DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

COURSE MATERIAL

Faculty Details

Name of the Faculty	:	Dr.C.Sunitharam, Ms.E.Padma, Dr.M.Gayathri
Designation	:	Assistant Professor
Department	:	CSE

Course Details

Name of the Course	:	BE
Branch	:	CSE
Year	:	IV
Subject Code	:	BCSF187T40
Title of the Subject	:	Computational Biology
Batch	:	2018-2019

Course Code: BCSF187T40

Course Title: COMPUTATIONAL BIOLOGY

PRE-REQUISITE

Knowledge and awareness of the basic principles of biology, Mathematics.

OBJECTIVES

1. Bioinformatics is the science of storing, extracting, organizing, analysing, interpreting and using information.
2. Approaches to the discipline of bioinformatics incorporate expertise from the biological sciences, computer science and mathematics.
3. Bioinformatics is designed for students. Interested in molecular biology and genetics, information technologies and computer science.

COURSEOUTCOME

1. knowledge and awareness of the basic principles and concepts of biology, computer science and mathematics
2. Problem-solving skills, including the ability to develop new algorithms and analysis methods.
3. An understanding of the intersection of life and information sciences, the core of shared concepts, language and skills the ability to speak the language of structure-function relationships, information theory and database queries.

UNIT-I OVERVIEW OF MEDICAL INFORMATICS

Healthcare functions and information technology, Key Players in Health Information Technology (HIT), Organizations involved with HIT. Public Health Informatics-Information systems in public health. Internet based consumer health information — telehealth and telemedicine.

UNIT-II CLINICAL DECISION-SUPPORT SYSTEMS

The Nature of clinical decision making, types of decisions, the role of computers in decision support-examples of clinical decision-support systems.

UNIT—III DATABASES IN BIOINFOMATICS

Biological databases- Types of databases- Examples of databases: GenBank (Genetic Sequence Databank)-NCBI (National Center for Biotechnology Information) -EMB (European Molecular Biological Laboratory)-SwissProt.

UNIT—IV ALGORITHMS IN COMPUTING BIOLOGY

Decision tree algorithm, Bayesian network: Bayes Theorem, Random Forest algorithm, Genetic Algorithm.

UNIT-V BIOMEDICALDATA

Their acquisition, storage and use, electronic health Record (EHR), Information Retrieval from Digital Libraries-PubMed, Cleveland, GENECARD.

TEXTBOOK

1. An introduction to bioinformatics algorithms by Neil C. Jones, Pavel Pevzner. MITPress.2004.
2. Biomedical Informatics: computer applications in Health care and Biomedicine (3rded), by Shortliffe EH, CiminioJJ., 2000, New York Springer-Verlag, ISBN 0-387-28986-0.

UNIT I

1.Introduction:

- Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and using information.
- Approaches to the discipline of bioinformatics incorporate expertise from the biological sciences, computer science and mathematics.
- Bioinformatics is designed for students interested in molecular biology and genetics, information technologies and computer science.

2. Health Information Technology:

- Health IT (health information technology) is the area of IT involving the design, development, creation, use and maintenance of information systems for the healthcare industry.
- Automated and interoperable healthcare information systems will continue to improve medical care and public health, lower costs, increase efficiency, reduce errors and improve patient satisfaction, while also optimizing reimbursement for ambulatory and inpatient healthcare providers
- The importance of health IT results from the combination of evolving technology and changing government policies that influence the quality of patient care

2.1. Types of Health Information Technology:

- The EHR is the central component of the health IT infrastructure. An EHR, or electronic medical record (EMR), is a person's official, digital health record and is shared among multiple healthcare providers and agencies.
- personal health record (PHR), which is a person's self-maintained health record, and the health information exchange (HIE), a health data clearinghouse or a group of healthcare organizations that enter into an interoperability pact and agree to share data between their various health IT systems.
- As a result of the mountain of patient information that healthcare organizations uses data analytics has taken a greater role in day-to-day operations. The ability to aggregate patient information, analyse it and then base treatments on the results fits in well with population health management (PHM) and value-based healthcare.
- Improvements in health technology include patient portals and physician practices.

- In comparison, modern portals provide more context to a patient's care. The portals let patients securely communicate with their physicians, pay bills, check services against what an insurance plan allows, download full medical records, order prescriptions and possibly interact with a chatbot for other services.
- More recent innovations in health IT technology include the greater use of the application program interface (API) to improve interoperability, the ability to access and interact with health data via mobile devices and further exploration of blockchain as a way to better access and secure medical records.
- Security and privacy rules have long guided healthcare organizations to provide patients access to their medical records, while also safeguarding that information.
- In response, traditional health IT systems often integrate with data security and cybersecurity technology.

2.2. Benefits of Health Information Technology:

While some critics say EHRs have led to clinicians spending more time entering data than conversing with patients and produced cumbersome federal regulations, there is broad consensus on the benefits of health IT.

The advantages include:

- The ability to use data analytics and big data to effectively manage population health management programs and reduce the incidence of expensive chronic health conditions;
- The use of cognitive computing and analytics to perform precision medicine (PM) tailored to individual patients;
- The ability to share health data among academic researchers to develop new medical therapies and drugs; and
- The rights of patients to obtain and use their own health data and collaborate in their own care with clinicians

3.Key Players in Health Information Technology (HIT)

3.1. Health Information Systems:

Health information systems strongly influence quality and efficiency of health care, and technical progress offers advanced opportunities to support health care.

Significance of information systems

- Information processing is an important quality factor, but an enormous cost factor as well. It is also becoming a productivity factor.
- Information processing should offer a holistic view of the patient and of the hospital.
- A hospital information system can be regarded as the memory and nervous system of a hospital

Information and communication technology has become economically important and decisive for the quality of health care.

3.2. Information Processing:

- Decisions of health care professionals are based on vast amounts of information about the patient's health state
- It is essential for the quality of patient care and for the quality of hospital management to fulfill these information needs

For example:

- When a patient is admitted to a hospital, a physician or nurse first needs information about the reason for patient admission and about the patient history.
- Incorrect reports, e.g. lab report, may lead to erroneous and even harmful treatment decisions
- Repeated examinations or lost findings have to be searched for, the costs of health care may increase
- Information should be documented adequately, enabling health care professionals to access the information needed and to make sound decisions

4. Organisations involved with HIT

- Health Information Technology (Health IT) is a broad term that describes the technology and infrastructure used to record, analyse, and share patient health data.
- Various technologies include health record systems, including personal, paper, and electronic; personal health tools including smart devices and apps; and finally, communities to share and discuss information.

- The purpose of Health IT is to provide better care for patients and help achieve health equity.
- Health IT supports recording of patient data to improve healthcare delivery and allow for analysis of this information for both healthcare practitioners and ministry of health/government agencies.
- Health IT improves the quality of healthcare delivery, increases patient safety, decreases medical errors, and strengthens the interaction between patients and healthcare providers.
- In low and middle-income countries (LMIC) the need for reliable and affordable medical record software is paramount.
- The OpenMRS community helps meet this specific need by developing and supporting the Open Medical Record System – an open-source electronic health record (EHR) platform, specifically designed for low-resource environments, and is completely free.
- The use of Health IT in medical clinics improves the quality of healthcare that is delivered by providing accurate patient records and allows doctors to better understand the patient's medical history

5. Public Health Informatics:
Objectives

- Define public health informatics, its goals and uses in the context of a CDC program.
- Identify three ways in which informatics capacity is critical to the effectiveness of a CD program in an e-health era

Apply informatics principles, methods and tools in planning for EHR-based CD surveillance program. Public health informatics is pragmatic, progressive, dexterous and adaptive.

Figure 1.1. Public Informatics

Value Case Exercise

- How do you best articulate the value case for the health care organizations that would provide the data for this program?
- How do you frame that value case in terms of their priorities and drivers?
- How you will ensure that data capture minimally impacts workflows, systems or both in the health care setting?

Data and Workflow Exercise

- Go to https://db.tt/j5OAsevw for the worksheet.
- Select a CD condition or indicator you want to work with.
- Respond to the worksheet questions as best as you can.
- Note where you are not sure of the answer. What are possible risks of not knowing?
- Share any insights or questions.

6. Information systems in public health:

- Systematic processing of information contributes to high quality patient care and reduces costs

The integrated processing of information is important because:

- All groups of people and all areas of a hospital depend on its quality
- The amount of information processing in hospitals is considerable
- Health care professionals frequently work with the same data

7. Internet based consumer health information:

- Information processing has to integrate the partly overlapping information needs of the different groups and areas of a hospital
- Systematic, integrated information processing in a hospital has advantages not only for the patient, but also for the health care professionals, the health insurance companies, and the hospital's owners
- Integration of information processing should consider not only information processing in one health care organization, but also information processing among different institutions

Raising the quality of patient care and reducing cost

- Systematic information processing is the key factor for raising quality and reducing costs
- Systematic in this context means purposeful and effective, and with great benefit regarding the costs
- Bearing this in mind, it is obvious that information processing in a health care institution should be managed systematically

8. Telehealth and telemedicine:

8.1. Barriers in Telehealth:

- Perspective of medical practitioners: Doctors are not fully convinced and familiar with e-medicine.
- Patients' fear and unfamiliarity: There is a lack of confidence in patients about the outcome of e-Medicine.

- Financial unavailability: The technology and communication costs being too high, sometimes make Telemedicine financially unfeasible.
- Lack of basic amenities: In India, nearly 40% of population lives below the poverty level.
- Literacy rate and diversity in languages: Only 65.38% of India's population is literate with only 2% being well-versed in English.
- Technical constraints: e-medicine supported by various types of software and hardware still needs to mature.
- For correct diagnosis and pacing of data, we require advanced biological sensors and more bandwidth support.

8.2. Telemedicine:

- Telemedicine is a subset of telehealth that refers solely to the provision of health care services and education over a distance, through the use of telecommunications technology.
- Telemedicine involves the use of electronic communications and software to provide clinical services to patients without an in-person visit.
- Telemedicine technology is frequently used for follow-up visits, management of chronic conditions, medication management, specialist consultation and a host of other clinical services
- In short, all telemedicine is telehealth, but not all telehealth is telemedicine. Both are part of the larger effort to expand access to care, make health management easier for patients and improve the efficiency of the healthcare delivery network.
- “Telematics for health is a composite term for both telemedicine and telehealth, or any health-related activities carried out over distance by means of information communication technologies.”

8.3. Advantage of Telemedicine System

- Easy access to remote areas
- Using telemedicine in peripheral health set-ups can significantly reduce the time and costs of patient transportation
- Monitoring home care and ambulatory monitoring
- Improves communications between health providers separated by distance
- Critical care monitoring where it is not possible to transfer the patient

UNIT II

2.Overview Scope of Clinical Decision Support Systems

Issues for success or failure Evaluation of Clinical Decision Support Systems Computing techniques used to create DSS Design Cycle for the development of DSS Early AI/Decision Support Systems.

2.1. Scope of Clinical Decision Support Systems

Definition

A clinical decision-support system is any computer program designed to help health professionals make clinical decisions. In a sense, any computer system that deals with clinical data or medical knowledge is intended to provide decision support. Three types of decision-support function, ranging from generalized to patient specific.

2.2 Characterizing Decision-Support Systems

- System function
- Determining what is true about a patient (e.g. correct diagnosis)
- Determining what to do (what test to order, to treat or not, what therapy plan …)
- The mode for giving advice
- Passive role (physician uses the system when advice needed)
- Active role (the system gives advice automatically under certain conditions)

2.2.1. Need for CDSS

- Limited resources - increased demand Physicians are overwhelmed.
- Insufficient time available for diagnosis and treatment.
- Need for systems that can improve health care processes and their outcomes in this scenario

2.2.2. Disadvantages of CDSS

- Changing relation between patient and the physician
- Limiting professionals’ possibilities for independent problem solving
- Legal implications - with whom does the onus of responsibility lie?

2.2.3. Issues for success or failure

- Evaluation of User Needs
- Top management support
- Commitment of expert
- Integration Issues
- Human Computer Interface
- Incorporation of domain knowledge

2.2.4. Evaluation of Clinical Decision Support Systems

- Criteria for success of CDSS
- Aspects for consideration during evaluation Criteria for a clinically useful DSS
- Knowledge based on best evidence
- Knowledge fully covers problem
- Knowledge can be updated
- Data actively used to draw from existing sources
- Performance validated rigorously

2.2.5. Aspects for Evaluation of a CDSS

- The process used to develop the system
- The system's essential structure
- Evidence of accuracy, generality and clinical effectiveness
- The impact of the resource on patients and other aspects of the health care environment

2.2.6. Computing techniques used to create DSS

- Machine Learning and Adaptive Computing
- Inductive Tree Methods
- Case Based Reasoning
- Artificial Neural Networks
- Expert Systems - Knowledge based Methods
- Rule based Systems

2.3. Assignment and Case Study had been given for the following:

- Collect Information about Clinical Support System through Web Site
- Write a case study for the data collection of various search data values
- Enumerate the features of Information System with Clinical Support System

UNIT III

3.Bioinformatics:

- In biology, bioinformatics is defined as, "the use of computer to store, retrieve, analyse or predict the composition or structure of bio-molecules". Bioinformatics is the application of computational techniques and information technology to the organisation and management of biological data. Classical bioinformatics deals primarily with sequence analysis.

3.1. Aim of Bioinformatics:

- Development of database containing all biological information.
- Development of better tools for data designing, annotation and mining.
- Design and development of drugs by using simulation software.
- Design and development of software tools for protein structure prediction function, annotation and docking analysis.
- Creation and development of software to improve tools for analysing sequences for their function and similarity with other sequences.

3.2. Biological Database:

- Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology and computational analysis.
- They contain information from genomics, proteomics, and microarray gene expression.
- Information contained in biological databases includes gene Function, structure, localization (both cellular and Chromosomal), biological sequences and structures.
- Biological data are complex, exception-ridden, vast and incomplete. Therefore several databases has been created and interpreted to ensure unambiguous results. A collection of biological data arranged in computer readable form that enhances the speed of searchand retrieval and convenient to use is called biological database. A good database must have updated information.

3.3. Importance of Biological Database:

- A range of information like biological sequences, structures, binding sites, metabolic interactions, molecular action, functional relationships, protein families, motifs and homologous can be retrieved by using biological databases. The main purpose of a biological database is to store and manage biological data and information in computer readable forms.
- Databases act as a store house of information.
- Databases are used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.
- It allows knowledge discovery, which refers to the identification of connections between pieces of information that were not known when the information was first entered. This facilitates the discovery of new biological insights from raw data.
- Secondary databases have become the molecular biologist's reference library over the past decade or so, providing a wealth of information on just about any gene or gene product that has been investigated by the research community.
- It helps to solve cases where many users want to access the same entries of data.
- Allows the indexing of data.
- It helps to remove redundancy of data.
- It helps the researchers to study the available data and form a new thesis, anti-virus, helpful bacteria, medicines, etc.
- It helps scientists to understand the concepts of biological phenomena.

3.3.1. Types of Biological Database:

There are basically 2 types of biological databases are as follows.

3.3.1.1. Primary databases:

It can also be called an archival database since it archives the experimental results submitted by the scientists. The primary database is populated with experimentally derived data like genome sequence, macromolecular structure, etc. The data entered here remains uncurated (no modifications are performed over the data). Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature. It obtains unique data obtained from the laboratory and these data are made accessible to normal users without any change. The data are given accession numbers when they are entered into the database. The same data can later be retrieved using the accession number. Accession number identifies each data uniquely andit never changes. They form part of the scientific record.

Examples

NCBI (The National Centre for Biotechnology Information)
GenBank
DDBJ (DNA data bank of Japan)
SWISS-PROT(Swiss-Prot)
PIR (Protein Information Resource)
PDB(Protein Data Bank)

3.3.1.2. Secondary Database:

The data stored in these types of databases are the analyzed result of the primary database. Computational algorithms are applied to the primary database and meaningful and informative data is stored inside the secondary database. The data here are highly curated (processing the data before it is presented in the database), often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science. A secondary database is better and contains more valuable knowledge compared to the primary database.

Examples:

- TrEMBL
- Pfam
- PROSITE
- Profiles
- SCOP
- CATH

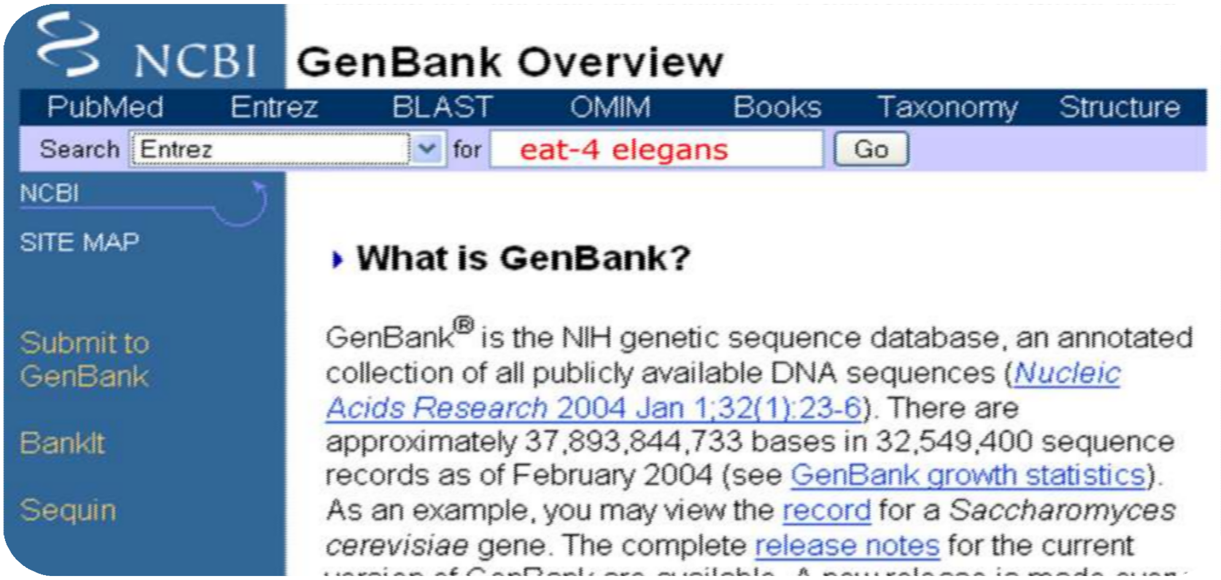
Examples of Database:

3.4. GenBank

GenBank (R) is a public database of all known nucleotide and protein sequences with supporting bibliographic and biological annotation, built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH). As of Release 119 in August 2000, GenBank contained more than 9.5 billion nucleotide bases from over 8.2 million DNA sequences representing more than 70,000 different organisms.

The GenBank sequences are derived primarily through direct submission of sequence data from individual laboratories and large-scale sequencing projects. Most submissions are made using the BankIt (Web) or Sequin programs and accession numbers are assigned by GenBank staff upon receipt. Data exchange with the EMBL Data Library and the DNA Data Bank of Japan helps ensure comprehensive worldwide coverage. GenBank data is accessible through NCBI's integrated retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, and protein structure information plus the biomedical literature via PubMed.

Sequence similarity searching is provided by the BLAST family of programs. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. NCBI also offers a wide range of World Wide Web retrieval and analysis services based on GenBank data.



What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research 2004 Jan 1;32(1):23-6*). There are approximately 37,893,844,733 bases in 32,549,400 sequence records as of February 2004 (see GenBank growth statistics). As an example, you may view the record for a *Saccharomyces cerevisiae* gene. The complete release notes for the current

Figure 3.1 GenBank - Representation

3.4.1. GenBank Data Usage:

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

3.4.1.1. Confidentiality

Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work. GenBank will, upon request, withhold release of new submissions for a specified period of time. However, if the accession number or sequence data appears in print or online prior to the specified date, your sequence will be released. In order to prevent the delay in the appearance of published sequence data, we urge authors to inform us of the appearance of the published data.

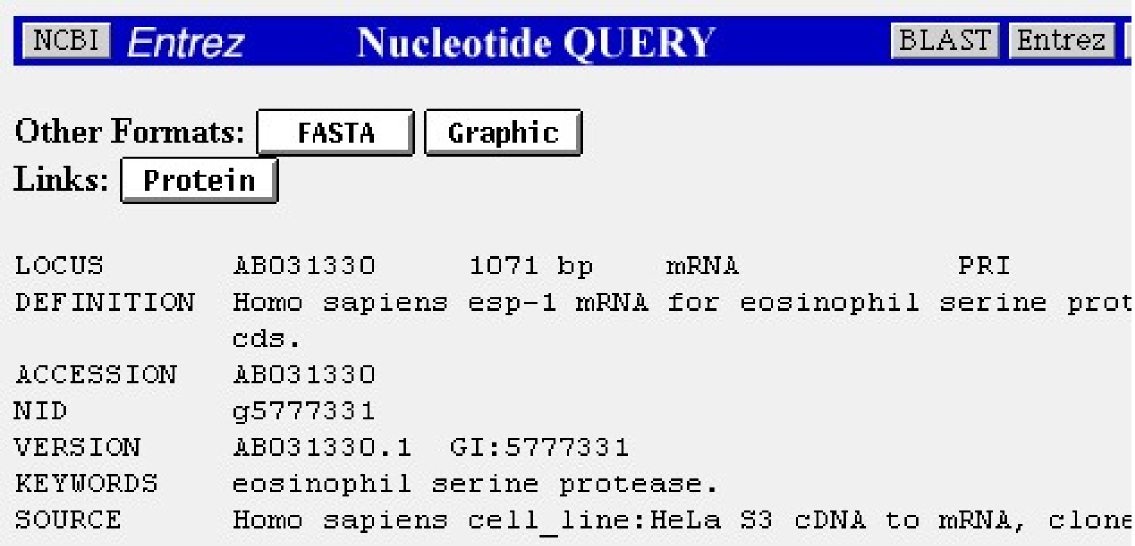
3.4.1.2. Privacy

If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. GenBank assumes that the submitter has received any necessary informed consent authorizations required prior to submitting sequences.

3.4.1.3. NCBI

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology. NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and many other kinds of biological data. In most cases, the data underlying these resources and executables for the software described are available for download at ftp.ncbi.nlm.nih.gov.

As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.



```
LOCUS       AB031330     1071 bp    mRNA            PRI
DEFINITION  Homo sapiens esp-1 mRNA for eosinophil serine prot
            cds.
ACCESSION   AB031330
NID         g5777331
VERSION     AB031330.1  GI:5777331
KEYWORDS    eosinophil serine protease.
SOURCE      Homo sapiens cell_line:HeLa S3 cDNA to mRNA, clone
```

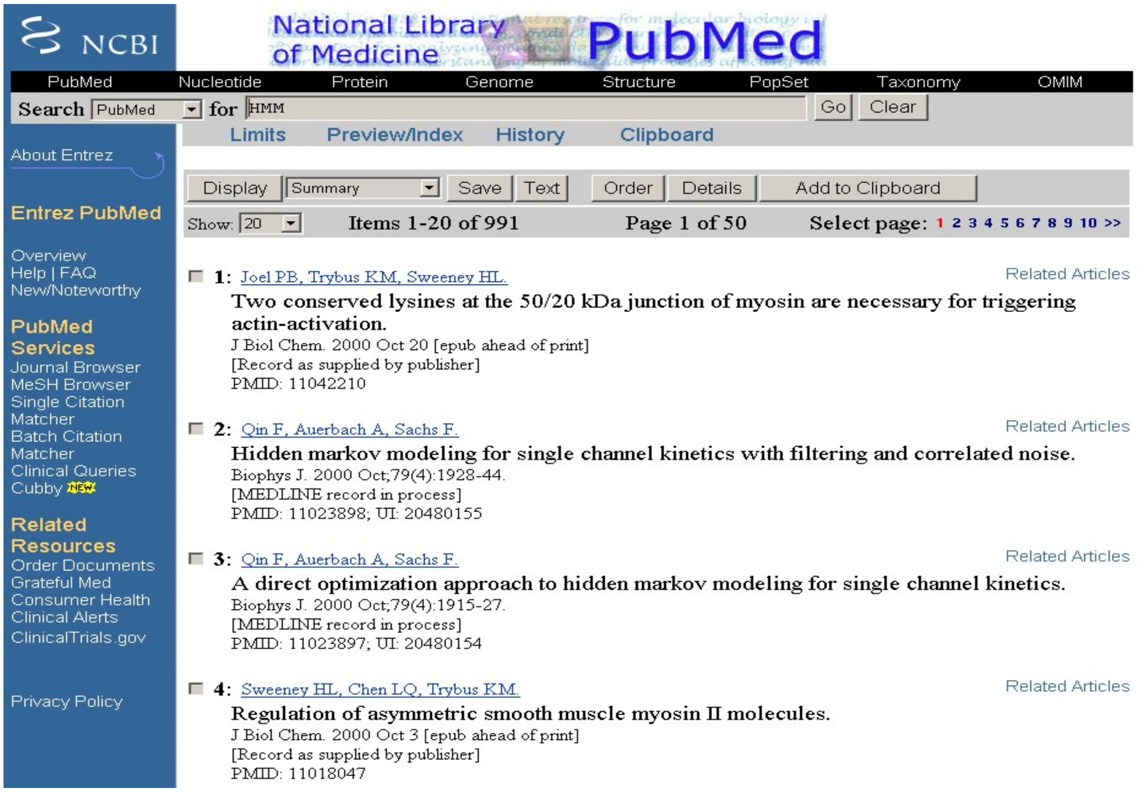
Figure 3.2. GenBank File Format

To carry out its diverse responsibilities, NCBI:

- conducts research on fundamental biomedical problems at the molecular level using mathematical and computational methods
- maintains collaborations with several NIH institutes, academia, industry, and other governmental agencies
- fosters scientific communication by sponsoring meetings, workshops, and lecture series
- supports training on basic and applied research in computational biology for postdoctoral fellows through the NIH Intramural Research Program
- engages members of the international scientific community in informatics research and training through the Scientific Visitors Program
- develops, distributes, supports, and coordinates access to a variety of databases and software for the scientific and medical communities
- develops and promotes standards for databases, data deposition and exchange, and biological nomenclature

3.4.1.4. Popular NCBI Databases:

- BLAST (Basic Local Alignment Search Tool) compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
- Entrez Gene is a searchable database of genes, from RefSeq genomes, and defined by sequence and/or located in the NCBI Map Viewer.
- Nucleotide is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.
- Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.
- PubMed is a bibliographic database of more than 19 million citations for biomedical literature from MEDLINE, life science journals, and online books.
- The power of NCBI's resources is found in their relationship to one another, as most are linked together, providing a comprehensive toolkit for researchers in biomedicine. Online tutorials and help are available at each site, and a nice collection of tutorials can be found on NCBI's YouTube channel.



□ 1: Joel PB, Trybus KM, Sweeney HL.

Two conserved lysines at the 50/20 kDa junction of myosin are necessary for triggering actin-activation.
J Biol Chem. 2000 Oct 20 [epub ahead of print]
[Record as supplied by publisher]
PMID: 11042210

□ 2: Qin F, Auerbach A, Sachs F.

Hidden markov modeling for single channel kinetics with filtering and correlated noise.
Biophys J. 2000 Oct;79(4):1928-44.
[MEDLINE record in process]
PMID: 11023898; UI: 20480155

□ 3: Qin F, Auerbach A, Sachs F.

A direct optimization approach to hidden markov modeling for single channel kinetics.
Biophys J. 2000 Oct;79(4):1915-27.
[MEDLINE record in process]
PMID: 11023897; UI: 20480154

□ 4: Sweeney HL, Chen LQ, Trybus KM.

Regulation of asymmetric smooth muscle myosin II molecules.
J Biol Chem. 2000 Oct 3 [epub ahead of print]
[Record as supplied by publisher]
PMID: 11018047

Figure 3.3 Sample NCBI

3.5. EMBL:

The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/index.html) is a comprehensive collection of primary nucleotide sequences maintained at the European Bioinformatics Institute (EBI). Data are received from genome sequencing centres, individual scientists and patent offices. New data are released daily into the EMBLNEW database and are immediately available. The EMBL and EMBLNEW databases are stored and maintained in an ORACLE data management system and can be searched on the Internet with the Sequence Retrieval System (SRS) (1), the EBI search engine for molecular biology databanks. It was established in 1980, the database was historically tightly coupled to the publication of sequences in the scientific literature, but quickly electronic submissions became usual practice. Today, the volume of data submitted by direct transfer of data from major sequencing centres, such as the Sanger Centre, overshadows all other input. Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages and also provides a structure usable by human readers. The EMBL flat-file comprises of a series of strictly controlled line types (for details see User Manual) that are presented in a tabular manner and consists of four major blocks of data.

- Descriptions and identifiers. Entry name, confidential status, molecule type, taxonomic division, and total sequence length (found in the ID line); accession number (AC); sequence version (SV); date of creation and last update (DT); brief description of the sequence (DE); keywords (KW); taxonomic classification (OS, OC) and links to related database entries (DR).
- Citations. The citation details (RX, RA, RT and RL) of the associated publication and the name (RA) and contact details (RL) of the original submitter.
- Features. Detailed source information, biological features, feature locations and feature qualifiers (multiple FT lines).
- Sequence. Total sequence length, base composition (SQ) and sequence.

3.6. SWISS PROT

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, structure of its domains, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. SWISS-PROT is now an equal partnership between the EMBL and the Swiss Institute of Bioinformatics (SIB). The SWISS-PROT protein sequence database consists of sequence entries. Sequence entries are composed of different line types, each with their own format. The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with other databases.

3.6.1. Annotation

In SWISS-PROT two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein), while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example, carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.

- Domains and sites. For example, calcium binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains, etc.
- Secondary structure. For example, alpha helix, beta sheet, etc.
- Quaternary structure. For example, homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

3.6.2. Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. SWISS-PROT tries to merge all these data so as to minimise the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding SWISS-PROT entry.

3.6.3. Integration with other databases

It is important to provide the users of bimolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialised data collections. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT.

UNIT IV

4.1. Definition – Tree:

A tree has many analogies in real life, and turns out that it has influenced a wide area of **machine learning**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning.

4.1.1. Types of Classification:

- Decision Tree
- Random Forest
- Naïve Bayes
- K Nearest Neighbor
- Logistic Regression

- A classification tree will determine a set of logical if then conditions to classify problems. For example, discriminating between three types of flowers based on certain features.
- A decision tree is a graphical representation of all possible solutions to a decision based on certain conditions.
- It is a tree shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, occurrence or relation.

4.1.2. Advantages of Decision Tree:

- Simple to understand, interpret and visualize
- Little effort for data preparation and less requirement of data cleaning
- Can handle both numerical and categorical data
- Useful for solving decision related problems.
- Nonlinear parameters don’t affect its performance

4.1.3. Disadvantages:

- Over fitting occurs when the algorithm captures noise in the data
- High variance- The model can get unstable due to small variation in data
- Low biased tree- A highly complication DT tends to have a low bias which makes it difficult for the model to work with new data

4.1.4. Motivation for tree-based models:

- Handling of categorical variables
- Handling of missing values and unknown levels
- Detection of nonlinear relationships
- Visualization and interpretation in decision trees.

4.1.5. Decision Tree- Terminology:

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

4.1.6. Building a Decision tree:

There are several algorithms to build a decision tree.

- CART-Classification And Regression Trees –Gini Index
- ID3-Iterative Dichotomiser 3
  - Entrophy function
  - Information Gain
- C4.5
- CHAID-Chi-squared Automatic Interaction Detection

A Decision tree is tree which each node represents a Feature(Attribute) , each link (branch) represents a Decision (Rule) and each leaf represents an outcome.

How does the Decision Tree algorithm Work?

- In a decision tree, for predicting the class of the given dataset, the algorithm starts from the **root node of the tree.** This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
- For the next node, the algorithm again compares the **attribute value** with the other **sub-nodes** and move further. It continues the process until it reaches the **leaf node** of the tree.

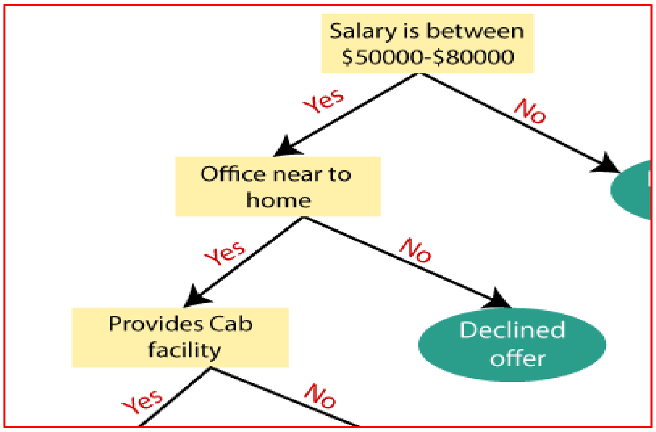


Figure 4.1 Decision Tree

The complete process can be better understood using the below algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step - 3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

4.1.7. Attribute Selection Measures:

- While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.**
- By this measurement, it can easily select the best attribute for the nodes of the tree.

These are popular techniques for ASM:

- Information Gain-ID3
- Gini Index-CART
- Entropy-ID3
- Gain Ratio-C4.5
- Reduction in Variance-C4.5
- Chi-Square-CHAID

Information Gain:

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, split the node and build the decision tree.

ie., decide which attribute should be selected as the decision node

- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)-[{Weighted avg)* Entropy(Each feature)]

For example

IG(T,X)=Entropy(T)-Entropy(T,X)

IG(PlayGolf,Outlook)=E(PlayGolf)-E(PlayGolf,Outlook)

= 0.940-0.693=0.247

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data.

Entropy can be calculated as:

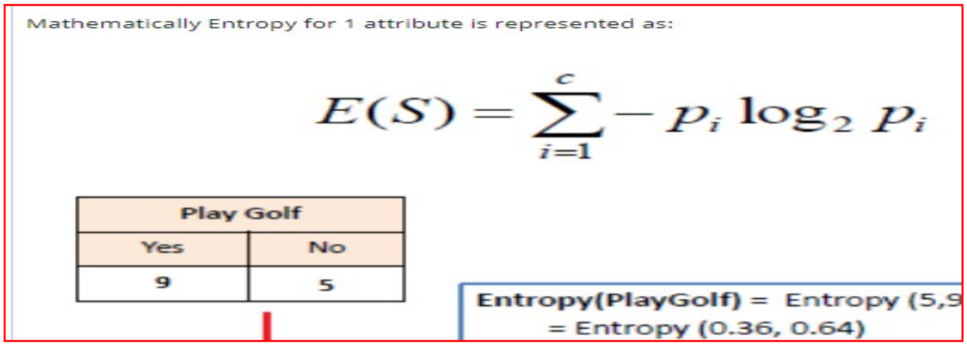
Entropy(s)= P(yes)$\log_2$ P(yes)- P(no) $\log_2$ P(no)

Where,

S= Total number of samples

P(yes)= probability of yes

P(no)= probability of no



Mathematically Entropy for 1 attribute is represented as:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9

= Entropy (0.36, 0.64)

Figure 4.2. Entropy

Gini Index:

Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.

An attribute with the low Gini index should be preferred as compared to the high Gini index.

It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.

Gini index can be calculated using the below formula:

$$\text{Gini Index}= 1- \sum_j P_j^2$$

Steps to Calculate Gini index for a split

- Calculate Gini for sub-nodes, using the above formula for success(p) and failure(q) (p^2+q^2).
- Calculate the Gini index for split using the weighted Gini score of each node of that split.

CART (Classification and Regression Tree) uses the Gini index method to create split points.

Pruning: Getting an Optimal Decision tree

- *Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.*
- A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning.

There are mainly two types of trees **pruning** technology used:

- Cost Complexity Pruning
- Reduced Error Pruning.

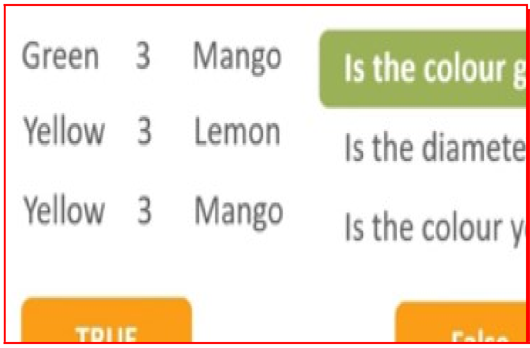


Figure 4.3(a) Tree Pruning

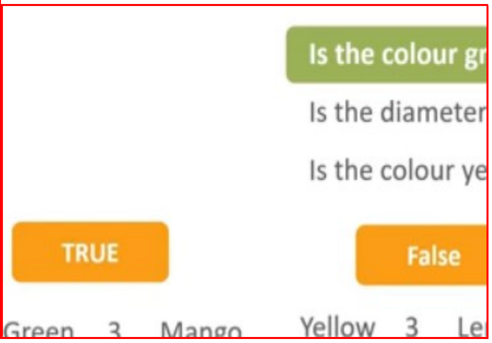


Figure 4.3(b) Tree Pruning

4.1.8. What is a CART in Machine Learning?

- A Classification and Regression Tree(CART) is a predictive algorithm used in machine learning. It explains how a target variable's values can be predicted based on other values.
- It is a decision tree where each fork is a split in a predictor variable and each node at the end has a prediction for the target variable.
- The CART algorithm is an important decision tree algorithm that lies at the foundation of machine learning.
- Moreover, it is also the basis for other powerful machine learning algorithms like bagged decision trees, random forest and boosted decision trees.
- The Classification and regression tree (CART) methodology is one of the oldest and most fundamental algorithms. It is used to predict outcomes based on certain predictor variables.

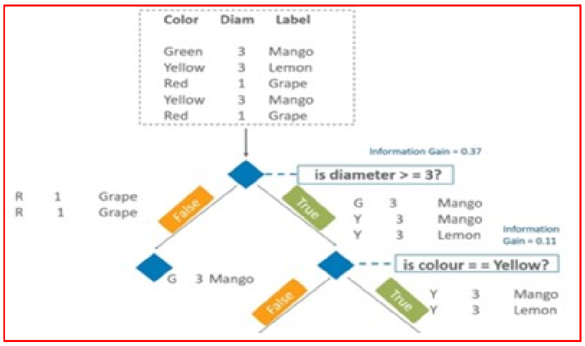
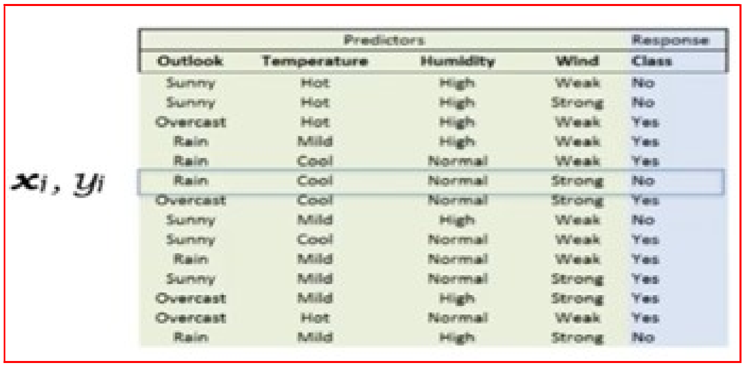


Figure 4.4 CART

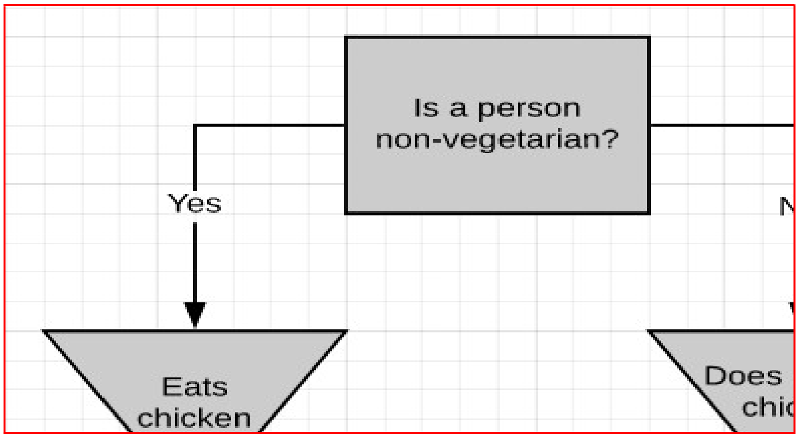


x_i , y_i

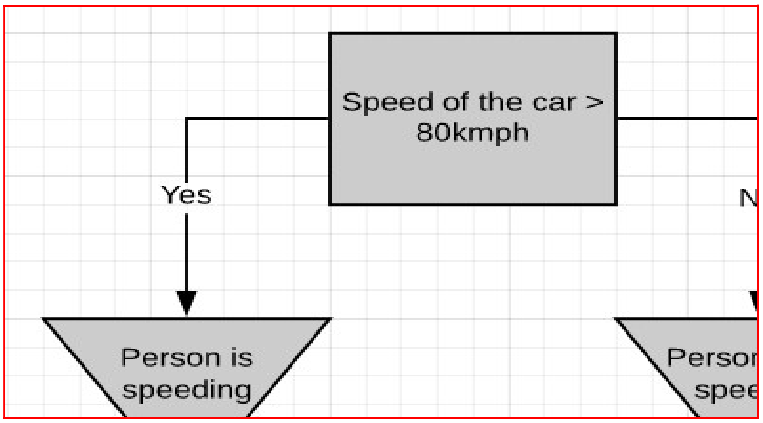
Predictors				Response
Outlook	Temperature	Humidity	Wind	Class
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Fgiure 4.2 Classification - CART

4.2. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.



Example 2: Decision tree is based on numeric data. If a person is driving above 80kmph, we can consider it as over-speeding, else not.



Example 3: If a person is driving above 80kmph, we can consider it as over-speeding, else not.Here is one more simple decision tree. This decision tree is based on ranked data, where 1 means the speed is too high, 2 corresponds to a much lesser speed. If a person is speeding above rank 1 then he/she is highly over-speeding. If the person is above speed rank 2 but below speed rank 1 then he/she is over-speeding but not that much. If the person is below speed rank 2 then he/she is driving well within speed limits.

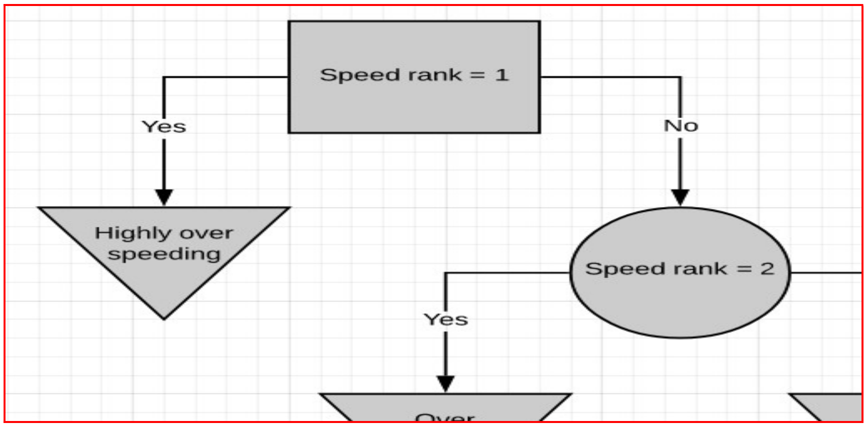


Figure 4.3 Classification – Decision Tree

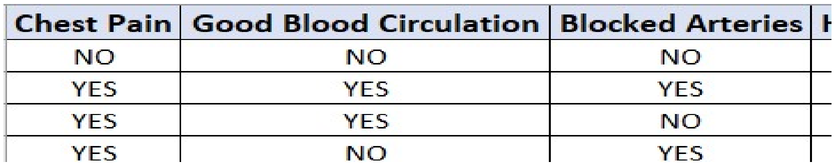
The tree is called the ***root node***. The nodes in between are called ***internal nodes***. Internal nodes have arrows pointing to them and arrows pointing away from them. The end nodes are called the ***leaf nodes*** or just ***leaves***. Leaf nodes have arrows pointing to them but no arrows pointing away from them.

In the above diagrams, root nodes are represented by rectangles, internal nodes by circles, and leaf nodes by inverted-triangles.

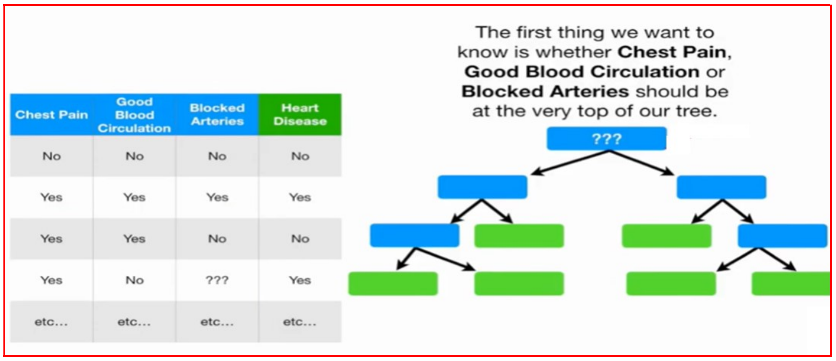
4.2.1. Classification and Regression Trees:

CART is a DT algorithm that produces binary Classification or Regression Trees, depending on whether the dependent (or target) variable is categorical or numeric, respectively. It handles data in its raw form (no pre-processing needed) and can use the same variables more than once in different parts of the same DT, which may uncover complex interdependencies between sets of variables.

In the example given, build a decision tree that uses chest pain, good blood circulation, and the status of blocked arteries to predict if a person has heart disease or not.



Chest Pain	Good Blood Circulation	Blocked Arteries	H
NO	NO	NO	
YES	YES	YES	
YES	YES	NO	
YES	NO	YES	



Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

The first thing we want to know is whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be at the very top of our tree.

Figure 4.4 Classification & Regression Tree

There are two leaf nodes, one each for the two outcomes of chest pain. Each of the leaves contains the no. of patients having heart disease and not having heart disease for the corresponding entry of chest pain

Chest pain as the root node

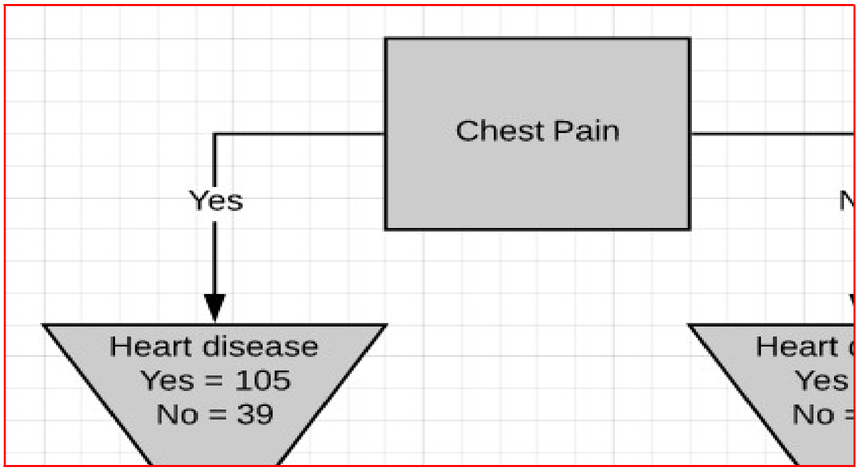


Figure 4.5 Root Node Representation

There are two leaf nodes, one each for the two outcomes of chest pain. Each of the leaves contains the no. of patients having heart disease and not having heart disease for the corresponding entry of chest pain.

The same thing for good blood circulation and blocked arteries.

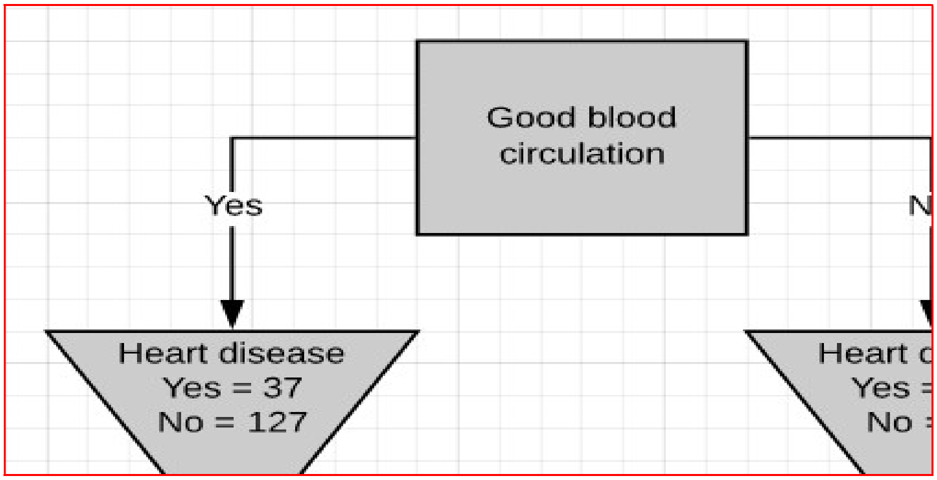


Figure 4.6 (a) blood circulation as the root node

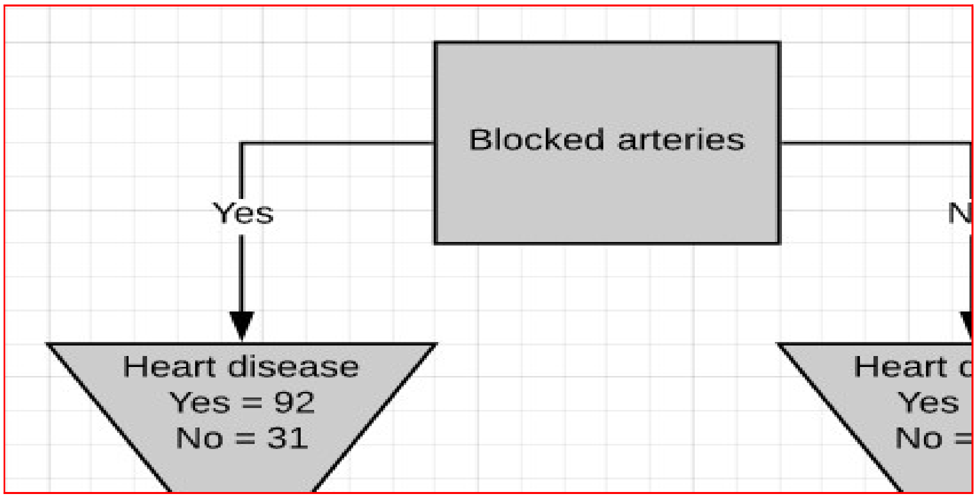


Figure 4.6 (b) Blocked arteries as the root node

The 3 features separate the patients having heart disease from the patients not having heart disease perfectly. It is to be noted that the total no. of patients having heart disease is different in all three cases. This is done to simulate the missing values present in real-world datasets. Because none of the leaf nodes is either 100% 'yes heart disease' or 100% 'no heart disease', they are all considered *impure.* To decide on which separation is the best, we need a method to measure and compare *impurity.*

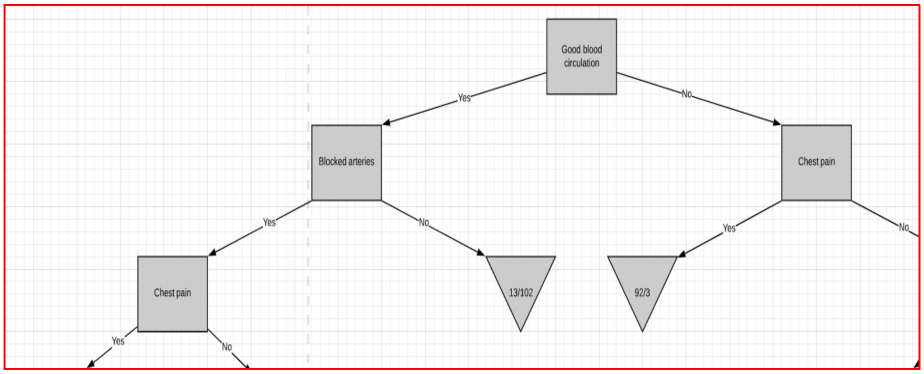
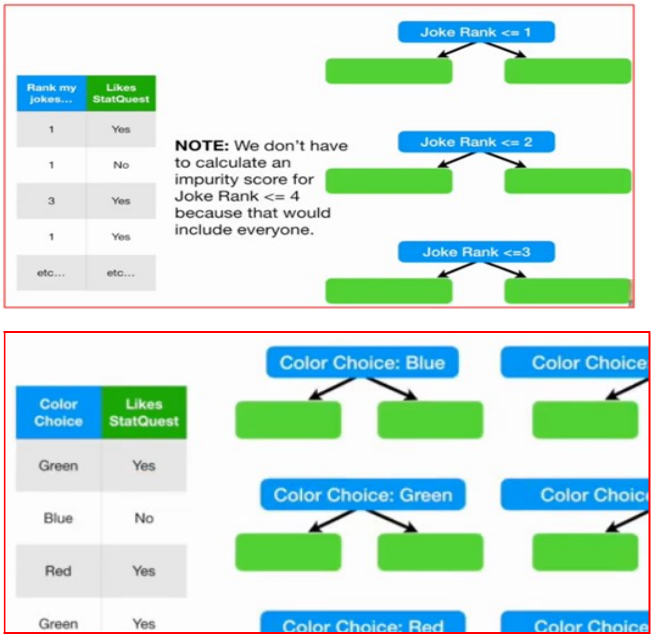


Figure 4. Complete Decision tree

Some examples



Rank my jokes...	Likes StatQuest
1	Yes
1	No
3	Yes
1	Yes
etc...	etc...

NOTE: We don't have to calculate an impurity score for Joke Rank <= 4 because that would include everyone.

Color Choice	Likes StatQuest
Green	Yes
Blue	No
Red	Yes
Green	Yes

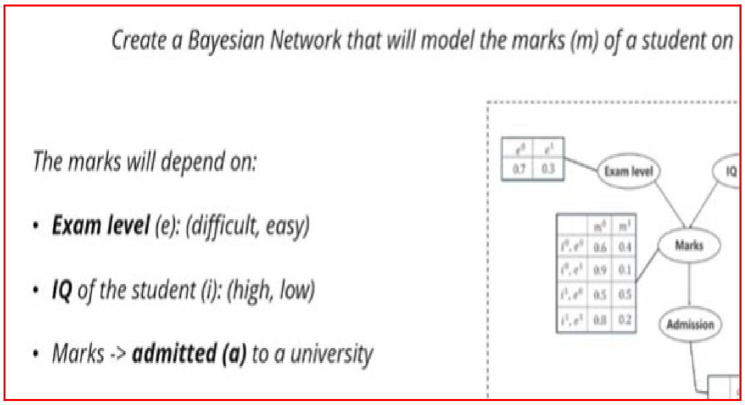
Figure 4.7 Representation of Decision Tree

4.3 Bayesian networks:

A Bayesian network is a probabilistic graphical model (PGM) that is used to compute uncertainties by using the concept of probability ie., that represents a set of variables and their conditional dependencies via a directed acyclic graph.

It consists of two components:

- A network structure in the form of a directed acyclic graph (DAG). In this graph, nodes represent the random variables and directed edges represent stochastic dependencies among variables.
- A set of conditional probability distributions, one for each variable, characterizing the stochastic dependencies represented by the edges. These conditional distributions are specified by the network parameters
- A DAG is used to represent a Bayesian network and like any other statistical graph, a DAG contains a set of nodes and links, where the links denote the relationship between the nodes



Create a Bayesian Network that will model the marks (m) of a student on

The marks will depend on:

- ***Exam level** (e): (difficult, easy)*
- ***IQ** of the student (i): (high, low)*
- *Marks -> **admitted (a)** to a university*

e^0	e^1
0.7	0.3

	m^0	m^1
i^0, e^0	0.6	0.4
i^0, e^1	0.9	0.1
i^1, e^0	0.5	0.5
i^1, e^1	0.8	0.2

Figure 4.8 Example – Bayesian Network

4.3.1. Bayesian Belief Networks:

- Learning Probabilities (local distributions)
    - Introduction to Bayesian statistics: Learning a probability
    - Learning probabilities in a Bayes net
    - Applications
- Learning Bayes-net structure
    - Bayesian model selection/averaging
    - Applications
- Incomplete data makes parameters dependent
- Parameter Learning for incomplete data
- Monte-Carlo integration
    - Investigate properties of the posterior and perform prediction
- Large-sample Approx. (Laplace/Gaussian approx.)
    - Expectation-maximization (EM) algorithm and inference to compute mean and variance.

4.3.2. Variational methods

- Learning Correct Model:
- True graph G and P is the generative distribution
- Markov Assumption: P satisfies the independencies implied by G
- Faithfulness Assumption: P satisfies only the independencies implied by G
- Theorem: Under Markov and Faithfulness, with enough data generated from P one can recover G (up to equivalence). Even with the greedy method!

4.4. Random forest:

Random forest or Random decision forest is a method that operates by constructing multiple decision trees during training phase. The decision of the majority of the tree is chosen by the random forest as the final decision.

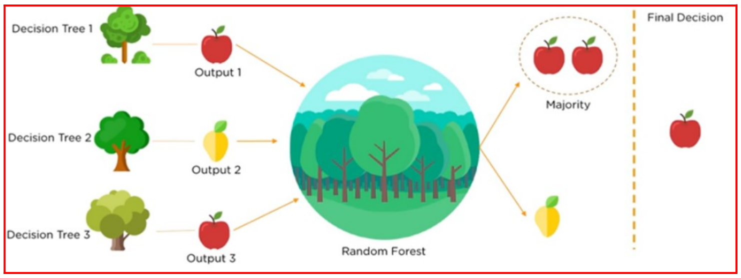


Figure 4.9 Random Forest - Representation

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.* Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4.4.1. Assumptions for Random Forest:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

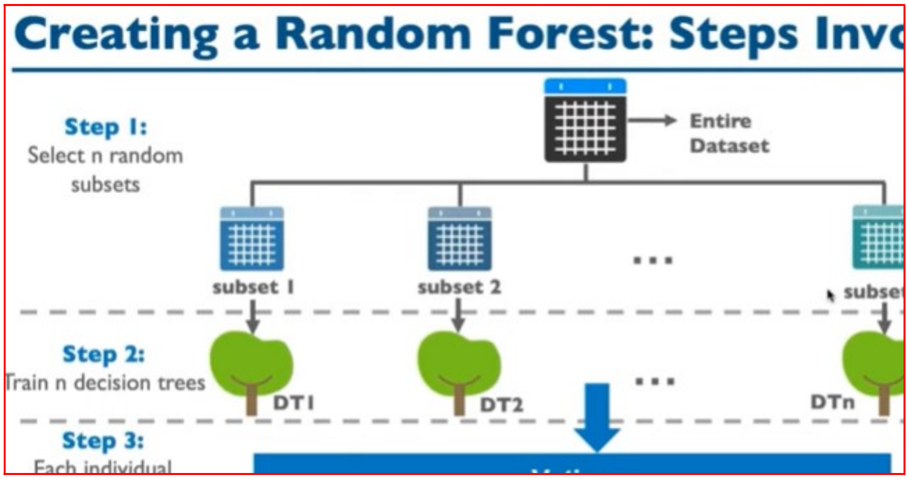
- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Why Random Forest:

- Random forest is the most used supervised machine learning algorithm for classification and regression
- RF uses ensemble learning method in which the predictions are based on the combined results of various individual models
- No Overfitting
    - Use of multiple trees reduce the risk of overfitting

- Training time is less

- High Accuracy
  - Runs efficiently on large database
  - For large data. It produces highly accurate predictions
- Estimates missing data
  - Random forest can maintain accuracy when a large proportion of data is missing



Creating a Random Forest: Steps Inv

Step 1: Select n random subsets

Step 2: Train n decision trees

Step 3: Each individual

Figure 4.10 Steps in Random Forest

How does Random Forest algorithm work?

- Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

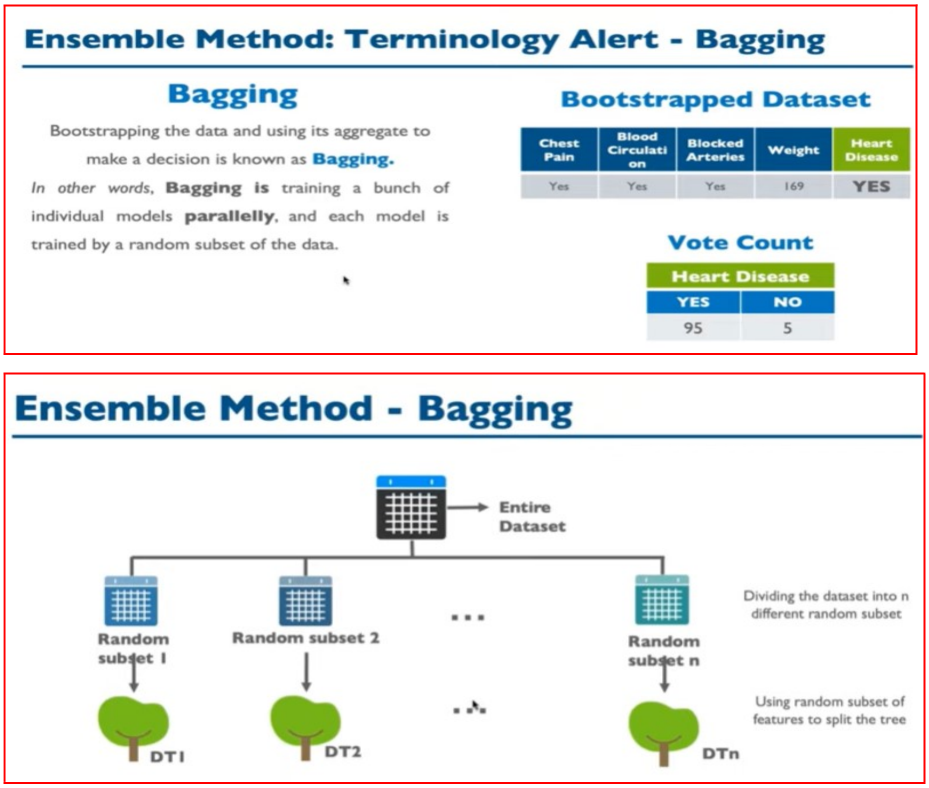
Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.



Ensemble Method: Terminology Alert - Bagging

Bagging

Bootstrapping the data and using its aggregate to make a decision is known as **Bagging.**

In other words, **Bagging is** training a bunch of individual models **parallelly**, and each model is trained by a random subset of the data.

Bootstrapped Dataset

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	169	YES

Vote Count

Heart Disease	
YES	NO
95	5

Ensemble Method - Bagging

Entire Dataset

Random subset 1

Random subset 2

...

Random subset n

Dividing the dataset into n different random subset

Using random subset of features to split the tree

Figure 4.11 Bagging Method

How Boosting Algorithm Works?

- The basic principle behind the working of the boosting algorithm is to generate multiple weak learners and combine their predictions to form one strong rule.
- These weak rules are generated by applying base Machine Learning algorithms on different distributions of the data set. These algorithms generate weak rules for each iteration.
- After multiple iterations, the weak learners are combined to form a strong learner that will predict a more accurate outcome.

The algorithm:

Step 1: The base algorithm reads the data and assigns equal weight to each sample observation.

Step 2: False predictions made by the base learner are identified. In the next iteration, these false predictions are assigned to the next base learner with a higher weightage on these incorrect predictions.

Step 3: Repeat step 2 until the algorithm can correctly classify the output.

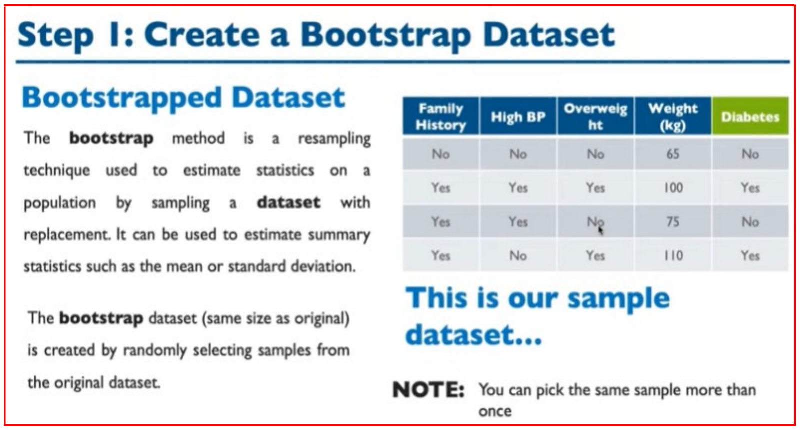
Therefore, the main aim of Boosting is to focus more on miss-classified predictions

Algorithms based on Bagging and Boosting:

- Bagging and Boosting are two of the most commonly used techniques in machine learning. Following are the algorithms will be focusing on:
- Bagging algorithms:
    - Bagging meta-estimator
    - **Random forest**
- Boosting algorithms:
    - **AdaBoost(Adaptive Boosting)**
    - GBM(Gradient Boosting)
    - XGBM
    - Light GBM
    - CatBoost

Similarities Between Bagging and Boosting:

- Both are ensemble methods to get N learners from 1 learner.
- Both generate several training data sets by random sampling.
- Both make the final decision by averaging the N learners (or taking the majority of them i.e Majority Voting).
- Both are good at reducing variance and provide higher stability.



Step 1: Create a Bootstrap Dataset

Bootstrapped Dataset

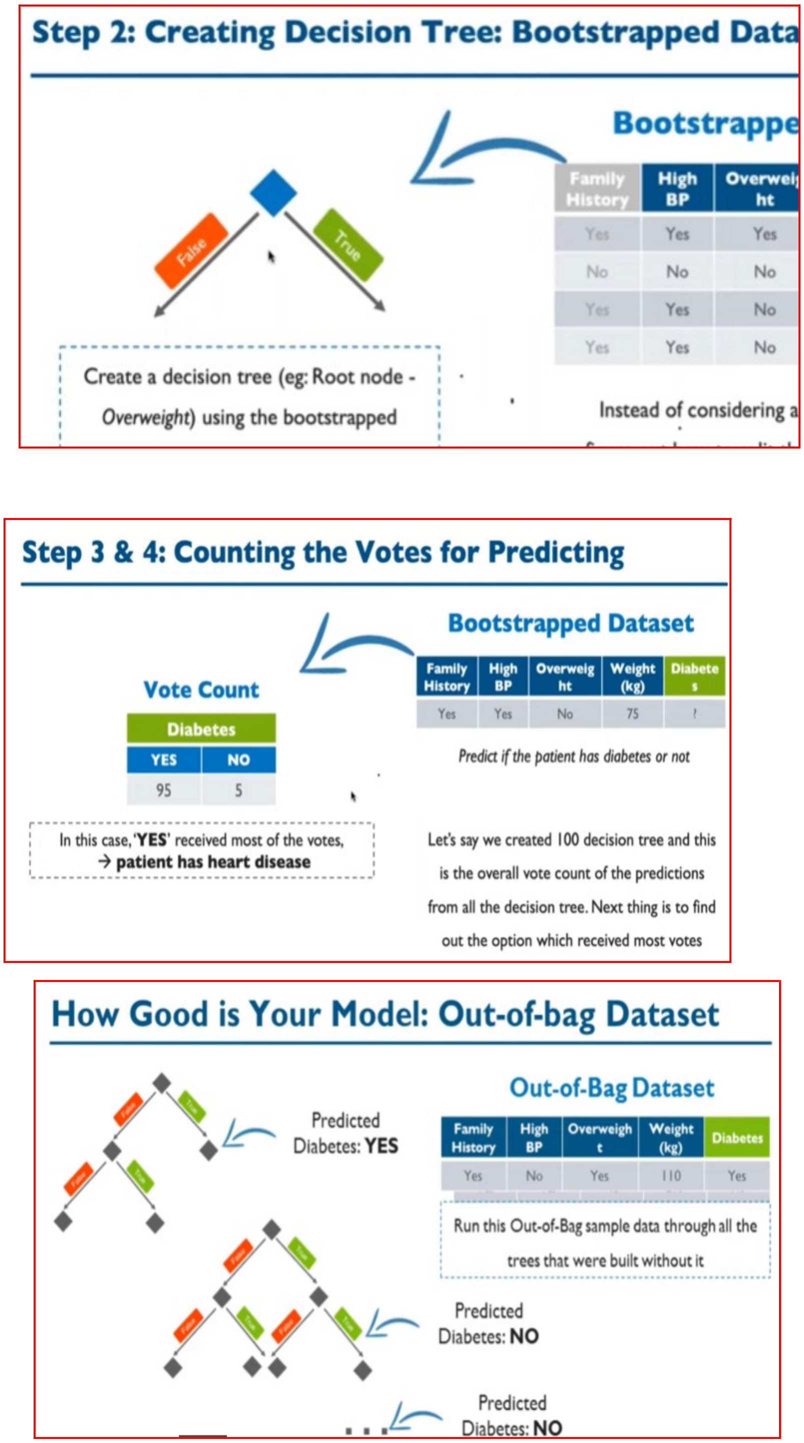
The **bootstrap** method is a resampling technique used to estimate statistics on a population by sampling a **dataset** with replacement. It can be used to estimate summary statistics such as the mean or standard deviation.

The **bootstrap** dataset (same size as original) is created by randomly selecting samples from the original dataset.

Family History	High BP	Overweig ht	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes

This is our sample dataset...

NOTE: You can pick the same sample more than once



Step 2: Creating Decision Tree: Bootstrapped Data

Bootstrappe

Family History	High BP	Overwei ht
Yes	Yes	Yes
No	No	No
Yes	Yes	No
Yes	Yes	No

Create a decision tree (eg: Root node - *Overweight*) using the bootstrapped

Instead of considering a

Step 3 & 4: Counting the Votes for Predicting

Bootstrapped Dataset

Family History	High BP	Overweig ht	Weight (kg)	Diabete s
Yes	Yes	No	75	?

Predict if the patient has diabetes or not

Vote Count

Diabetes	
YES	NO
95	5

In this case, '**YES**' received most of the votes,
→ **patient has heart disease**

Let's say we created 100 decision tree and this is the overall vote count of the predictions from all the decision tree. Next thing is to find out the option which received most votes

How Good is Your Model: Out-of-bag Dataset

Predicted Diabetes: **YES**

Out-of-Bag Dataset

Family History	High BP	Overweigh t	Weight (kg)	Diabetes
Yes	No	Yes	110	Yes

Run this Out-of-Bag sample data through all the trees that were built without it

Predicted Diabetes: **NO**

Predicted Diabetes: **NO**

Figure 4.12 Working Principle - Dataset

Example:

Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random Forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

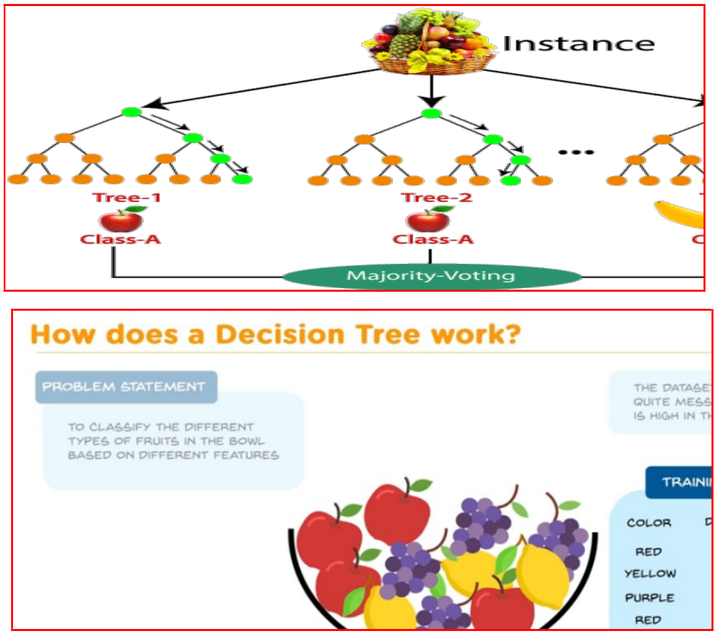
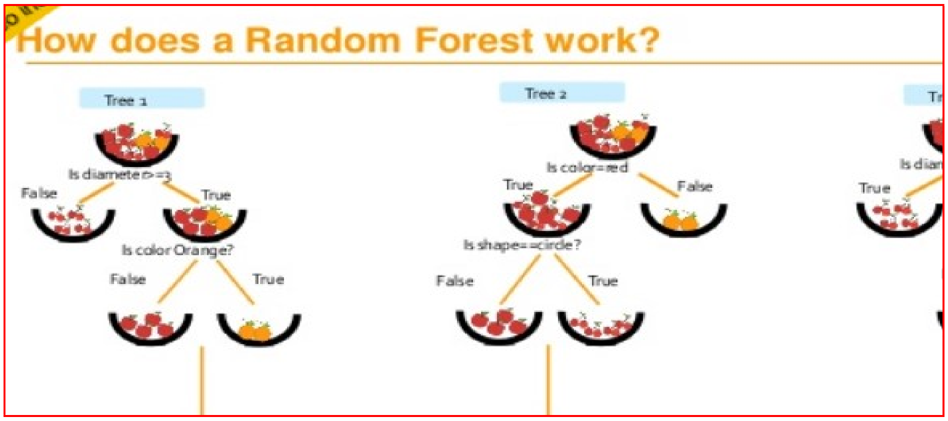


Figure 4.13 Decision Tree - Working Principle



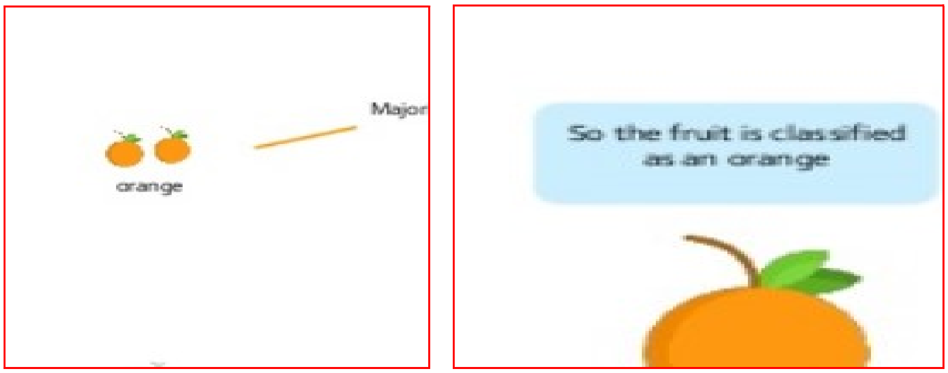


Figure 4.14 Random Forest – Working Principle

4.4.2. Bioinformatics Applications of Random Forest and Variants:

- In the past decade, random forest has been successfully applied to various problems in computational biology.
- The popularity of random forest in this field arises from the fact that RF can be applied to a wide range of data types, even if the problems are nonlinear or involve complex high-order interaction effects.
- RF and its variants have been applied on a variety of bioinformatic problems, such as gene expression classification, mass spectrum protein expression analysis, biomarker discovery, sequence annotation, protein-protein interaction prediction, or statistical genetics.
- RF is especially useful to identify features that show small marginal contributions individually, but gives a larger effect when combined together.
- For example, the initial attempt from utilized random forest permutation importance as a screening procedure to identify small numbers of risk-associated SNPs among large numbers of unassociated SNPs using 16 complex disease models.
- RF was concluded to outperformer Fisher’s exact test when interactions between SNPs exist.
- Later, similar RF importance measure and extended the concept on pairs of predictors, in order to capture joint effects.
- These early studies normally limited the number of SNPs under analysis to a relatively small range.

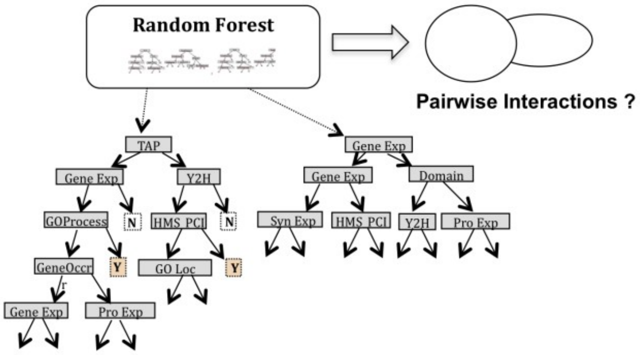


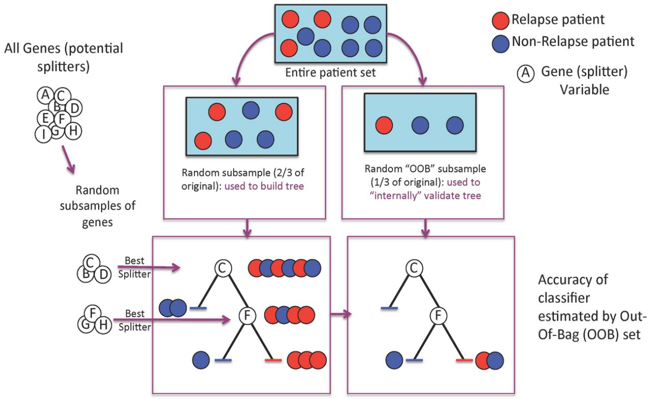
Figure 4.15 shows evidence was integrated using a random forest classifier for protein-protein interaction prediction

RF was used in three specific ways:

- feature selection of drug gene-expression signatures based on RF permutation importance,
- removing outlier cell-lines based on RF proximity, and
- RF multivariate regression model for predicting continuous drug response.
- More applications of random forests can be found in other different fields like quantitative structure-activity relationship modeling, nuclear magnetic resonance spectroscopy, or clinical decision supports in medicine in general

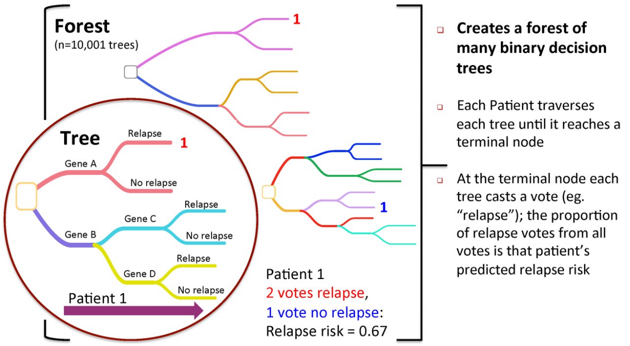
4.4.3. Random Forest algorithm applied in Bio informatics:

A Random Forest is a collection of decision trees. Each tree gets a "vote" in classifying. There are two components of randomness involved in the building of a Random Forest. First, at the creation of each tree, a random subsample of the total data set is selected to grow the tree. Second, at each node of the tree, a well-performing gene from a random subset of all genes is chosen as a "splitter variable". The splitter variable attempts to separate patients in one class (e.g., Response) from those in the other class (e.g., Non-Response). The tree is grown with additional splitter variables until all terminal nodes (leaves) of the tree are purely one class or the other. This tree is then "tested" against the 1/3 of patients set aside, the "out of bag" (OOB) patients. Each OOB patient traverses the tree, going down one branch or another depending on his/her gene expression values for each splitter variable. These OOB patients are assigned a predicted class based on where they land in the tree (a vote). The entire process is repeated with new random divisions into 2/3 and 1/3 patient sets and new random gene sets for selection of splitter variables to produce additional trees and ultimately a forest. In each case a different subset of patients is used to build the tree and test its performance. At the end, each patient will have contributed to the construction of ~2/3 of all trees and been tested in the other ~1/3. Each "test" tree gives a vote for whether the patient will relapse or not relapse. The fraction of votes for relapse is an estimate of the probability of relapse and all patients will be predicted as either a relapse or non-relapse (using probability of 0.5 as the threshold). By comparing these predictions based on the OOB data to their known class, estimates of the accuracy of the overall forest can be obtained. The forest can then also be applied to independent test data or patients of unknown class.



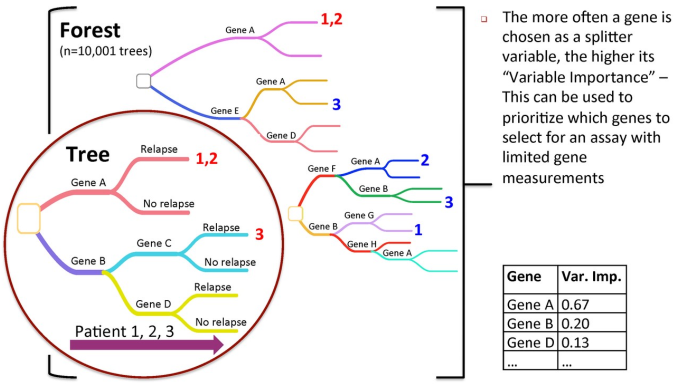
Relapse patient
Non-Relapse patient
A Gene (splitter) Variable
All Genes (potential splitters)
A C B D E F G H I
Random subsamples of genes
Entire patient set
Random subsample (2/3 of original): used to build tree
Random "OOB" subsample (1/3 of original): used to "internally" validate tree
C B D Best Splitter
F G H Best Splitter
C
F
Accuracy of classifier estimated by Out-Of-Bag (OOB) set

Figure 4.16 Creation of the Random Forest tree



Forest
(n=10,001 trees)
1
Tree
Gene A
Relapse
1
No relapse
Gene B
Gene C
Relapse
No relapse
Gene D
Relapse
No relapse
Patient 1
1
Patient 1
2 votes relapse,
1 vote no relapse:
Relapse risk = 0.67
Creates a forest of many binary decision trees
Each Patient traverses each tree until it reaches a terminal node
At the terminal node each tree casts a vote (eg. "relapse"); the proportion of relapse votes from all votes is that patient's predicted relapse risk

Figure 4.17 shows to predict new patients, each tree gets a vote...



- The more often a gene is chosen as a splitter variable, the higher its “Variable Importance” – This can be used to prioritize which genes to select for an assay with limited gene measurements

Gene	Var. Imp.
Gene A	0.67
Gene B	0.20
Gene D	0.13
…	…

The above figure 4.18 shows variable importance is a feature of random forests

4.4.4. Applications of Random Forest:

There are mainly four sectors where Random Forest mostly used:

- **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
- **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
- **Land Use:** We can identify the areas of similar land use by this algorithm.
- **Marketing:** Marketing trends can be identified using this algorithm.

4.4.5. Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

4.4.6.Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

4.5 Evolutionary Learning-Genetic algorithms - Genetic Offspring: - Genetic Operators-Using Genetic Algorithms

- Genetic Algorithm- uses concept from evolutionary Biology(Natural Genetics & Natural selection).

- John Holland introduced in 1975
- Populations of possible solutions to the given problem.
- It is a search-based optimization technique

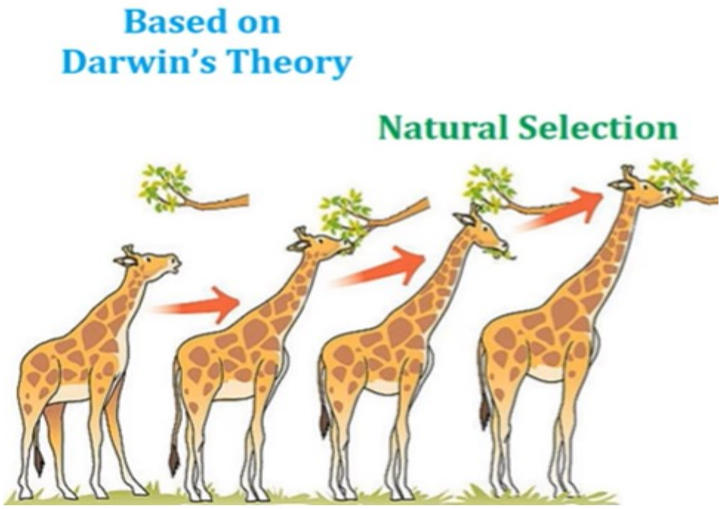


Figure 4.19 Darwin's Representation

Principle of natural selection "**Select the best, discard the rest**"

- Genetic Algorithms are the heuristic search and optimization techniques that mimic the process of natural evolution.
- Thus genetic algorithms implement the optimization strategies by simulating evolution of species through natural selection

Applications:

- DNA analysis
- Robotics
- Game playing
- Business
- Machine learning
- Image processing
- Vehicle Routing
- Neural Network

4.6. Genetic Algorithm:

- Each iteration in the cycle produces a new generation of chromosomes

- The entire set of chromosomes is called a run
- Typical GA run is from 50 to 500 or more generations
- At the end of a run often there is at least one highly fit chromosome in the population

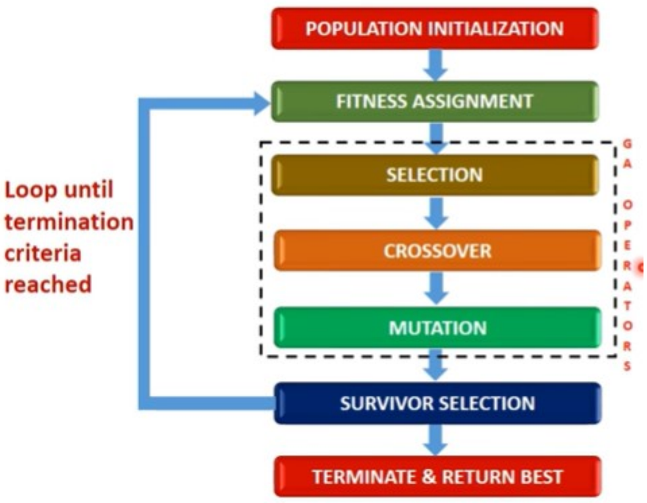


Figure 4.20 Representation of Genetic Algorithm

4.6.1. Basic Terminology of GA:

- Population- subset of all the possible solutions to the given problem
- Chromosomes-one such solution to given problem
- Gene- one element position of a chromosome
- Allele- value a gene takes for particular chromosome
- Genotype- population in the computation space.
- Phenotype-population in the actual real world solution space
- Decoding- transforming a solution from the genotype to the phenotype space
- Encoding- transforming from the phenotype to genotype space

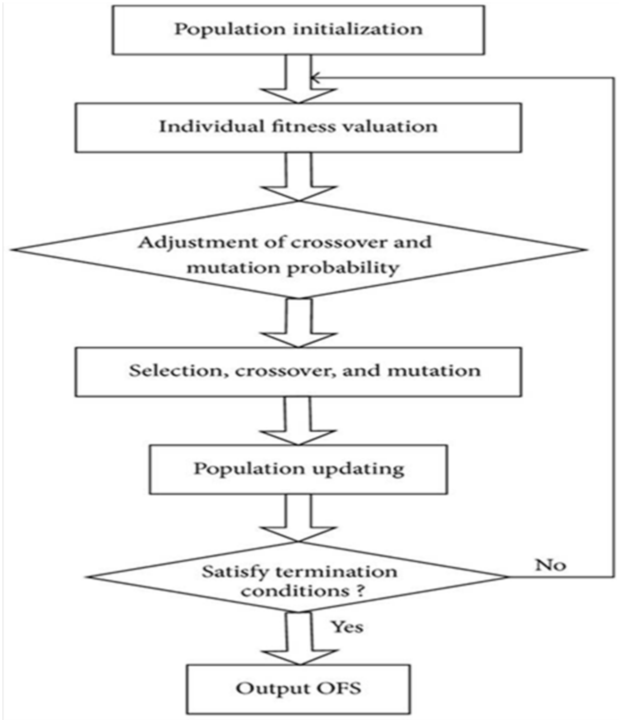

Population initialization
Individual fitness valuation
Adjustment of crossover and mutation probability
Selection, crossover, and mutation
Population updating
Satisfy termination conditions ?
No
Yes
Output OFS

Figure 4.21 Selection, Crossover and Mutation

4.6.1.1. Selection:

- The process that determines which solutions are to be preserved and allowed to reproduce and which ones deserve to die out.
- The primary objective of the selection operator is to emphasize the good solutions and eliminate the bad solutions in a population while keeping the population size constant.
- “Selects the best, discards the rest”

4.6.1.2. Tournament selection:

- In tournament selection several tournaments are played among a few individuals. The individuals are chosen at random from the population.
- The winner of each tournament is selected for next generation.
- Selection pressure can be adjusted by changing the tournament size.
- Weak individuals have a smaller chance to be selected if tournament size is large.

4.6.1.3. Fitness function:

- A fitness value can be assigned to evaluate the solutions

- A fitness function value quantifies the optimality of a solution. The value is used to rank a particular solution against all the other solutions
- A fitness value is assigned to each solution depending on how close it is actually to the optimal solution of the problem

4.6.1.4. Crossover operator:

- The most popular crossover selects any two solutions strings randomly from the mating pool and some portion of the strings is exchanged between the strings.
- The selection point is selected randomly.
- A probability of crossover is also introduced in order to give freedom to an individual solution string to determine whether the solution would go for crossover or not.

4.6.1.5. Binary Crossover:

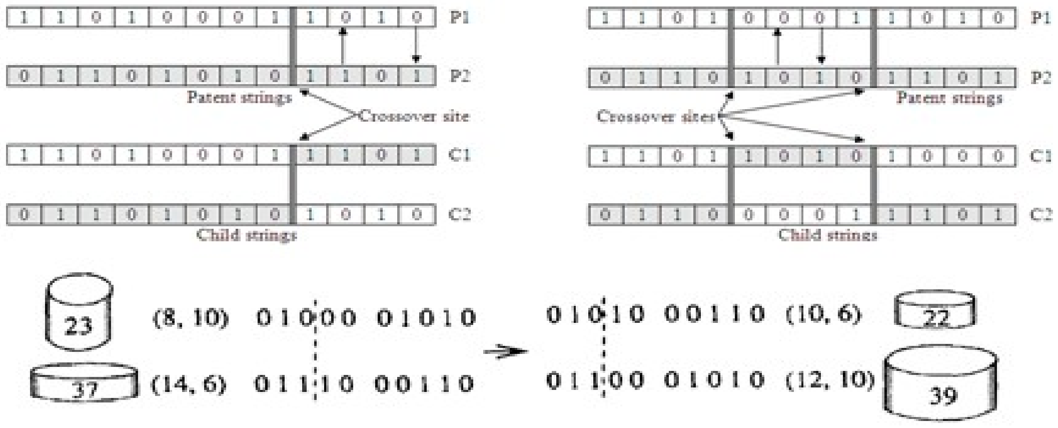

1 1 0 1 0 0 0 1 1 0 1 0 P1
0 1 1 0 1 0 1 0 1 1 0 1 P2
Patent strings
Crossover site
1 1 0 1 0 0 0 1 1 1 0 1 C1
0 1 1 0 1 0 1 0 1 0 1 0 C2
Child strings
1 1 0 1 0 0 0 1 1 0 1 0 P1
0 1 1 0 1 0 1 0 1 1 0 1 P2
Patent strings
Crossover sites
1 1 0 1 1 0 1 0 1 0 0 0 C1
0 1 1 0 0 0 0 1 1 1 0 1 C2
Child strings
23
(8, 10) 01000 01010
37
(14, 6) 01110 00110
01010 00110 (10, 6)
22
01100 01010 (12, 10)
39

Figure 4.22 Binary Crossover

4.6.1.6. Mutation operator:

- Mutation is the occasional introduction of new features in to the solution strings of the population pool to maintain diversity in the population.
- Though crossover has the main responsibility to search for the optimal solution, mutation is also used for this purpose**.**

4.6.1.7. Binary Mutation:

- The mutation probability is generally kept low for steady convergence.
- A high value of mutation probability would search here and there like a random search technique

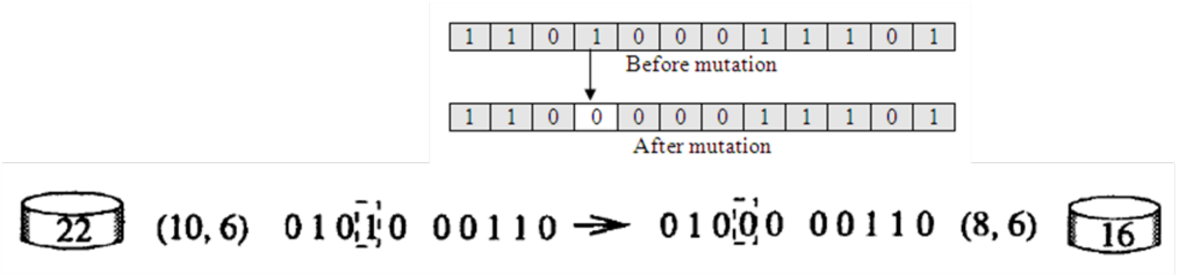


Figure 4.23 Binary Mutation

4.6.2. Advantages of Genetic Algorithm:

- Does not require any derivative information which may not be available for many real-world problems
- Faster and more efficient as compared to the traditional methods
- Optimizes both continuous and discrete functions and also multi objective problems
- Provides a list of ‘good’ solutions and not just a single solution. Always gets an answer which gets better over the time.
- Useful when the search space is very large and there are a number of parameters involved.

UNIT V

5.1. Medical Data:

- Data provide the basis for categorizing the problems a patient may be having or for identifying subgroups within a population of patients. They also help a physician to decide what additional information is needed and what actions should be taken to gain a greater understanding of a patient's problem or to treat most effectively the problem that has been diagnosed.

- We consider a medical datum to be any single observation of a patient.
- If a medical *datum is a single observation about a patient, medical data are multiple* observations. Such data may involve several different observations made concurrently, the observation of the same patient parameter made at several points in time, or both.

Consider the following examples:

An adult patient reports a childhood illness with fevers and a red rash in addition to joint swelling. Could he or she have had scarlet fever? The patient does not know what his or her pediatrician called the disease nor whether anyone thought that he or she had scarlet fever.

- A physician listens to the heart of an asthmatic child and thinks that he or she hears a heart murmur—but is not certain because of the patient's loud wheezing.
- A radiologist looking at a shadow on a chest X-ray film is not sure whether it represents overlapping blood vessels or a lung tumor.
- A confused patient is able to respond to simple questions about his or her illness, but under the circumstances the physician is uncertain how much of the history being reported is reliable.

5.2. Types of Medical Data:

- There is a broad range of data types in the practice of medicine and the allied health sciences. They range from narrative, textual data to numerical measurements, recorded signals, drawings, and even photographs.
- Narrative data account for a large component of the information that is gathered in the care of patients. For example, the patient's description of his or her present illness, including responses to focused questions from the physician, generally is gathered verbally and is recorded as text in the medical record.
- The electronic versions of such reports can also easily be integrated into electronic health records (EHRs) and clinical data repositories so that clinicians can access important clinical information even when the paper record is not available.

- Electronically stored transcriptions of dictated information often include not only patient histories and physical examinations but also other narrative descriptions such as reports of specialty consultations, surgical procedures, pathologic examinations of tissues, and hospitalization summaries when a patient is discharged.
- Some narrative data are loosely coded with shorthand conventions known to health personnel, particularly data collected during the physical examination, in which recorded observations reflect the stereotypic examination process taught to all practitioners.
- The narrative summaries are dictated and then transcribed by typists who work with word processors to produce printed summaries for inclusion in medical records.
- The electronic versions of such reports can also easily be integrated into electronic health records (EHRs) and clinical data repositories so that clinicians can access important clinical information even when the paper record is not available.

- Electronically stored transcriptions of dictated information often include not only patient histories and physical examinations but also other narrative descriptions such as reports of specialty consultations, surgical procedures, pathologic examinations of tissues, and hospitalization summaries when a patient is discharged.

- Many data used in medicine take on discrete numeric values. These include such parameters as laboratory tests, vital signs (such as temperature and pulse rate), and certain measurements taken during the physical examination.
- In some fields of medicine, analog data in the form of continuous signals are particularly Important Perhaps the best-known example is an electrocardiogram (ECG), a tracing of the electrical activity from a patient’s heart.
- Visual images—either acquired from machines or sketched by the physician—are another important category of data. Radiologic images are obvious examples. It also is common for physicians to draw simple pictures to represent abnormalities that they have observed; such drawings may serve as a basis for comparison when they or another physician next see the patient.
- They must record their observations, as well as the actions they have taken and the rationales for those actions, for later communication to themselves and other people.
- A glance at a medical record will quickly reveal the wide variety of data-recording techniques that have evolved. The range goes from hand-written text to commonly understood shorthand notation to cryptic symbols that only specialists can understand; few physicians know how to interpret the data-recording conventions of an

ophthalmologist.

A physician’s hand-drawn sketch of a prostate nodule. A drawing may convey precise information more easily and compactly than a textual description.An ophthalmologist’s report of an eye examination. Most physicians trained in other specialties would have difficulty deciphering the symbols that the ophthalmologist has used.

5.2.1. Uses of Medical Data:

- Medical data are recorded for a variety of purposes. They may be needed to support the proper care of the patient from whom they were obtained.
- Traditional data-recording techniques and a paper record may have worked reasonably well when care was given by a single physician over the life of a patient.
- However, given the increased complexity of modern health care, the broadly trained team of individuals who are involved in a patient’s care, and the need for multiple providers to access a patient’s data and to communicate effectively with one another through the chart, the paper record no longer adequately supports optimal care of individual patients.
- Another problem occurs because traditional paper-based data recording techniques have made clinical research across populations of patients extremely cumbersome.

5.3. Electronic Health Record (EHR):

An electronic health record (EHR) is a digital version of a patient’s paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider’s office and can be inclusive of a broader view of a patient’s care.

EHRs are a vital part of health IT and can:

- Contain a patient’s medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results.
- Allow access to evidence-based tools that providers can use to make decisions about a patient’s care
- Automate and streamline provider workflow

An electronic health record (EHR) is an individual's official health document that is shared among multiple facilities and agencies. The role of EHRs is becoming increasing influential as more patient information becomes digital and larger numbers of consumers express a desire to have mobile access to their health records. Among other types of data, an EHR typically includes:

- Contact information
- Information about visits to healthcare professionals
- Allergies
- Insurance information
- Family history
- Immunization status
- Information about any conditions or diseases
- A list of medications
- Records of hospitalization
- Information about any surgeries or procedures performed

It is also becoming more common to see medical images attached to EHRs.

5.3.1. Key Features of an HHR:

One of the key features of an EHR is that health information can be created and managed by authorized providers in a digital format capable of being shared with other providers across more than one health care organization. EHRs are built to share information with other health care providers and organizations – such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and school and workplace clinics – so they contain information from all clinicians involved in a patient’s care.

There are a series of common and essential features that any EHR system offers. For starters, EHR platforms often set up a patient portal for consumers to access information as well as allow for secure data sharing and data access from other healthcare organizations.

EHRs also typically place patient care orders for clinicians, such as medication orders and diagnostic test requests. In terms of medications, EHRs can manage doses for specific patients and alert physicians to any possible drug interactions. The systems can additionally manage order sets, results and patient consents and authorizations.

Further, electronic health record systems often help coordination clinician workflow management and scheduling. Finally, these systems offer assistance in completing clinical, financial and administrative coding. This feature includes support of service requests and claims for reimbursement.

5.3.2. Advantages of Electronic Health Records:

EHRs and the ability to exchange health information electronically can help you provide higher quality and safer care for patients while creating tangible enhancements for your organization. EHRs help providers better manage care for patients and provide better health care by:

- Providing accurate, up-to-date, and complete information about patients at the point of care.
- Enabling quick access to patient records for more coordinated, efficient care.
- Securely sharing electronic information with patients and other clinicians.
- Helping providers more effectively diagnose patients, reduce medical errors, and provide safer care.
- Improving patient and provider interaction and communication, as well as health care convenience.
- Enabling safer, more reliable prescribing.
- Helping promote legible, complete documentation and accurate, streamlined coding and billing.
- Enhancing privacy and security of patient data.
- Helping providers improve productivity and work-life balance.
- Enabling providers to improve efficiency and meet their business goals.
- Reducing costs through decreased paperwork, improved safety, reduced duplication of testing, and improved health.

5.3.3. Benefits of Electronic Health Records:

Electronic Health Records (EHRs) are the first step to transformed health care.

The benefits of electronic health records include:

- Better health care by improving all aspects of patient care, including safety, effectiveness, patient-centeredness, communication, education, timeliness, efficiency, and equity.
- Better health by encouraging healthier lifestyles in the entire population, including increased physical activity, better nutrition, avoidance of behavioral risks, and wider use of preventative care.
- Improved efficiencies and lower health care costs by promoting preventative medicine and improved coordination of health care services, as well as by reducing waste and redundant tests.
- Better clinical decision making by integrating patient information from multiple sources.

With fully functional EHRs, all members of the team have ready access to the latest information allowing for more coordinated, patient-centered care. With EHRs:

- The information gathered by the primary care provider tells the emergency department clinician about the patient’s life threatening allergy, so that care can be adjusted appropriately, even if the patient is unconscious.
- A patient can log on to his own record and see the trend of the lab results over the last year, which can help motivate him to take his medications and keep up with the lifestyle changes that have improved the numbers.
- The lab results run last week are already in the record to tell the specialist what she needs to know without running duplicate tests.
- The clinician’s notes from the patient’s hospital stay can help inform the discharge instructions and follow-up care and enable the patient to move from one care setting to another more smoothly.

5.4. Difference Between EHR and EMR:

Electronic medical records **(EMRs)** are a digital version of the paper charts in the clinician’s office. An EMR contains the medical and treatment history of the patients in one practice. EMRs have advantages over paper records. For example, EMRs allow clinicians to:

- Track data over time
- Easily identify which patients are due for preventive screenings or checkups

- Check how their patients are doing on certain parameters—such as blood pressure readings or vaccinations
- Monitor and improve overall quality of care within the practice

Electronic health records **(EHRs)** do all those things—and more. EHRs focus on the total health of the patient—going beyond standard clinical data collected in the provider's office and inclusive of a broader view on a patient's care. EHRs are designed to reach out beyond the health organization that originally collects and compiles the information.

They are built to share information with other health care providers, such as laboratories and specialists, so they contain information from all the clinicians involved in the patient's care. The National Alliance for Health Information Technology stated that EHR data "can be created, managed, and consulted by authorized clinicians and staff across more than one healthcare organization."

The information moves with the patient—to the specialist, the hospital, the nursing home, the next state or even across the country. In comparing the differences between record types, HIMSS Analytics stated that, "The EHR represents the ability to easily share medical information among stakeholders and to have a patient's information follow him or her through the various modalities of care engaged by that individual." EHRs are designed to be accessed by all people involved in the patients care—including the patients themselves. Indeed, that is an explicit expectation in the Stage 1 definition of "meaningful use" of EHRs.

And that makes all the difference. Because when information is shared in a secure way, it becomes more powerful. Health care is a team effort, and shared information supports that effort. After all, much of the value derived from the health care delivery system results from the effective communication of information from one party to another and, ultimately, the ability of multiple parties to engage in interactive communication of information.

5.5. PUBMED:

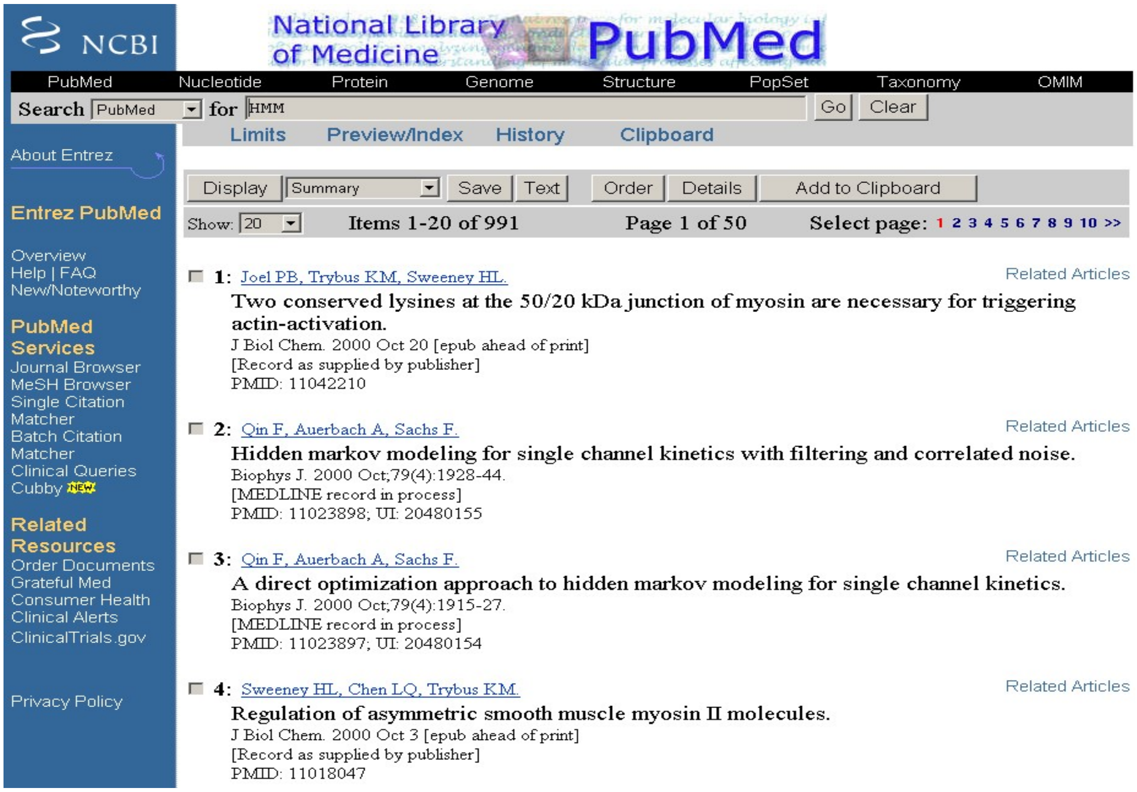
PubMed is a centralized NCBI repository containing millions of articles which can be accessed through a web-based interface provided within NCBI. It can be accessed from the site www.ncbi.nlm.nih.gov/pubmed. PubMed is developed and maintained by the National Centre for Biotechnology Information (NCBI) at the U.S National Library of Medicines (NLM), located at the National Institute of Health.

Pubmed is one of the databases in NCBI and it contains about 22 million records of articles published from MEDLINE and various other biomedical, life science journals. Online books containing links to full text from PubMed Central and links to other NCBI molecular biology resources are also present in Pubmed database.

PubMed offers different search options according to the user's convenience, which is very easy to use even for a beginner. One can search articles by different searching methods. Search can be done by any query word, by the name of the author, by the Journal name or by their PMID. PMID (PubMed identifier or PubMed unique identifier) is a unique number assigned to each PubMed record. PubMed offers to sort the records according to the date of publishment of papers. Results are displayed year wise which can be accessed from a graphical view displayed on the right side of the PubMed result page.

PubMed offers additional filter options to make our search more specific. User can sort their search by choosing options from Species (Humans/Animals), Sex, Article types, Languages, Subjects, Ages, and Journal Categories. After selecting filters, user can activate his search by clicking the "Search" button, in which result will be displaying as a message on the top of the page .The activated filters will be remain until it get deactivated by selecting the "clear" button. PubMed also offers to save our search records. PubMed Central (PMC) images can also be viewed from PubMed result page. It also allows downloading of documents in many standard formats that are usable through citation managers such as BibTeX, or some XML based citation extractors. It is an archive of free journal publication in life sciences field. It allows searching number of journals thereby providing the list of publications and their full text content without copyright violation.

PubMed is a free web-based public access resource that supports the search and retrieval of literature from the National Library of Medicine's MEDLINE database. In the past 2 years, PubMed has been updated to improve functionality and add important new features, including a set of search statements to identify COVID-19 articles and a new publication filter to allow PubMed searches to retrieve early-release preprints. This Viewpoint describes how these important new PubMed features and functions could allow clinicians to use a 3-step literature searching process to obtain real-time answers to important clinical questions.



Items 1-20 of 991 Page 1 of 50

☐ 1: Joel PB, Trybus KM, Sweeney HL.
Two conserved lysines at the 50/20 kDa junction of myosin are necessary for triggering actin-activation.
J Biol Chem. 2000 Oct 20 [epub ahead of print]
[Record as supplied by publisher]
PMID: 11042210

☐ 2: Qin F, Auerbach A, Sachs F.
Hidden markov modeling for single channel kinetics with filtering and correlated noise.
Biophys J. 2000 Oct;79(4):1928-44.
[MEDLINE record in process]
PMID: 11023898; UI: 20480155

☐ 3: Qin F, Auerbach A, Sachs F.
A direct optimization approach to hidden markov modeling for single channel kinetics.
Biophys J. 2000 Oct;79(4):1915-27.
[MEDLINE record in process]
PMID: 11023897; UI: 20480154

☐ 4: Sweeney HL, Chen LQ, Trybus KM.
Regulation of asymmetric smooth muscle myosin II molecules.
J Biol Chem. 2000 Oct 3 [epub ahead of print]
[Record as supplied by publisher]
PMID: 11018047

Figure 5.1. PubMed - NCBI

5.5.1. Effective Pubmed Searching In 3 Simple Steps

Step 1: Focus the Clinical Question

Use the PICO format (patient, intervention, comparator, outcome) to concisely express the type of clinical question to be answered (eg, therapy, diagnosis, etiology, prognosis).

Start by writing down common medical phrases that describe each PICO component of the clinical question.

Step 2: Use PubMed Clinical Query Filters

PubMed clinical query filters are complex, multiterm search strategies implemented by using simple key terms or drop-down menus. Identify updated systematic reviews or meta-analyses with key search phrases and the statement (systematic review [filter] OR meta-analysis [filter]) typed into the main PubMed search page.

Search for recent primary research studies with the clinical study categories filter most appropriate for the clinical question type (therapy, diagnosis, etiology, or prognosis) from the clinical queries page drop-down menu.

To explicitly search for NIH-sponsored COVID-19 preprints, use key search phrases and the term preprint[filter] typed into any PubMed search box.

Step 3: Refine the Search Terms

Every PubMed article is indexed with a controlled medical vocabulary called MeSH. PubMed automatically translates common medical phrases used in any search to appropriate MeSH terms before it conducts the search.

5.6. Genecards - Human Gene Database:

GeneCards is a searchable, integrative human gene database that provides comprehensive, user-friendly information on all annotated and predicted human genes. The knowledgebase automatically integrates gene-centric data from ~150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information. It is being developed and maintained by the Crown Human Genome Center at the Weizmann Institute of Science. The database aims at providing a quick overview of the current available biomedical information about the searched gene, including the human genes, the encoded proteins, and the relevant diseases.

The GeneCards database provides access to free Web resources about more than 7000 all known human genes that integrated from >90 data resources, such as HGNC, Ensembl, and NCBI. The core gene list is based on approved gene symbols published by the HUGO Gene Nomenclature Committee (HGNC). The information is carefully gathered and selected from these databases by its engine. If the search does not return any results, this database will give several suggestions to help users accomplish their search depending on the type of query and offer direct links to other databases' search engine.

Initially, the GeneCards database had two main features: delivery of integrated biomedical information for a gene in 'card' format, and a text-based search engine. Since 1998, the database has integrated more data resources and data types, such as protein expression and gene network information. It has also improved the speed and sophistication of the search engine, and expanded from a gene-centric dogma to contain gene-set analyses. Version 3 of the database gathers information from more than 90 database resources based on a consolidated gene list. It has also added a suite of GeneCards tools which focus on more specific purposes. GeneNote and GeneAnnot for transcriptome analyses, GeneLoc for genomic locations and markers, GeneALaCart for batch queries and GeneDecks for finding functional partners and for gene set distillations. The database updates on a 3-year cycle of planning, implementation, development, semi-automated quality assurance, and deployment.

5.6.1. Gene Card Suite:

GeneDecks is a novel analysis tool to identify similar or partner genes, which provides a similarity metric by highlighting shared descriptors between genes, based on GeneCards' unique wealth of combinatorial annotations of human genes.

5.6.1.1. Annotation combinatory: Using GeneDecks, one can get a set of similar genes for a particular gene with a selected combinatorial annotation. The summary table result in ranking the different level of similarity between the identified genes and the probe gene.

5.6.1.2. Annotation unification: Different data sources often offer annotations with heterogeneous naming system. Annotation unification of GeneDecks is based on the similarity in GeneCards gene-content space detection algorithms.

5.6.1.3. Partner hunting: In GeneDecks's Partner Hunter, users give a query gene, and the system seeks similar genes based on combinatorial similarity of weighted attributes.

5.6.1.4. Set distillation: In Set distiller, users give a set of genes, and the system ranks attributes by their degree of sharing within a given gene set. Like Partner Hunter, it enables sophisticated investigation of a variety of gene sets, of diverse origins, for discovering and elucidating relevant biological patterns, thus enhancing systematic genomics and systems biology scrutiny.