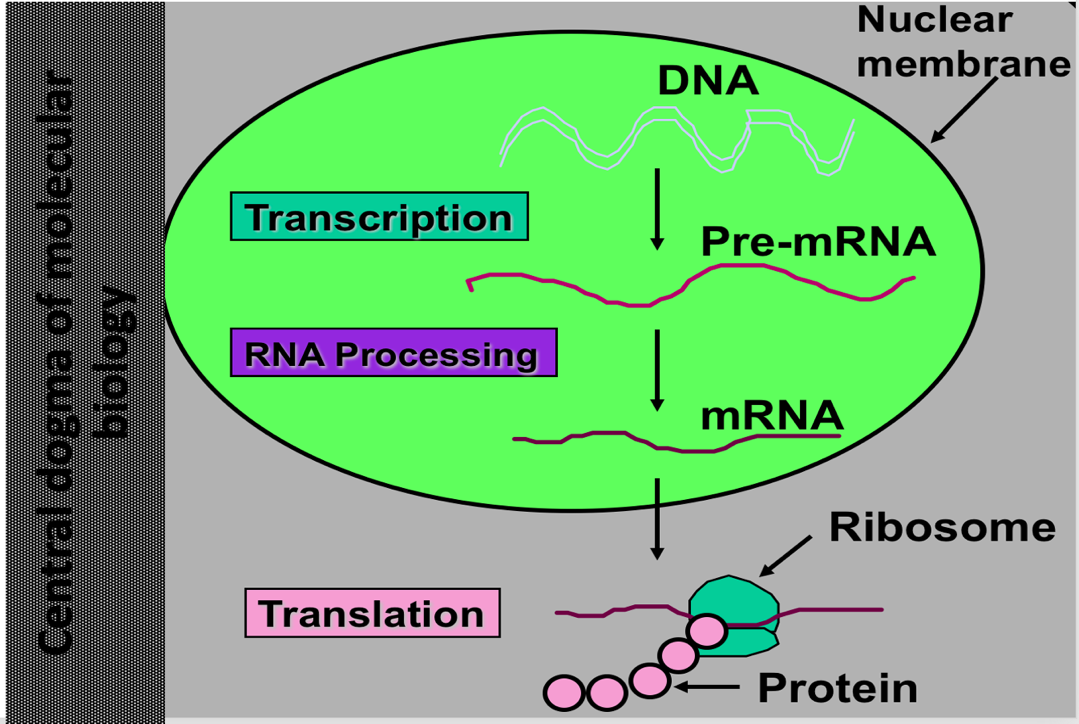
Answer 17: It is used to predict the allergenicity of chemical compounds.

1. **Write the name of two formats used for storing chemicals.**

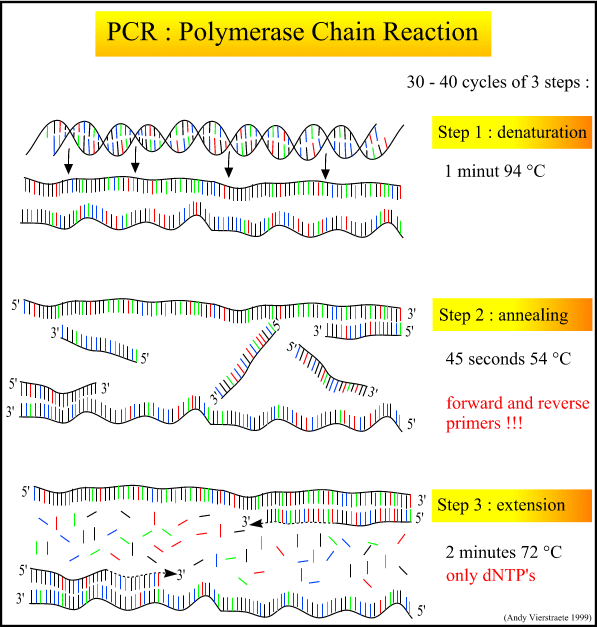
Answer 18: SDF File (Standard Data File) , SMILES (Simplified Molecular Input Line Entry System) or MDL mol format:

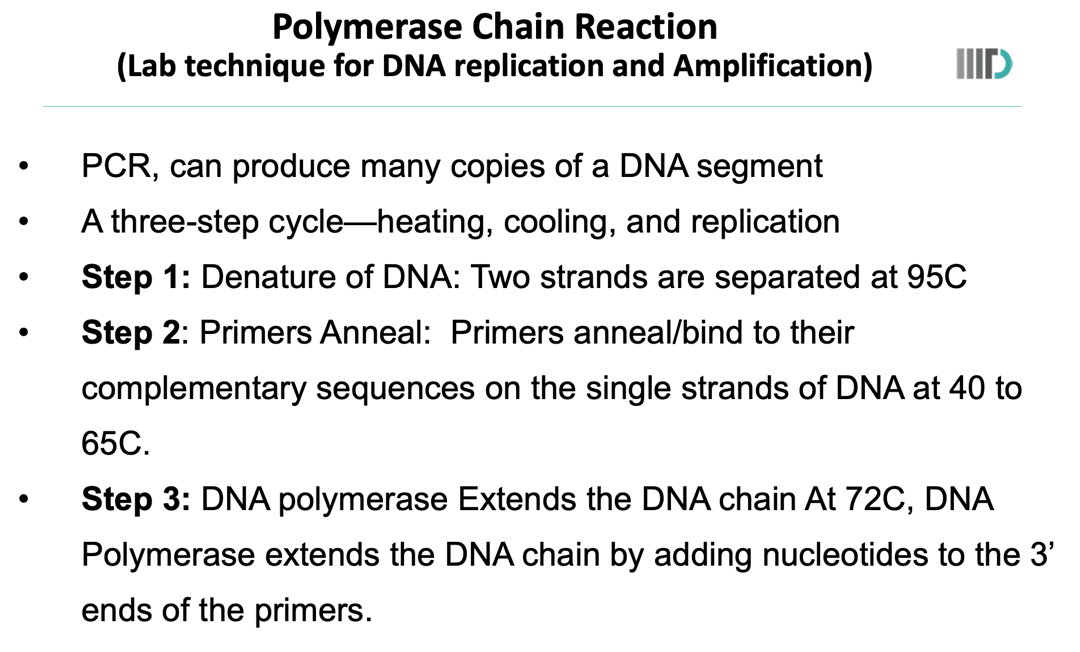
**Section B – (4 marks each)**

1. **Draw a diagram of the central dogma of molecular biology with labels.**



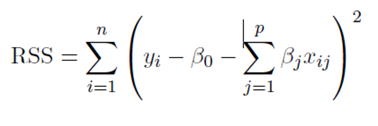
1. **Show the Polymerase Chain Reaction graphically and describe each step in detail.**



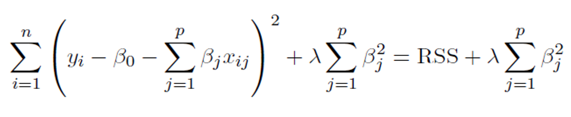


1. **Write the formula for lasso and ridge regression in which condition they become OLS.**

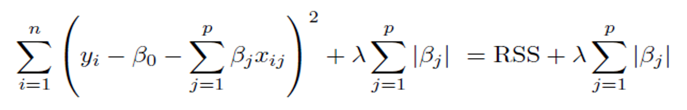
Formula for OLS:



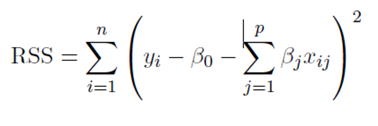
Formula for Ridge Regression:



Formula for Lasso Regression:



Lasso and Ridge regression continue to become OLS when λ=0, for which the formula will be:



1. **Give examples for the following operations in tuples: a) indexing, b) slicing, c) repetition, d) concatenation.**
2. Indexing:

my\_tuple = (10, 20, 30, 40, 50)

element = my\_tuple[2]

print(element) # Output: 30

1. Slicing:

my\_tuple = (10, 20, 30, 40, 50)

sub\_tuple = my\_tuple[1:4]

print(sub\_tuple) # Output: (20, 30, 40)

1. Repetition:

my\_tuple = (10, 20)

# Repeating the tuple 3 times

rep\_tuple = my\_tuple \* 3

print(rep\_tuple) # Output: (10, 20, 10, 20, 10, 20)

1. Concatenation

tuple1 = (10, 20)

tuple2 = (30, 40)

# Concatenating the tuples

concat\_tuple = tuple1 + tuple2

print(concat\_tuple) # Output: (10, 20, 30, 40)

1. **Write code using the Python library Scikit-learn to train the KNN model.**

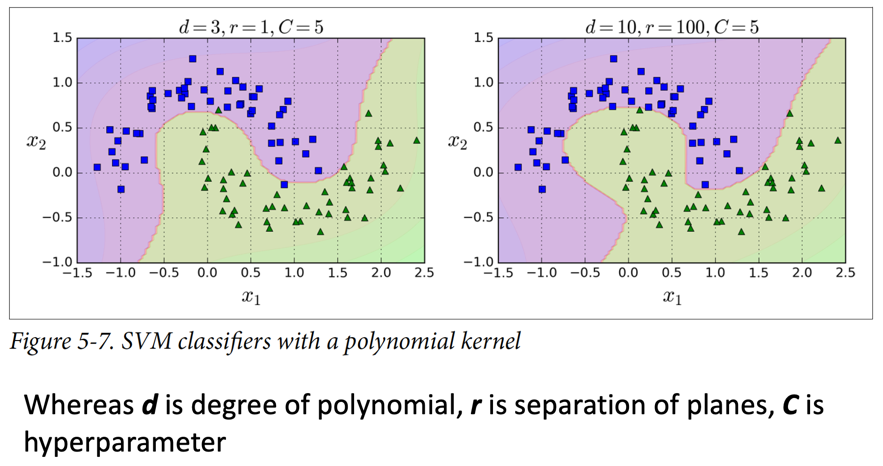
from sklearn.neighbors import KNeighborsClassifier

clf = KNeighborsClassifier()

clf.fit(X\_train, y\_train)

y\_pred = clf.predict(X\_test)

1. **In SVM, show the polynomial kernel graphically for degrees d = 3 and d = 10.**



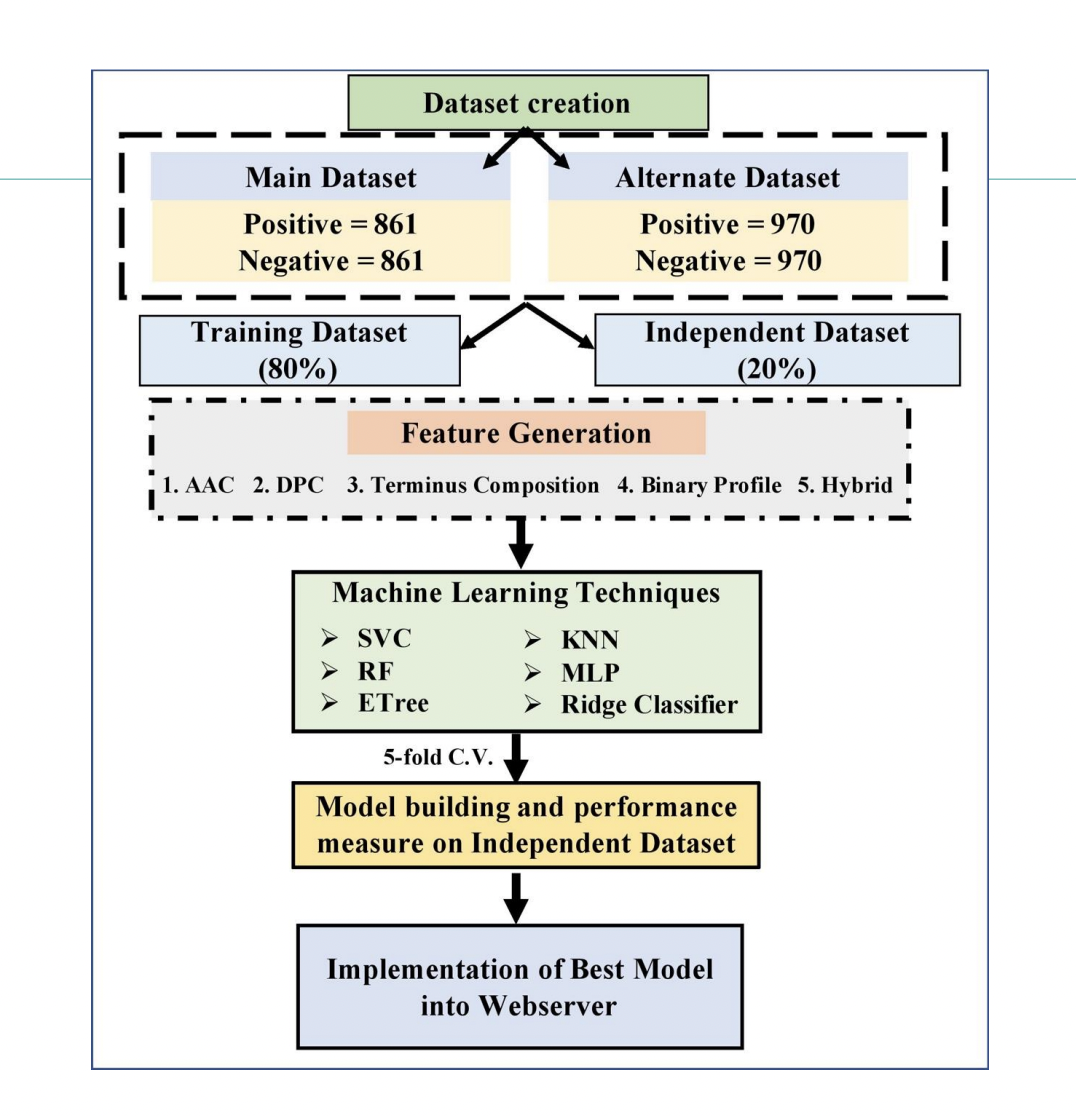
1. **If TP = 21, FP = 30, TN = 12, FN = 5, compute Sensitivity, Specificity, Accuracy, PPV and NPV.**

Answer 7:

Sensitivity= *TP/*(*TP*+*FN)* =21/(21+5)=21/26 ≈0.8077 or 80.77%  
Specificity= TN/(TN+FP) = 12/(12+30) = 12/42 ≈0.2857 or 28.57%  
Accuracy=*(TP*+*TN)/*(*TP*+*TN*+*FP*+*FN)*=(21+12)/(21+12+30+5)=33/68 ≈0.4853 or 48.53%  
PPV=*TP/*(*TP*+*FP)*=21/(21+30)=21/51 ≈0.4118 or 41.18%  
NPV=*TN/*(*TN*+*FN)*=12/(12+5)=12/17 ≈0.7059 or 70.59%

1. **Show the algorithm of AntiCP2.0 by flowchart.**

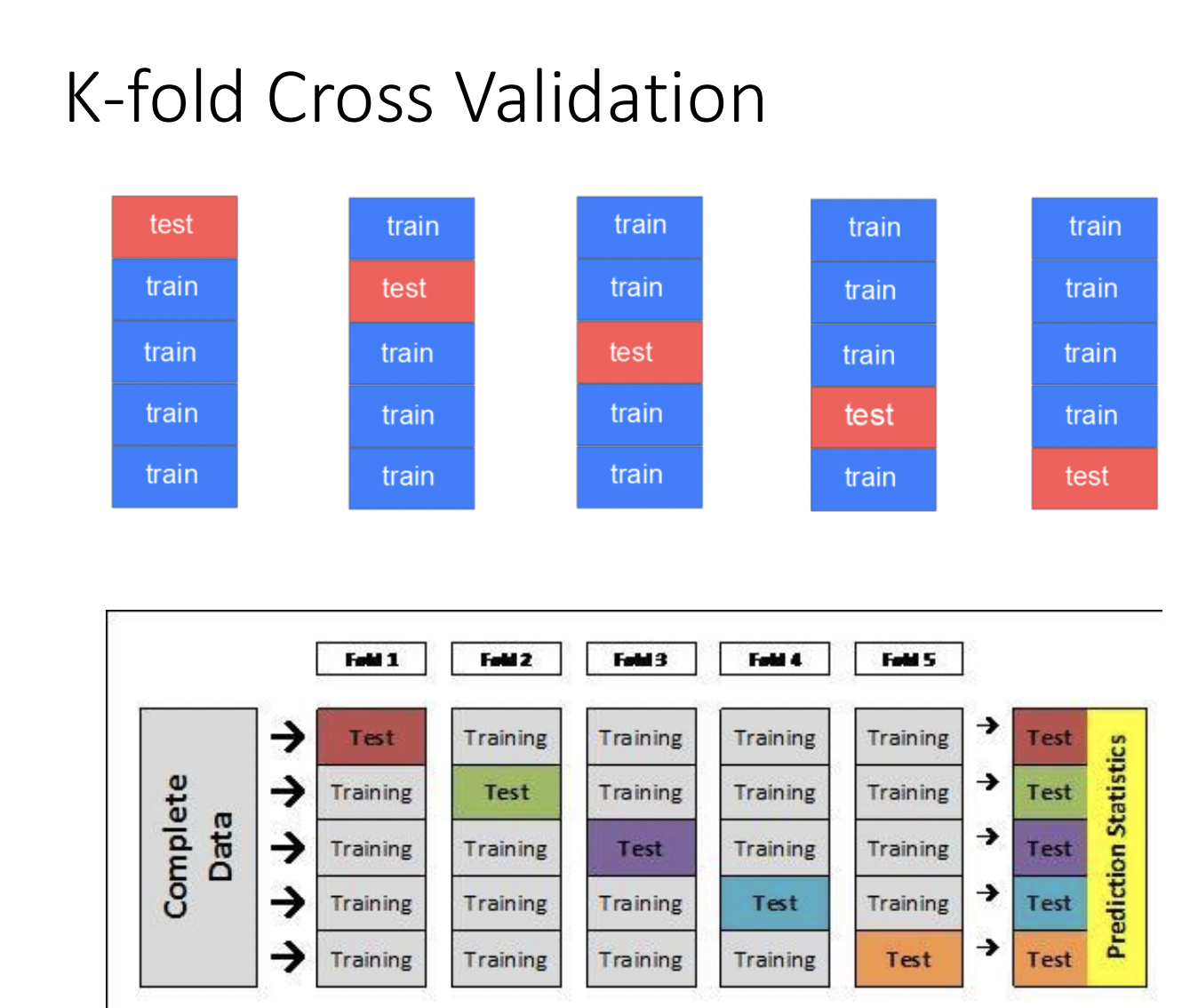
Answer 8:



1. **Show the 7-fold cross-validation technique graphically.**

Answer 9: Here k = 7

The following diagram is for 5 fold but the question will be checked according to 7 folds



1. **Briefly describes three types of drugs.**

Answer 10:

**• Therapeutic Drugs:** These drugs are used for the treatment, prevention, or

management of various medical conditions.

**• Recreational Drugs:** These drugs are often used for non-medical purposes,

primarily for their psychoactive effects. Examples include alcohol, nicotine,

marijuana, and various illegal substances like cocaine and ecstasy.

**• Pharmaceutical Drugs:** Pharmaceutical drugs are substances that have undergone

extensive research, development, testing, and regulatory approval processes before

being marketed and prescribed for medical use.