1. **LITERATURE REVIEW**

1. **Yomna Omar et. al. (2018)**, described a Lung Cancer Prognosis System (LCPS) which provides oncologists with an accurate estimate of the health status of patients. They developed a prototype of a Lung Cancer Prognosis System (LCPS) which were making techniques accessible in a user-friendly manner in order to augment the knowledge of the expert oncologist and improve patient care and treatments. The decision tree was generated relevant input variables for prediction which determined using information gain. [37]

2. **Yeo B et. al. (2018)**, had investigated the influence of Information and communication technologies (ICTs) on service industry performance and given inadequate attention to service industries. This research has been carried out by Survey method which involves of 15 financial variables, out of these six constraints has been selected to avoid over fitting and for predictive analysis that affects company performance and growth. They have developed a Chi-Square Automatic Interaction Detector (CHAID) decision tree using SPPS Modeler v. 15 and used Descriptive statistics for the decision tree analysis.[38]

3. **Shaikhina T et. al. (2017)**, had used the applications of Machine Learning (ML) techniques for predictive modelling in clinical research and organ transplantation and constructed the group of Decision Tree models that was based on the standard CART algorithm with the highest prediction power. This decision tree was implemented using MATLAB and developed logistic regression likelihood multivariate model and explored the Decision Tree (DT) and Random Forest (RF) classification models in the 80 samples for prediction in high-risk kidney transplantation. The purpose of research was to improve strength of the classification tool, classifying incomplete and missing values by using Decision tree and RF model. They have suggested a model that helps a simulation tool to discover various clinical situations, identify the risk of ABMR of patients and also helps to patients for leaving more time to make essential adjustment to treatment.[17]

4. **Topiceanu A et.al. (2017)**, have used Decision tree learning for the classification of student archetypes in online courses namely, MOOC’s and presented six decision trees for participation rate, finalization rate and certification using recursive partitioning. The
Aim of the study was to set of archetypes which could quickly assess any student for to participate and finalize a course. They have proposed a classification tree for extracting valuable knowledge about advantages, disadvantages of participation in eLearning and reasons for not joining the courses and discussed supervised data mining techniques using random partitioning decision trees by using R language.

5. **Nichat A et.al. (2017)**, explained the predicting and analysis of student’s performance by using various data mining techniques and focused on the improvement of prediction and classification techniques such as Decision Tree, based on their academic performance. The purpose of this research was that when the student completes their test then the system would calculate the performance of the student by using Decision tree algorithm. They have constructed a classification model which classifies student performance, guide the students towards the achievement of good score and would help them to taking timely decisions. They have used C4.5, ID3 and has adapted a greedy approach. In this algorithm there was no backtracking and this design was based upon cross-sectional area. This research would help to the teacher that on which topics student was weak to prevent academic risk and desertion. [21]

6. **Liu X et.al. (2017)**, had used the data mining methods for the study of the different conditions of loss of grain storage and predicts the grain loss analysis based on decision tree algorithm. The aim of this research was to investigate, evaluation of the grain loss situation in country. Also, this research paper compares the application of classification forecasting algorithm in data mining. Regression (LR), Gaussian NB (GNB) and Decision Tree Classifier (DTC) in data mining were used. This model would help to different grain storage conditions analysis of the food losses caused by the situation.[15]

7. **Bondarenko et.al. (2017)**, described comparative existing rules extraction algorithms along with overviews of extraction process, how it was constructed into trained network training, pruning and knowledge extraction in the form of decision trees. Also proposed a recursive algorithm application and C4.5 Tree algorithm for construction of classification tree. The workflow for decompositional rule extraction has described by RX algorithm. They had focused on the improvements on the minimization of computational costs of pruning, clustarization process and suggested that extracted decision trees had good classification, accuracy and sizes comparable to C4.5 trees.[5]
8. **Ren L et.al. (2017)**, used decision tree technique to predict surface currents in a ¼-scaled marine renewable test site Galway Bay using the observations from the CODAR system. The decision tree model had constructed into training and testing model. They compared Scatter plots between CODAR measurements and predictions from Decision Tree models and evaluated prediction performance by calculating Statistics, bias, RMSE, Scatter Index (SI) and correlation (R). They have concluded that Decision tree model would generate satisfactory results for both surface velocity components at three analysis location using predicting dataset.[26]

9. **Daud, A et. al. (2017)**, has predicted the student performance by using different types of characteristics like academic record, family assets, income of a family and personal information. Study has been observed that expenditure of a family and their personal information might be affects the performance of the students and decrease the interest and concentration in their studies. The effectiveness of study has tested on scholarship holding students which provides both discriminative and generative classification models. The BN and NB classifiers were performed better than C4.5 and CART.[9]

10. **Osmanović A et. al. (2017)**, had implemented decision tree classifiers and artificial neural network to predict whether the patient will live with ovary cancer or not. They have used decision tree classifiers and artificial neural network to predict whether the person will live with that type of cancer. The feature selection would give which features were relevant or irrelevant for such type of decision for improved detection. The decision trees were generated by C4.5 that can be used for classification and referred as a statistical classifier. Cross validation techniques were used to evaluate predictive models by partitioning the original sample into training set to train the model.[22]

11. **Leena Patil et. al. (2017)**, have given new approach to detect the lung cancer by decision tree algorithm which provide effective result as compared to other algorithms. To minimize the process time and unnecessary checkups, the output of this prediction and classification model will help to patients to know that they are having Lung cancer or not. The various parameters namely, age, patient history and lung condition value were selected by correlation analysis. The information gain was obtained by comparing entropy before and after the split using particular feature. The decision tree was generated
by splitting and aggregation of data and useful for implementation of Lung cancer Disease with very much accuracy and fast. [16]

12. **Ortega-Campos et. al. (2016)**, identified profiles of risk and protective factors in juvenile offenders and showed a study of profiles of risk factors that impact young offenders towards committing sanctionable antisocial behavior. They have given multivariate approach to the phenomenon of repetitive sanctionable antisocial behavior in juvenile offenders and presented a descriptive study of the individual, criminological and contextual variables of the total group of juveniles and recidivist juveniles. A decision tree was constructed for total group of juveniles and applied for creating homogeneous groups based on the value of dependent variables and grouping subjects into two or more groups as a function of the predictive variables. The decision tree showed the relationship between strengthening and mitigating factors in predicting S-ASB recidivism.[23]

13. **Spanakis G et. al. (2016)**, presented a framework for warning people when they were at risk of unhealthy eating and predicts under which conditions participants were more possible consume unhealthy food. Data was collected through a mobile application called “ThinkSlim” which was developed for the purpose of studying eating behavior using Ecological Momentary Assessment (EMA) principles. This had performed in Random sampling and Event sampling. A decision tree algorithm has developed to predict healthy versus unhealthy eating events and proposed a methodology for construction of the decision tree containing warnings over possible unhealthy eating moments. The significance of the various tests were tested by Fisher Test and comparisons were tested between groups by Euclidean distance, standard Hierarchical Agglomerative Clustering (HAC) algorithm. [28]

14. **Bhatkande S S et.al. (2016)**, investigated the forecasting main attributes like maximum temperature, minimum temperature and for deleting the irrelevant data and developed a model using decision tree to predict weather phenomena like full cold, full hot and snow fall. In this research, they have used research procedure like Data Cleaning, Data Selection, Data Transformation and Data Mining for construction of a Decision tree and also used Regression, artificial neural network, fuzzy logic and group method of data handling. They have got decision tree model which separates a datasets into smaller datasets in the form of decision nodes and leaf nodes. From this research, they have
showed the influence of these parameters observed over the study period and identify changes in climatic patterns. [6]

15. **Agaoglu M (2016)**, explored the potential of Educational Data Mining (EDM) in enlightening the criteria or measures of effective instructor performance as perceived by the students. Course evaluation questionnaires based on students’ perception was used to evaluate instructor’s performance. Seven classification models were constructed using different classification algorithms namely C5.0 and CART decision trees; support vector machines; and neural network. Discriminant analysis was used to test importance of variables in the questionnaire and showed that it was useful to get rid of unnecessary variables, hence useful to identify important variables for tool construction. The research is useful to study instructor’s performance using student’s perceptions, which was depended on student’s interest in the respective subject.[1]

16. **Fraser J et. al. (2016)**, had examined the relations of Classification and regression tree (CART) to develop and validate a clinical tool to discriminate between patients at high risk of early death and longer term survival. This analysis was carried out using 29 variables with 18-month survival as the dependent variable. Decision tree analysis was performed on the whole data set using chi-square automatic detection. This method was used a chi-square significance test at each split for categorical or continuous variables. A Cox proportional hazards regression was performed to analyze the association of the different risk groups with survival. The comparisons were made using the Mann-Whitney U test and multiple comparisons were performed using the Kruskal-Wallis test. The model was internally validated using 10-fold cross-validation method.[7]

17. **Yoo K et. al. (2016)**, developed a new field based approach that could estimate patterns of groundwater pollution sensitivity using data mining algorithms. The objective of study was to evaluate validity of decision tree and rule induction in predicting groundwater TCE sensitivity using hydro geological input variables for a TCE contaminated site. They have developed a method for identifying key hydro geological input variables which influencing the groundwater TCE sensitivity and used four representative data mining methods DT, ANN, MLR, and CBR to estimate the groundwater pollution sensitivity to TCE contamination. They have also evaluated the performance of the data mining models where the data were partitioned into training and a testing set and applied Ordinal
Pairwise Partitioning (OPP) method to improve the performance of the data mining results. In this study, Decision tree has identified hydro geological parameters which have influence the groundwater TCE sensitivity. [39]

18. **Al-Obeidat F et. al. (2015)**, proposed a decision tree (DT) algorithm based on a fuzzy approach and used multi-spectral values of pixels and classify these central pixels in each neighbourhood from satellite images. They have investigated and evaluated the performance of proposed method against other DT classifiers namely with C4.5 and ID3 algorithms, and comparative and analytical study has been conducted. This study has introduced a new classification algorithm from the two main disciplines machine learning and MCDA method PROAFTN. This method was a purely numerical methodology that would be implemented to predict the class of a pixel of the given population of multi-spectral values of Landsat Multi-Spectral Scanner image data and implemented in Weka software using the stratified 10-fold cross-validation.[2]

19. **Song Y Y et. al. (2015)** described decision tree as effective method for data mining and used for developing prediction algorithms for a target variable. Decision tree has used for classification of missing values as a separate category and replaced those values with predicted values. The decision tree has been developed from historical data and predicts the future records. They also used many statistical algorithm for construction of decision tree namely, CART, C4.5,CHAID, QUEST in this research. Also, Study described concepts of pre-pruning, post-pruning which improved the accuracy of classification. Research has suggested that the disadvantage was related to over fitting and under fitting for small datasets.[29]

20. **Shahiri et. al. (2015)**, had calculated prediction of students performance by predictive modelling. To construction of predictive modelling, Regression, classification and categorization were used. There were different algorithms used in classification could be applied for predicting the student performances. Among the all algorithms, Decision tree, Artificial Neural Networks, Naïve Bayes, K- Neighbour and Support Vector Machine are used.[30]

21. **Khobragade et. al. (2015)**, proposed a model on prediction of academic failure of students using Data mining Techniques. The objective of the research was to identify the failure of students to avoid them from dropping out and improve their academic
performance. They have applied different classification techniques namely, induction rules, decision rules and Naïve Bayes for prediction. The Decision tree algorithm would help the Teacher and students to improve their performance by changing some regulation in their teaching methodologies. This study would help to teachers and Principal to make suitable provisions to increase the ability of students in their academics.[13]

22. Venkatesan E. et. al. (2015), has used Data mining approach to quick recognition of disease over symptoms. A classification technique was used to predict the hidden information in the medical domain. They had analyzed the breast cancer data using classification algorithms namely J48, Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree) and Best First Tree (BF Tree) for breast cancer data to classify the data set for classification and prediction. They have compared the four decision tree algorithms in the prediction of the performance accuracy in breast cancer data. Among these four methods, they have concluded that the decision tree is best algorithm for the chosen input data for supervised learning algorithms to predict the best classifier.[34]

23. Lee J S et.al. (2014), investigated that the location can be predicted by observing a sequence of human activities. They have proposed a set of significant rules which showed a sequence of user’s activities through which proper location was finally possible by using decision tree. This research had explained about user could performed various specific activities at specific moments by using decision tree. These activities were computed by using mobile devices, such as smart phones, tablet computers, and laptops, naïve location Prediction. This research would help to predict an activity of a person would engage in after having performed another specific activity and easy to make timely and suitable decisions about location prediction.[17]

24. Acharya A et. al. (2014), applied different Machine Learning Techniques for prediction of student results and to predict the performance of student early by using previous semester results. The objectives of the study were to identify the attributes and examination pattern of students, perform and classify features selection which were used for classification. The model was constructed into two steps firstly, classifier model was constructed using dataset and secondly tested an accuracy of classification rules. They have calculated feature evaluation and information gain attribute evaluation. Selection of
various features Wrapper based feature selection methods were used. To analyze the various algorithms, Paired t-test were used and to check the efficiency of model, Kappa statistic and F-test values were used. The degree of significance was determined using P-value into four groups as Not Statistically Significant, Not Quite statistically significant, statistically significant and statistically very significant.[3]

25. Zeng et. al. (2014), had used survey data to develop a decision Tree model for forecasting the popularity of large number of Chinese colleges with average passing score. Here popularity of colleges were considered as time series forecasting problem. They have been proposed efficient decision tree based on splitting criterion and decision tree pruning. In the construction of decision tree, passing scores and population change ratio were considered as attributes. The model were compared with two classification methods namely, Naïve Bayes and SVM and investigational effects showed that the algorithm was more effective as compared with previous methods and has practical approach for predicting the status of colleges.[40]

26. Taruna S et.al. (2014), had compared five Classification Techniques namely, Decision tree, Naïve Bayes, Bayesian Network, NB Tree and K-Nearest Neighbor methods for engineering students for predicting performance of grades. To develop a decision tree algorithm, they have used previous academic performance of students. Attributes were selected by Gini index, Information gain and Gain ratio for decision tree model. [32]

27. Geetha, A et. al. (2014), has used Decision tree for Weather prediction to predict the dependent variables like temperature, rainfall, humidity, wind speed and direction. They have used Decision tree for modeling of prediction of Rainfall and as a visual and analytical tool, to predict the target values. This historical data for a classification project has divided into two data sets: one for building the model; the other for testing the model. The original dataset contained nearly 20 parameters, but they have considered only the 12 parameters, which were relevant and vital for rainfall prediction. Decision Trees were generated by recursive partitioning. This paper also has measured significant suggested that if the weather warnings would inform to the common man about natural calamity so that it would act as a lifesaving message, a value added service and as a disaster prevention tool.[10]
28. Chourasia et. al. (2013), have used tree structure as decision tree which divides input sample into its possible classes. Decision tree was used to absorb valuable information by creating a decision rules from a large number of available database. A decision tree classifies data into simpler form which can be efficiently stored and classifies effectively in new data. This paper reviews different algorithms using decision tree to classify various datasets.[8]

29. V Krishnaiah et. al. (2013), have examined data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network for health care data. For data processing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) were used. They have proposed model for early detection and correct diagnosis of the disease which would help the doctors in saving the life of the patient. This method was based on systematic study of symptoms and risk factors. The effective model was formed to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. They found the target group of people who needs further screening for Lung cancer disease and used other data mining techniques like Time Series, Clustering and Association Rules.[35]

30. G. Sujatha et. al. (2013), had compared the accuracy of various decision tree classifiers on Tumor Data sets. The unsupervised discrimination was used for converting continuous attributes to categorical attributes. They showed the accuracy of ID3, C4.5 and CART algorithms for classification using 10-fold cross validation. C4.5 algorithm was used for a set of Primary tumour and ID3 was used for Colon tumor data.[11]

31. Xiaohu W et. al. (2012), were related to the application of Decision tree ID3 of data mining for actual sales data. Data was cleaned using noise smoothing processing. If User choose the most common area then correlation analysis used for irrelevant attributes. The decision tree was constructed for maximum entropy that’s based on information gain measure for the experimental information. This study was applied on online user to buy data mining the probable buying power of the user.[36]

32. Li et. al. (2012), used Fuzzy ID3 algorithm in evaluation of performance of employees and that could implement to staffs appraisal and management. This algorithm would help to Fuzzy logic toolbox in Matlab software were used for blur the original data to put the
regularity between the data. Trapezoidal Membership was calculated to unity, the data was converted from quantitative to qualitative. Study would help to encourage the staff to continuously improve work efficiently. The generating fuzzy ID3 had used in a wide area of customer satisfaction evaluation, customer relationship management, economic analysis, rule extraction and so on.[18]

33. **Takada M et.al. (2012)**, have proposed to develop a new mathematical model to predict AxLN metastasis in patients with primary breast cancer using preoperative clinic pathological information by using decision tree method. The prediction model contains multiple decision trees which were trained on generated datasets. To eliminate multicollinearity, they have used Machine learning techniques as decision tree. The modelling method showed accurate and versatile prediction using large number of variables. They have used WEKA software for resampling and Mann-Whitney test for calculating confidence interval. [33]

34. **Li et. al. (2010)**, used various algorithms for classification of a huge data namely, Linear programming, neural network, statistics and Decision tree etc. Among all of these, Decision tree was one of most popular and powerful approach in data mining. Now a days, all science and technology were exploring large and compound type of data to find some relevant and useful information for modeling and making the process more efficient, cost-effective and accurate. Basically, Decision tree was executed in decision theory and statistics. The decision tree algorithm was very beneficial in data mining for handling a variety of nominal, numeric and textual type of response documents and various datasets where datasets having a large number of errors and missing values.[19]

35. **Pumpuang et. al. (2008)**, proposed and compared the classifier algorithms for model of Course Registration Planning. The proposed model was used to predict the grade point average (GPA) of undergraduate students. In this study, they have used four classifiers for comparison of performances namely, Bayesian Network, Decision forest, C4.5, NM Tree. Out of these, NB Tree was showed the best predictive measures and selected as Course Registration Planning model for the students. [24]

36. **Lee et. al. (2007)**, dedicated on the encouragement of customer’s for use of e-commerce through some attractive characteristics. To understand the complex relationships between the online preference and service feature, various DT algorithms such as CHAID, C4.5,
C5.0, and CART were applied. They have used Chi-square test value for splitting criteria and variables which showed the highest accuracy of classification. Further, in this research, a prediction model was developed for online success which was based on the customer’s appreciations. They have developed a decision tree for prediction model for customer’s preferences of online channel over offline channel from collected data. Also, they have developed another two prediction models using logistic regression and discriminant analysis and evaluated its accuracy and compared their results. To combine interrelated variables into independent factor a principal component analysis (PCA) was used. [20]

37. Podgorelec V et. al. (2005), had presented a new outlier prediction method for improving the classification performance and taken the filtering of the two cardiovascular training datasets for study. They have also suggested that it would help the clinicians in diagnosing the various types of the cardiovascular problems in young patients and used classification algorithms like ID3, C4.5, Naïve-Bayes (N-B), and instance-based classifier (IB) for both original and filtered training set for comparing classification results. All the results were the averages of 10-fold cross-validation and the decision tree classifiers showed an improvement of the classification performance on a smaller, less complex dataset, where outliers have negative effect on the training process, whereas on a bigger and more complex datasets, the classification performances were decreased.[25]