Chapter 1 INTRODUCTION

TechieAid also offers one-on-one coaching and mentoring program. They offer their clients the best candidates at any level with the fastest turnaround time. TechieAid also provides a reliable cost effective temporary staffing solutions that offer the ability to build the staff strength without absorbing them full time, assist overloaded employees during critical times and keep projects moving.

**1.2 Introduction to Data Science**

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from [data](https://en.wikipedia.org/wiki/Data) in various forms, both structured and unstructured, similar to [data mining.](https://en.wikipedia.org/wiki/Data_mining) It is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields within the context of [mathematics,](https://en.wikipedia.org/wiki/Mathematics) [statistics,](https://en.wikipedia.org/wiki/Statistics) [information science,](https://en.wikipedia.org/wiki/Information_science) and [computer science.](https://en.wikipedia.org/wiki/Computer_science)

Data analysis is a process of inspecting, [cleansing,](https://en.wikipedia.org/wiki/Data_cleansing) [transforming,](https://en.wikipedia.org/wiki/Data_transformation) and [modeling](https://en.wikipedia.org/wiki/Data_modeling) [data](https://en.wikipedia.org/wiki/Data) with the goal of discovering useful information, informing conclusions and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

Data visualization is viewed by many disciplines as a modern equivalent of [visual](https://en.wikipedia.org/wiki/Visual_communication) [communication.](https://en.wikipedia.org/wiki/Visual_communication) To communicate information clearly and efficiently, data visualization uses [statistical and information graphics,](https://en.wikipedia.org/wiki/Statistical_graphics) [plots,](https://en.wikipedia.org/wiki/Plot_(graphics)) and other tools. Numerical data may be encoded to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may make comparisons or understanding [causality](https://en.wikipedia.org/wiki/Causality) and the design principle of the graphic follows the task. Tables are used where users will look up a specific measurement, while charts are used to show patterns or relationships in the data for one or more variables.

[Data mining](https://en.wikipedia.org/wiki/Data_mining) is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. [Predictive](https://en.wikipedia.org/wiki/Predictive_analytics) [analytics](https://en.wikipedia.org/wiki/Predictive_analytics) focuses on application of statistical models for predictive forecasting or classification, while [text analytics](https://en.wikipedia.org/wiki/Text_analytics) applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of [unstructured data.](https://en.wikipedia.org/wiki/Unstructured_data) All of the above are varieties of data analysis.

|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 2 |

Chapter 1 INTRODUCTION

**1.3 Overview of Machine learning**

Machine learning is a field of [computer science](https://en.wikipedia.org/wiki/Computer_science) that uses statistical techniques to give [computer](https://en.wikipedia.org/wiki/Computer_systems) [systems](https://en.wikipedia.org/wiki/Computer_systems) the ability to learn (e.g., progressively improve performance on a specific task) with [data,](https://en.wikipedia.org/wiki/Data) without being explicitly programmed. Machine learning is closely related to [computational statistics,](https://en.wikipedia.org/wiki/Computational_statistics) which also focuses on prediction-making through the use of computers. It has strong ties to [mathematical optimization,](https://en.wikipedia.org/wiki/Mathematical_optimization) which delivers methods, theory and application domains to the field.

Within the field of [data analytics,](https://en.wikipedia.org/wiki/Data_analytics) machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as [predictive analytics.](https://en.wikipedia.org/wiki/Predictive_analytics) These analytical models allow researchers, [data scientists,](https://en.wikipedia.org/wiki/Data_science) engineers, and analysts to produce reliable, repeatable decisions and results and uncover hidden insights through learning from historical relationships and trends in the data.

Machine learning tasks are typically classified into two broad categories, depending on whether there is a learning signal or feedback available to a learning system:

* + [**Supervised learning:**](https://en.wikipedia.org/wiki/Supervised_learning) Thecomputer is presented with example inputs and theirdesired outputs, given by a teacher, and the goal is to learn a general rule that [maps](https://en.wikipedia.org/wiki/Map_(mathematics)) inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback.
  + [**Semi-supervised learning:**](https://en.wikipedia.org/wiki/Semi-supervised_learning) Thecomputer is given only an incomplete trainingsignal, a training set with some (often many) of the target outputs missing.
  + [**Unsupervised learning:**](https://en.wikipedia.org/wiki/Unsupervised_learning) Nolabels are given to the learning algorithm, leaving it onits own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end [(feature learning)](https://en.wikipedia.org/wiki/Feature_learning).
* The most widely used learning algorithms are:
  + [Support Vector Machines](https://en.wikipedia.org/wiki/Support_Vector_Machines)
  + [Linear regression](https://en.wikipedia.org/wiki/Linear_regression)
  + [Logistic regression](https://en.wikipedia.org/wiki/Logistic_regression)
  + Polynomial regression
  + [Decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning)
  + [k-nearest neighbour algorithm](https://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm)
  + [Neural Networks](https://en.wikipedia.org/wiki/Artificial_neural_network)
  + Clustering analysis

|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 3 |

Chapter 1 INTRODUCTION

* Some of the application of machine learning are:
  + In [classification,](https://en.wikipedia.org/wiki/Statistical_classification) inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more [(multi-label classification)](https://en.wikipedia.org/wiki/Multi-label_classification) of these classes. This is typically tackled in a supervised way.
  + In [clustering,](https://en.wikipedia.org/wiki/Cluster_analysis) a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
  + [Density estimation](https://en.wikipedia.org/wiki/Density_estimation) finds the [distribution](https://en.wikipedia.org/wiki/Probability_distribution) of inputs in some space.
  + [Dimensionality reduction](https://en.wikipedia.org/wiki/Dimensionality_reduction) simplifies inputs by mapping them into a lower-dimensional space. [Topic modelling](https://en.wikipedia.org/wiki/Topic_modeling) is a related problem, where a program is given a list of [human language](https://en.wikipedia.org/wiki/Natural_language) documents and is tasked to find out which documents cover similar topics.

**1.4 About the Project**

“Finding the Stage of Cancer malignant or benign” is a machine learning project which is used to predict the stage of cancer in a person. The Breast Cancer Wisconsin dataset is used for the implementation. The data set consists of 570 records which is split into training and testing samples. Each sample contains 33 features/attributes which describes the various biological features of a patient.

The machine learning algorithm used in the project is logistic regression which accepts the training samples as the basis to classify the test samples. Logistic regression is used to find the probability of event success and event failure. The dependent variable must be binary in nature (0 or 1). It is widely used for classification problems. Logistic regression as a special case of linear regression when the outcome variable is categorical. It is used to predict a discrete outcome based on variables which maybe discrete continuous or mixed. Some of the attributes on which the dataset is based are radius, smoothness, compactness, diagnosis, ID, etc. thus one can see that when the dependent variable has two or more discrete outcomes, logistic regression is a commonly used technique. Further one can see the use of Random Forest module which assures a high percentage of accuracy of the result obtained.

|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 4 |