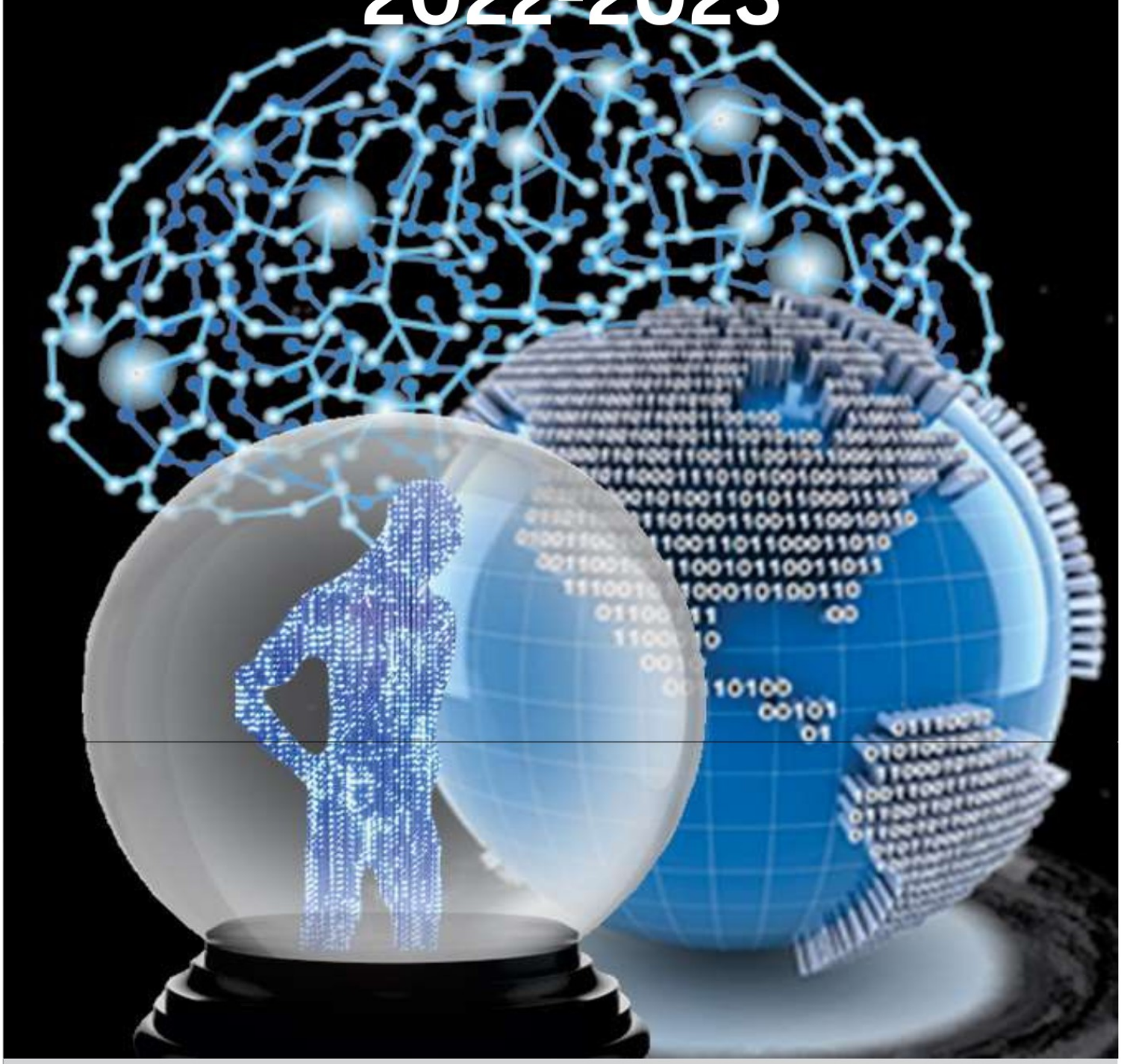
BHARATI VIDYAPEETH'S

Technical Magazine

2022-2023



COMPILATION OF BEST RESEARCH PAPERS

Table of Content

Machine Learning Based Developmental Capability Prediction: A Diagnosis to the Learning Capacity Disorder for Specially-Abled Children

Priya Chandran*, Suhasini Vijaykumar, Gunjan Behl, Shravani Pawar, Nidhi, Manish Dubey, and Vasudha Arora

Bharati Vidyapeeth's Institute of Management & Information Technology, Navi Mumbai, Maharashtra, India
Email: priyaci2005@gmail.com (P.C.); suhasini.kottur12@gmail.com (S.V.); mailto.gunjan83@gmail.com (G.B.); smitakapase@gmail.com (S.P.); mca.nidhipoonia@gmail.com (N.P.); dby.manish@gmail.com (M.D.); vasudha131999@gmail.com (V.A.)
*Corresponding author

Manuscript received July 3, 2023; revised August 8, 2023; accepted November 1, 2023; published February 15, 2024

***Abstract*—Specially-abled people are recognized and acknowledged for their issues such as hyperactivity, learning disorder, proprioceptive sensory issues, problems in self-help skills and problems in various motor skills such as Gross Motor Skills (GMS), Fine Motor Skills (FMS) and Oral Motor Skills (OMS), This study sought to identify effective machine-learning-based classification models to predict developmental capability disorders and thereby addressing of the learning disorder issue at opportune time. We have used machine learning classification algorithms Decision Tree, Random Forest, K-nearest neighbors, and Logistic Regression for the developmental capability prediction of individuals. The generalized progress monitoring datasets were carried out by interpreting and visualizing gender, age and disability-specific developmental competence. We have collected dataset from an occupational therapist for the study. The results of the study show that the Random Forest algorithm has a high accuracy of 95% compared to other algorithms that we have implemented.**

***Keywords*—learning disorders, ASD, disability, occupational therapy, speech therapy, machine learning**

I. Introduction

Developmental capability refers to the ability of an individual to respond to the external environment and develop accordingly. Each individual has a different rate of development. The development process slows down due to various reasons like genetic disorders, environmental or social conditions, etc., but one of the important impetus is towards learning disorder, which will gradually retard the learning process of the individuals. This is a complex neurodevelopmental disorder that can have long-lasting impacts on individuals and their families. Delayed diagnosis and intervention can hinder the progress of children with this disability, leading to missed opportunities for early intervention and support. The motivation leading to this predictive model is to identify if the individual is carrying any traits, if not arrested, would ultimately lead to any type of disorder. In this study we have done an examination of cognitive abilities in children with disabilities associated with learning. Predicting developmental capability accurately and promptly is a significant challenge faced by healthcare professionals. Traditional methods often lack the efficiency and accuracy needed to identify this at an early stage. This research tries to relate the degree of disability and how it impacts the development process. The World Health Organization (WHO) framework [1] on the International Classification of Functioning, Disability and Health (ICF) computes the degree of disability at both individual and population levels. It defines three stages for human functioning: the first stage is at body level (considering body formation and functional domain), second is at activity level (considering all ADL that is activities of daily living) and third is at participation level (considering the socialization of person in varied surroundings), which is shown in Fig. 1. It also describes the reasons for disability: first the environmental factors (encompassing both physical and social surroundings) and second the personal factors (considering the health issues, age and family history). Types of disabilities are shown in Table 1. The extent of disability is estimated on the basis inability to perform at one or more of these stages [2].

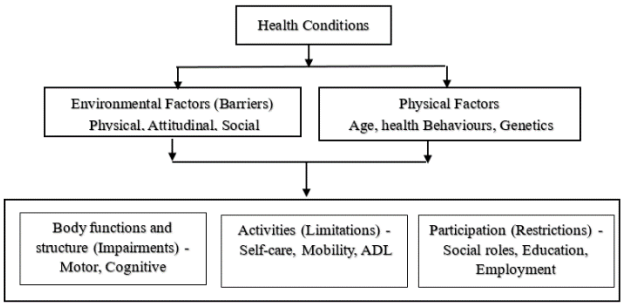


Fig. 1. ICF model.

Table 1. Types of disability

Disability	Description
Learning Disability (LD)	The ability to learn and apply the knowledge is delayed. The degree of impairment could vary. The learning process is slow. The prevalent problems are in reading and writing.
ADHD	The ability to pay attention is difficult due to hyperactivity. The sitting tolerance is also low. It is most prevalent in children. As a result, the learning process is also delayed. It is very important to utilize their excessive energy by indulging in physical activities.
Autism	Neuro-developmental disorder which affects the overall functioning and personality of an individual. Specialization of such individual is most effected. The self-time (time spent alone without interaction) is high.
Cerebral Palsy	The ability to maintain the balance and posture is delayed.

The developmental capability prediction model aims to achieve accurate predictions about whether an individual has a high developmental capability or not. The optimization of progress tracking and monitoring of individual is an essential

aspect of this system. Most of the research focused on the application of technical advancements in implementing new learning methods that can be adapted for specially-abled children. Some of the state-of-the-art mechanisms used for assisting these children are based on machine learning and deep learning techniques.

Learning disorders, problems in various motor skills like gross motor skills, fine motor skills and oral motor skills, proprioceptive sensory issues, and problems in self-help skills issues are recognized as etiological factors for identifying specially-abled people and their learning disorder. The purpose of the study is to speed up the diagnosis process and construction of a learning disorder therapy plan by predicting the developmental capability of individuals in an early stage. The therapy plan is like a blueprint for the therapy process. It enlists all the activities to be carried out during the therapy sessions. The optimization of the progress tracking and monitoring is an essential aspect of this system.

The goal of this research work is to present a developmental capability prediction model using machine learning algorithms implemented in Python. By harnessing the power of advanced algorithms and data analysis techniques, machine learning offers a ground-breaking solution for predicting disability with the assistance of machine learning models, healthcare professionals can accurately identify autism-like genetic disorders at an early stage, allowing for timely intervention and tailored support. By leveraging machine learning in developmental capability prediction, we can revolutionize the way we approach this disorder, ensuring that individuals receive the care they need from an early age. Based on the prediction results, the practitioner can identify the areas where a child excels and where they may require additional assistance and can also provide tailored support, interventions, and resources that cater specifically to their strengths and challenges. By utilizing our state-of-the-art machine learning tools, we can make a significant impact on early intervention and treatment strategies for children with disabilities.

In our study, the functional level value is taken as the target variable, which is used for computing the output. The output has two possible values High and Low. This study mainly focuses on the prediction of three important parameters, namely, functional level, duration of therapy and the development area to be focused on.

II. Literature Review

Because of the worldwide increase in autism rates, research towards early identification of disability has accelerated. The absence of normal activities, rather than the presence of abnormal ones, is one of autism's most fundamental symptoms. Several researches have been carried out in this domain. Most of the research focused mainly on one disability that is ASD (Autism Spectrum Disorder. In this study, we concentrate on all forms of disability that have an impact on people's daily lives. The concept of eye-tracking for ASD prediction and quantitative analysis of the social response to ASD diagnosis were discussed in [3, 4]. The researchers have used feature transformation techniques like log and Z-score in their study [4]. They have used different machine learning algorithms for the study of different age groups.

Even though, certain advancements were made like highlighting the age for developing a mobile application for autism screening to predict ASD traits among people of varied age groups [5]. Numerous feature selection techniques were carried out to determine which machine learning classifier gave the best results in categorizing ASD risk factors in toddlers, children, adolescents and adults [6–9]. Linear discriminant analysis classifier and K-nearest neighbor techniques were used for ASD diagnosis.

The main objective was to find out whether children have ASD or not. Researchers also used a deep-learning model for ASD trait classification [10–12]. They used a 4-layer neural network to implement the deep-learning model. The activation functions applied are basic rectified linear unit (ReLU), Hyperbolic Tangent activation function (Tanh) and sigmoid activation function. The majority of the research used the pre-defined ASD dataset called the AQ-10 dataset is used for building the prediction model [13,14]. In this research, a web application is used for live data capturing and dataset creation. AI augment-based learning technique is used for behavior analysis and as an assistant for making decisions [15]. A machine learning framework with four feature scaling strategies like quantile, power, normalizer and MaxABS scaler [16].

In the study, the authors have proposed a phased methodology for early prediction of child disability [17]. The three-phase approach includes dataset identification and preprocessing techniques, data modeling using machine learning approaches like multilayer perceptron and evaluating the performance of the model. This study in [18, 19] focuses on creating significant feature signatures for the early detection of autism by applying automated machine learning along with feature ranking approaches on the Q-chat-based dataset.

The authors examined the use of technology in diagnosing and evaluating abilities in socialization, emotion management, communication and addressing behavioral issues [20]. The article also suggests standards for gaining knowledge of potential technologies with applications for the care and understanding of people with autism spectrum disorders. The Deep Convolution Neural Network (DCNN) ensemble-based classification framework was proposed in [21] to detect ASD disabilities.

Most of the current studies used the existing dataset like the AQ-10 dataset, which is mainly used for the detection of autism traits. In our study, we have collected a primary dataset from occupational therapist, based on disability traits in toddlers, kids, and teens.

III. Research Methodology

The proposed system is implemented in Python. The framework of the proposed study is depicted in Fig. 2. We have used Decision Tree (DT), Random Forest (RF), KNN classifier and Logistic Regression (LR) machine learning algorithms to study, predict and analyze developmental capability prediction of individuals. The collected dataset consists of the features of the age group between 3 and 17.

A. Individual Assessment

Each individual's performance was assessed in various development areas. The five areas of development that were

taken into account in our study are Gross-Motor skills (GMS), Fie-Motor skills (FMS), Oral-motor skills (OMS), Cognitive Perceptual skills and Self-Help skills. These development areas help to understand and identify the co-occurrences of the disability conditions. For further examination, five different activities of each development area are taken into consideration. These areas and corresponding activities are shown in Table 2.

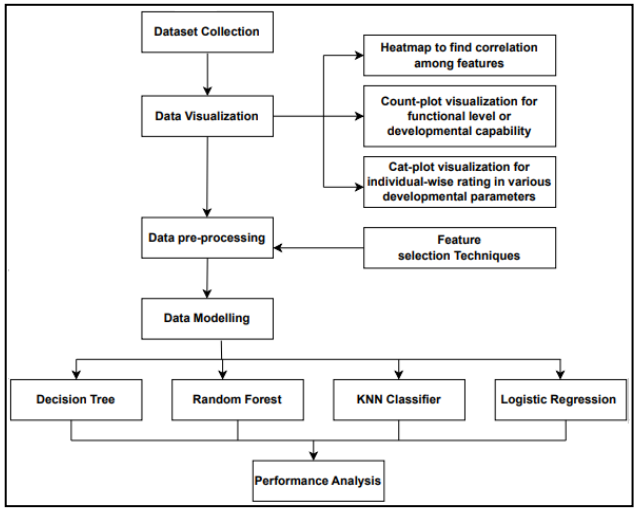


Fig. 2. Proposed study.

Table 2. Development areas

Area	Description	Activity
Gross-Motor skills (GMS)	Associated with mobility or body movements	Running, Bouncing, Ladder climbing, walking on balance beam, Going through Tunnel.
Fine-Motor skills (FMS)	Associated with more detailed or refined movements of wrists and hands.	String beads, Cloth Clip activity, Clay activity, building blocks, Scribbling with crayon
Oral-motor skills (OMS)	Associated with the oral cavity movements.	Sucking through straw, blowing (general), blowing candle, blowing whistle, protruding tongue
Cognitive Perceptual skills	Associated with overall thinking, understanding and applying of pre-learned knowledge	Matching objects, Putting shapes, Solving-puzzles, Identification of objects, identification of pictures
Self-Help Skills	Associated with the ability of the child to do self-care tasks and routine daily tasks.	Dressing, eating, bathing, tooth brushing, Hygiene maintenance

In this research we have done an examination of cognitive abilities in children with disabilities associated with learning. Cognitive perceptual skills associated with overall thinking, understanding and applying of pre-learned knowledge are the parameters for this examination. A rating scale approach was used where the lowest value being 1 and the highest being 4, which is shown in Table 3.

Table 3. Rating scale

Rating	Description
1	Does not Initiate
2	Partially Initiates
3	Initiates
4	Completes

B. Prediction

This research predicts three output values, functional level, duration and development areas to be focused on, which is shown in Table 4.

Table 4. Prediction parameters

Sr. No.	Value	Possible value
1	Functional level	High or Low
2	Duration	6 months, 1 to 2 years, 2 to 3 years, more than 3 years
3	Development Areas to be focused on	Gross-motor skills, Fine-motor skills, Oral-motor skills, Cognitive perceptual skills, Foundation development established in every area

1) Functional level

This output feature predicts the developmental capability of an individual. The functional level helps to identify the degree of disability of a child. For example, if there is an autistic child then with the help of the functional level we can identify at which level of disability spectrum the child is. The value of functional level high means that all the motor skills like GMS, FMS, etc. are properly developed. The child with proper therapy can carry out a normal life. A functional level value low means that all the motor skills of the child are not properly developed. Extensive therapy and care are necessary. With the help of this analysis, the therapist can plan the therapy process and decide to make modifications to the ongoing therapy process.

There are five development areas. Each area comprises five activities. For each activity, the highest rating possible is 4 and the lowest rating is 1. The therapy process takes into account the total of all the ratings. Then, the maximum possible total rating of an individual is computed as in Eq. (1):

$$M = n \times p \times h \tag{1}$$

where n is the total number of development areas, p is the total number of activities of each development area and h is the highest possible rating given to the individuals. In our study, h is equal to 4. Since each of the five development areas in our study has five distinct activities, the value of n and p is 5. M is calculated as follows:

$$M = 5 \times 5 \times 4 = 100$$

The disability spectrum can be classified as high-functional and low-functional. High functional individuals' brain growth will be taken into account as being greater than 50% [22, 23]. The threshold value, number and type of developmental areas and corresponding functional activities were identified with the help of an occupational therapist. Based on the value computed for M, the threshold value, T, is computed for identifying the functional level value using Eq. (2).

$$T = \mathrm{M}/2 = 100/2 = 50 \tag{2}$$

This threshold value decides whether an individual has high developmental capability or low developmental capability.

The equation for computing an individual's total rating, TR, is given in Eq. (3).

$$TR = \sum_{i,j=1}^{5} R_{i,j} \tag{3}$$

where $R_{i,j}$ is the rating of j^{th} activity of i^{th} development area.

Then the average value of TR is calculated to identify the diagnostic evaluation by the occupational Therapist. This developmental capability prediction process is shown in Fig. 3.

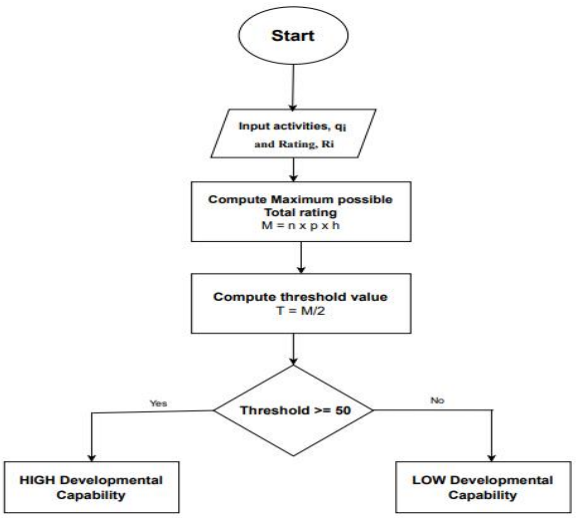


Fig. 3. Developmental capability prediction process.

2) Duration

Duration is a prediction parameter which predicts the time required to complete the basic foundation therapy of the individual. It is an estimated value based on the average rating computed. Since therapy is an ongoing process, the duration can be increased or decreased as per user's requirement. Therapy is a support given to an individual to normalize the learning and behavioural disorder. It can reduce the impact of the disability and help in the normal functioning of an individual. The total rating computed is used for calculating the average rating. Average rating= TR / p, where TR is the total rating and p is the number of activities of each development area. Based on this average the duration value estimate is computed as shown in Fig. 4. The occupational therapist begins diagnostic evaluation and decides the therapeutic strategies based on the duration identified.

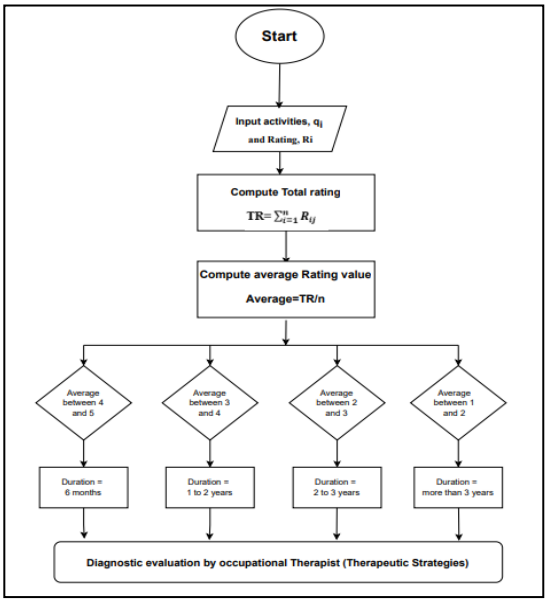


Fig. 4. Foundation therapy duration identification of the individual.

3) Development area

This predicts the development area that requires most attention. This will help in designing a better therapy plan for the individual. This takes into account the rating of each activity in that area. There are four areas. Each area has five activities respectively. The rating in each activity is considered for the prediction. The rating scale varies from 1 to 4. If the rating is less than or equal to two in any activity, then that area needs to be focussed on. Otherwise, the foundation development is established in every area.

IV. Implementation of Machine Learning Approaches

We have used primary data for the analysis and prediction of developmental capability. The dataset collected from the occupational therapist is used in the prediction model. In this research, we have taken 80% of the dataset as a training dataset and 20% as a testing dataset. This research used machine learning algorithms Decision Tree, Random Forest, KNN classifier and Logistic Regression for developmental capability prediction. The algorithms used multiple visualization technique heatmap for finding out the correlation among different features. The decision tree is further visualized using text representations and tree-like representations. The therapist assesses an individual on various development parameters like Gross-Motor Skills (GMS), Fine-Motor Skills (FMS), Oral-Motor Skills (OMS), and cognitive-perceptual skills with the help of a rating- scale. The main objective of the prediction system is to get a probabilistic value for the high developmental capability of an individual. This machine learning prediction model helps to identify whether there is improvement or there is a need to change the strategy of development.

A. Feature Description

We have collected data from an occupational therapist. The dataset collected consists of 28 features and 1 target value. These 28 features are defined by the therapist. From the dataset used to train our algorithm, only the name attribute, which has no bearing on prediction, has been eliminated. The functional level is the target value which has two possible values High and Low. This value is a probabilistic estimation of the developmental capability of an individual. If the calculated functional level value is greater than or equal to 0.5, then the developmental capability is considered as high and low otherwise. The feature description is depicted in Table 5.

Table 5. Feature description

Feature	Description
Name	Name of an individual
Jumping	Rating given in jumping activity under GMS
Bouncing	Rating given in bouncing activity under GMS
Ladder climbing	Rating given in ladder climbing activity under GMS
Walking on balance beam	Rating given in walking on balance beam activity under GMS
Going through tunnel	Rating given in going through tunnel activity under GMS
Stringing beads	Rating given in stringing beads activity under FMS
Cloth clip activity	Rating given in cloth clip activity under FMS
Clay activity	Rating given in clay activity under FMS

Building blocks	Rating given in building blocks activity under FMS
Scribbling through crayon	Rating given in scribbling through crayon activity under FMS
Sucking through straw	Rating given in sucking through straw activity under OMS
Blowing	Rating given in blowing activity under OMS
Blowing candle	Rating given in blowing candle activity
Blowing whistle	Rating given in blowing whistle activity
Protruding tongue	Rating given in protruding tongue activity
Matching objects	Rating given in matching objects activity
Putting shapes	Rating given in putting shapes activity
Solving puzzles	Rating given in solving puzzles activity
Identification of objects	Rating given in identification of objects
Identification of pictures	Rating given in identification of pictures
Total	Sum of all the ratings
Avg	Average of all the ratings given in each area
gender	Gender of Individual
duration	Estimated time for therapy
category	The specific age category
type	Type of disability
Functional level	The developmental capability prediction

B. Correlation among Features

Correlation between different features is visualized with the help of a heatmap, which is shown in Fig. 5. The darker shade represents a higher value and the lighter shade represents a lower value.

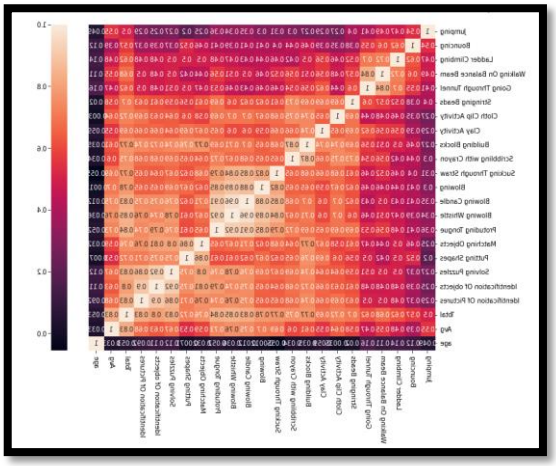


Fig. 5. Correlation among different features.

C. Feature Selection

Feature selection methods have a major role in maintaining system performance. The data is processed for missing values. We use the correlation heatmap to identify feature correlation. If the feature is highly correlated, then handle the missing value otherwise drop the unwanted feature. The dataset contains both numerical and categorical data. It is very important to convert categorical data into numerical form before data modelling. The conversion of all the object type features into numerical values is done using label encoding.

D. Machine Learning Models

1) Decision tree

A decision tree is a classification prediction algorithm. The Gini index is used to select the splitting attribute and based on this splitting attribute the decision tree is divided into two parts. The Gini index value is computed as follows:

Gini (D) =1−∑Pi, where Pi is the probability that a record D belongs to class C.

In this research, the target value, functional level, is used to compute the output. Output prediction depends on the functional level value. The condition is functional level ≧0.5 goes for class:0 (yes: functional level is high) and functional level <0.5 goes for class:1(no: functional level is not high).

2) Random forest

Random Forest algorithm builds decision trees on different samples and voting is carried out, as shown in Fig. 6. The majority vote is considered for classification. The RF approach uses numerous decision tree classifiers to improve the performance of the model DT are generated at random from the instances of the training set. As a result, each decision tree makes predictions. By majority vote, the model's final prediction is chosen.

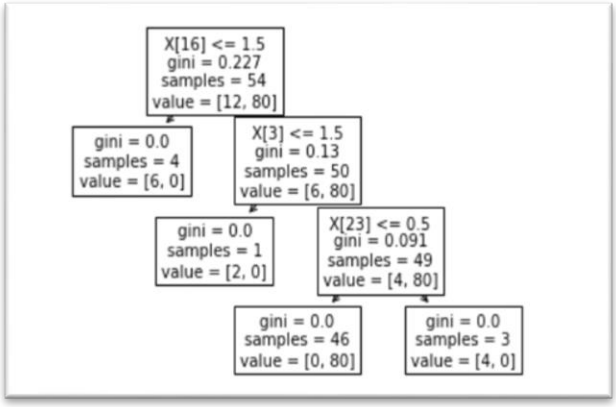


Fig. 6. Tree-like representation for Random Forest.

3) KNN classifier

K-Nearest Neighbor is used for classifying data based on the threshold value specified. The number of k-nearest values is computed using the Euclidean distance. Euclidean distance value is computed as follows:

$$\text{Euclidean Distance} = \sqrt{\sum(x_i - y_i)^2}$$

where x_i and y_i are the Euclidian vectors

4) Logistic regression

Logistic regression uses probabilistic estimation for classification. The logistic regression function $p(x)$ is the sigmoid function of

$$f(x)\colon p(x) = 1 / (1 + \exp(-f(x)))$$

This function p(x) is used to predict the probability that a given x is either closer to 0 or 1.

V. Results and Discussion

A. Dataset Analysis

The main objective is to track and monitor monthly data for the behavioural analysis of individuals. So, for the performance evaluation and model prediction, we have carried out a study using machine learning algorithms, on the dataset collected.

B. Machine Learning Algorithm Analysis

A confusion matrix was built to visualize how well this strategy performed. Confusion matrix of Decision Tree, Random Forest, KNN classifier and Logistic Regression on test data is shown in Fig. 7.

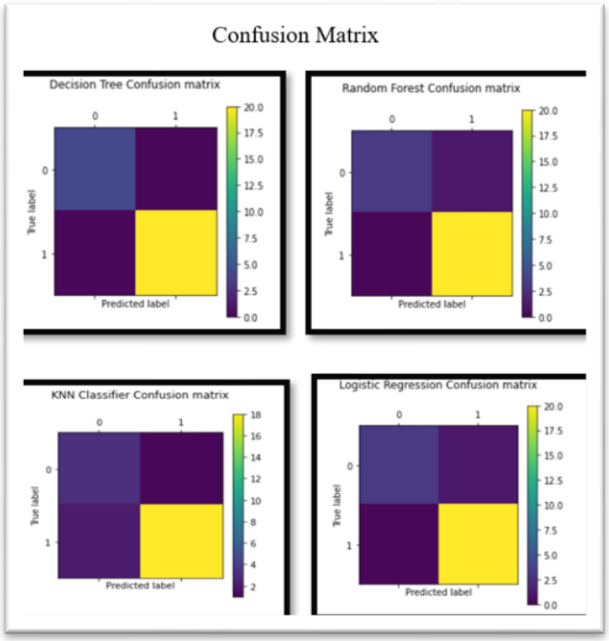


Fig. 7. Confusion matrix.

1) Text representation analysis

In this research, the functional level value is used to compute the output. The output can be visualized using text representation and tree-like structure, which is shown in Fig. 8.

```
|--- feature_23 <  0.50
|   |--- class: 1
|--- feature_23 >= 0.50
|   |--- class: 0
```

Fig. 8. Text-representation for decision tree.

The feature taken as a splitting attribute is "functional level". The values less than or equal to 0.5 are classified into class 1(no: functional level is not high) and the values greater than 0.5 are classified into class 0 (functional level is high).

2) Decision tree analysis

From the text representation analysis, it is observed that the splitting attribute is "functional level". In the decision tree shown in Fig. 9, 76 samples are classified into class 0 (developmental capability is high) and 16 samples are classified into class 1 (developmental capability is not high).

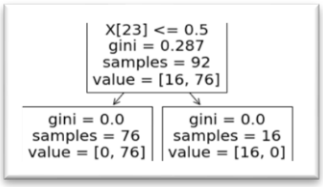


Fig. 9. Decision tree analysis.

3) Random forest analysis

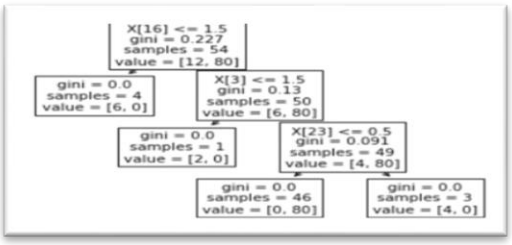


Fig. 10. Random forest analysis.

As shown in Fig. 10, the main splitting attribute is "Putting Shapes". Out of which 4 samples are classified (developmental capability is not high). These 4 samples have a rating less than or equal to 1.5. 50 samples are remaining. For this 50 sample, the splitting attribute is "Walking on Balance Beam". One sample has to have a rating of less than or equal to 1.5 and is classified into class 0. Class 0 indicates that the developmental capability is not high. For the rest of the 49 samples the splitting attribute is "functional level_high".46 samples having a value greater than 0.5 are classified into class:0 (developmental capability is high) and the 3 remaining samples are classified into class:1 (developmental capability is not high).

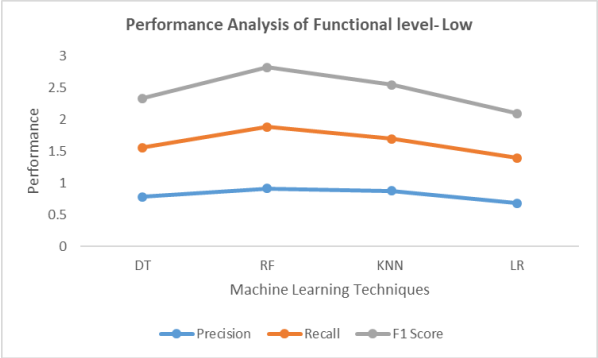
4) Performance analysis

The performance of the machine learning algorithms is evaluated using various parameters. These parameters are the accuracy of each algorithm, precision, recall and F1 score. The result of the evaluation parameters for each algorithm is tabulated in Table 6. The accuracy of DT, RF, KNN and LR is 78.15%, 95.38%, 84.54% and 72.63% respectively. The RF algorithm produces the most reliable prediction whereas the LR algorithm produces the least reliable prediction. The functional level performance analysis of different models is depicted in Fig. 11.

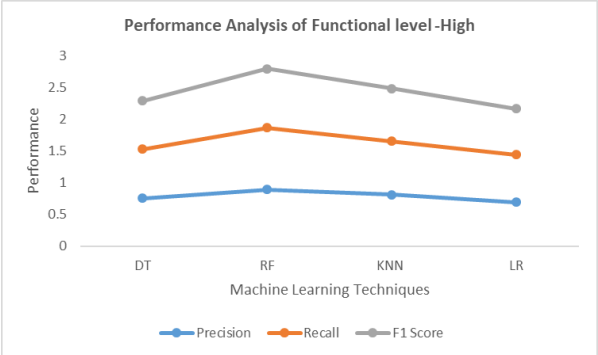
This model acts as a litmus test to further aid the initial investigation process towards the formulation of strategic plan for impactful occupational therapy. The findings highlight the strong relation between comorbidities of learning disorder and developmental capability. Our experimental study will give a rudimentary approach to carry forward their behavioural analysis and learning disorder without any further delay.

Table 6. Evaluation parameter result

Algorithm	Functional Level	Precision	Recall	F1 Score	Accuracy
DT	low	0.7831	0.7712	0.7771	0.7815
	high	0.7543	0.7747	0.7644	
RF	low	0.9151	0.9645	0.9392	0.9538
	high	0.8918	0.9731	0.9307	
KNN	low	0.8733	0.8233	0.8476	0.8454
	high	0.8144	0.8415	0.8277	
LR	low	0.6828	0.7117	0.6970	0.7263
	high	0.6947	0.7505	0.7215	



(a)



(b)

Fig. 11. Functional level performance analysis of different models: (a) Functional level value- Low; (b) Functional level value- High.

VI. Conclusion

The developmental capability prediction model aims to achieve accurate predictions about whether an individual has a high developmental capability or not. In this research, we have used and explored DT, RF, KNN, and LR classification algorithms for the developmental capability prediction of individuals. In our study, we have collected a primary dataset from occupational therapist, based on disability traits in toddlers, kids, and teens. To identify the disability traits, five developmental areas were identified and for each developmental area, five activities were recognized. The results of our research show that the RF algorithm produces the most reliable prediction, which is 95.38% whereas the DT algorithm produces the least reliable prediction, which is 72.63%. The main goal of this research was to advance the existing research by creating the model in a novel and creative method and to make the approach practical and simple to apply to real-world scenarios. This prediction is the precursor for the identification of the individual's disability and timely addressing of their learning disorder. We would like to extend our study using deep learning algorithms with more datasets for improving the performance of machine learning algorithm and further pursue the research on the learning disorders like dyslexia, dyscalculia and dysgraphia.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

GB and VA conceptualized and conducted the research, NP and MD analyzed the data, PC conducted the experiments and wrote the manuscript, SV revised the discussion section and SP proofread the paper. All authors had approved the final version.

References

[1] R. Surendiran, M. Thangamani, C. Narmatha, and M. Iswarya, "Effective autism spectrum disorder prediction to improve the clinical traits using machine learning techniques," *International Journal of Engineering Trends and Technology (IJETT)*, 2022, ISSN, 2231-5381.

[2] O. Khursheed, S. Gupta, and S. Sarkar, "Prevalence of malocclusion among 7-14 years old specially abled children attending various special schools in Mathura district, India," *Journal of Advanced Medical and Dental Sciences Research*, vol. 5, no. 4, April 2017

[3] W. Liu, X. Yu, B. Raj, L. Yi, X. Zou, and M. Li, "Efficient autism spectrum disorder prediction with eye movement: A machine learning framework," *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 649–655, 2015

[4] B. Scassellati, "Quantitative metrics of social response for autism diagnosis. In ROMAN 2005," *IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 585–590

[5] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A machine learning approach to predict autism spectrum disorder". *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6, IEEE, 2019.

[6] C.-O. M. Susana, P. P. Marrugo, and J. C. R. Ribón, "E-learning ecosystems for people with autism spectrum disorder: A systematic review," *IEEE Access*, 2023.

[7] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada and R. Mitkov, "Detecting High-Functioning Autism in Adults Using Eye Tracking and Machine Learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 6, pp. 1254–1261, June 2020

[8] Z. Zhao *et al*., "Applying machine learning to identify autism with restricted kinematic features," *IEEE Access*, vol. 7, pp. 157614–157622, 2019

[9] O. Altay and M. Ulas, "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children," in *Proc. 6th International Symposium on Digital Forensic and Security (ISDFS)*, pp. 1–4, IEEE, 2018

[10] A. Garg, A. Parashar, D. Barman, S. Jain, D. Singhal, M.Masud, and M. Abouhawwash, "Autism spectrum disorder prediction by an explainable deep learning approach," *Computers, Materials & Continua*,vol. 71, no. 1, pp.1459–1471, 2022

[11] I. P. Gowramma, E. Gangmei, and L. Behera "Research in education of children with disabilities," *Indian Educational Review*, vol. 56, no. 2, July 2018.

[12] U. B. Mahadevaswamy and C. Manjunath, "f-MRI based detection of autism using CNN algorithm," *IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 2022, pp. 1–5

[13] K. Rakhee, D. Panwar, and V. Singh, "Autism spectrum disorder study in a clinical sample using Autism Spectrum Quotient (AQ)-10 tools," in *Proc. Third International Conference on Sustainable Computing: SUSCOM* 2021, Springer Singapore, 2022.

[14] R. Suman and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, Jan. 2020

[15] S. Ghafghazi, A. Carnett, L. Neely, A. Das, and P. Rad, "AI-augmented behavior analysis for children with developmental disabilities: Building toward precision treatment," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 7, no. 4, pp. 4–12, Oct 2021

[16] S. M. Mahedy Hasan, M. P. Uddin, M. A. Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," *IEEE Access*, vol. 11, pp. 15038–15057, 2023

[17] A. S. Albahri *et al*., "Early automated prediction model for the diagnosis and detection of children with autism spectrum disorders based on effective sociodemographic and family characteristic features," *Neural Computing and Applications*, vol. 35, no. 1, pp. 921–947, 2023

[18] T. Akter, M. S. Satu, M. I. Khan, M. H. Ali, S. Uddin, P. Lio, and M. A. Moni, "Machine learning-based models for early stage detection of autism spectrum disorders," *IEEE Access*, vol. 7, pp. 166509–166527, Nov. 2019

[19] J. S. Gracia, M. M. B. A. Sulaiman, and B. Bennet, "Feature signature discovery for autism detection: An automated machine learning based feature ranking framework," *Computational Intelligence and Neuroscience*, Jan. 2023

[20] G. Shaurya, M. Chugh, and S. Vyas, "Understanding immersive technologies for autism detection: A study," *Automation and Computation*, pp. 364–370, 2023

[21] M. Mayank and U. C. Pati, "A classification framework for Autism Spectrum Disorder detection using sMRI: Optimizer based ensemble of deep convolution neural network with on-the-fly data augmentation," *Biomedical Signal Processing and Control*, vol. 84, 104686, 2023.

[22] A. Carrie, B. Auyeung, and B.-C. Simon, "Toward brief red flags for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 51, no. 2, 2012, pp. 202–212.

[23] R. Diana *et al*., "The modified checklist for Autism in toddlers: An initial study investigating the early detection of Autism and pervasive developmental disorders," *Journal of Autism and Developmental Disorders*, 2001, pp. 131–144.

Chapter 7
Analysis on Prediction of Crop Diseases Using TensorFlow with Keras and OpenCV Technique of Deep Learning

Jyoti Kharade, Pratibha Deshmukh, Gunjan Behl, Nidhi, Raje Fardin Rauf, and Kriti Gupta

Abstract Crop disease is a harmful deviation from the normal growth of plant which affects its appearance, function, or productivity and causes great damage in agriculture resulting in significant yield losses. Prediction of crop diseases can be important for agriculture field, as it can help farmers take preventive measures to protect crops and ensure a good yield. There are various methods that can be used to predict crop diseases, such as examining the plant for symptoms, using weather data to identify conditions that are conducive to the development of diseases, and implementing early warning systems that can alert farmers to the presence of potential diseases. By predicting and preventing crop diseases, farmers can improve the health and productivity of their crops, leading to better yields and increased profitability. This research paper proposes a model whose purpose is to classify the type of disease. Deep learning techniques are used in the context of categorizing plant diseases. Dataset of diseased and healthy crops is used. Leaf images of diseased crop are infused as training set in a deep learning model. TensorFlow with keras and OpenCV are used for prediction of crop disease. OpenCV is used to detect the pattern in the crop leaf image and translate it into data on which ML model can be build using keras. Firebase is used to efficiently manage the large number of images. The trained model achieved the accuracy of 97.80%. Model predicts the disease name, possible causes, and the solution.

Keywords Deep learning · OpenCV · Keras · TensorFlow · CNN · Crop disease prediction · Leaf image

J. Kharade (✉) · P. Deshmukh · G. Behl · Nidhi · R. F. Rauf · K. Gupta
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai, Maharashtra, India
e-mail: jyoti.kharade@bharatividyapeeth.edu

D. Mishra et al. (eds.), *Data Science and Big Data Analytics*, Data-Intensive Research,
https://doi.org/10.1007/978-981-99-9179-2_7

1 Introduction

Variety of reasons leads to decrease in agricultural productivity, but damage caused by pests and pathogens plays a significant role in crop losses throughout the world. The losses in crop yield due to pathogen infections range between 20 and 40% [1]. Crop losses due to pests and diseases are a major threat to incomes of rural families and to food security worldwide [2, 3], especially in the developing countries that depend on single crop or few crops.

Precision agriculture uses latest technologies for the decision-making process [4]. Various digital technologies are used for collecting large amount of data in real time, and various machine learning (ML) algorithms are used to provide optimal decisions, which had led to a minimization in costs.

Various algorithms and methods such as linear regression, logistic regression, random forest, clustering, decision trees, Naive Bayes, K-nearest neighbors (KNN), and support vector machines (SVM) among others are used for this purpose. Deep learning (DL) methods are also used in the agriculture. Computer vision and artificial intelligence can lead to new solutions. Various types of deep neural networks (DNNs) have achieved remarkable results in hyperspectral analysis [5]. Convolution neural networks (CNNs) have performed well in crop classification tasks [6], fruit counting, yield prediction [7], disease detection [8–10], and vision tasks in general [10, 11]. Additionally, it has been shown that better results are acquired if networks are pre-trained [10].

2 Literature Review

Classification will get good results depending on accurate extraction features and the study focuses on feature extraction [12]. The traditional methods require inspections of the symptoms, which is something best done manually. This can be impractical for larger crop fields and can be difficult for farmers. Recent improvements in remote sensing technologies can help solve these problems. Unmanned aerial vehicle [13] platforms with deep learning-based computer vision algorithms allow for early detection and identification of crop diseases which means increased food production.

As a result of remote sensing, the cotton root was deteriorating and was in need of site specific antibiotics. Solution was to inject a site specific fungicide to control the disease [14]. Through the survey [15] provides insight into the identification of crop pathogens and pests, also presented the general practical applications.

Diagnosing plant diseases, botanists require assistance from a variety of sources. Even they are reaching out to AI assistants [16]. As a result, a small-time farmer might not be able to detect the condition, and as a result may use excessive or incorrect pesticides. In order to prevent the crop from going waste, expert guidance can be seek to precisely diagnose the condition.

In [17], authors developed the Internet of Agro Things, which was based on a trained convolution neural network model to perform an analysis of the crop image captured by a health maintenance system.

Poor harvest conditions in rice cultivating areas have been caused by the rice blast disease, which is especially worsened by the surrounding environmental factors. In [18], authors set up the sensors in the field to collect data like temperature, humidity, pressure, and rainfall, and the data was used for spore germination model.

Due to the rapid growth in diseases and the farmer's limited knowledge of it, treatment for each disease becomes a major challenge. The leaves visually have similar attributes which allow for disease identification. To solve this, computer vision with deep learning is used [19].

When different types of data together is used, like combining spectral and spatial information [20] or visible and infrared images, it is often an effective way to improve crop disease identification. Many different types of cameras were implemented in agriculture, including RGB, its thermal substitute and other multispectral variants. The best technologies allow for identification of diseases from leaves level even in the visible spectrum [21].

The convolution neural network (CNN) [22, 23] is one of the most efficient deep learning models which produce state of the art results in many applications including agriculture. In [21], authors used the LeNet-5 [22] architecture to identify infected areas in grapevines from aerial images taken with a UAV, which had an RGB camera sensor.

CNNs have become the best choice for agricultural visual applications. Used in weed detection, crop classification [23], biotic, and abiotic stress monitoring. They are popular because of their impressive ability to extract the most relevant information from images. The other classifiers used were logistic regression, K-nearest neighbor, and support vector machine with accuracies of 66, 55, and 53% [24].

3 Research Methodology

The proposed method focuses on several important stages of developing a crop disease prediction model as shown in Fig. 1.

3.1 Dataset

Dataset of diseased and healthy crops is required. This dataset should include detailed information about the crops, such as their species, age, location, and symptoms. The dataset can be collected through field observations, expert annotations, or remote sensing techniques.

Once the dataset is collected, it must be pre-processed and cleaned to remove any irrelevant or missing data. The next step is to split the dataset into training and testing

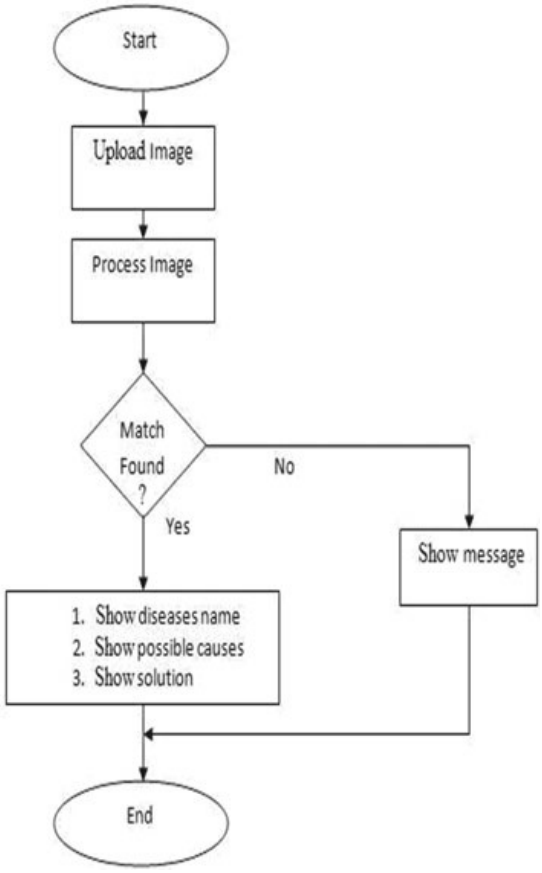


Fig. 1 Stages in crop disease prediction model

sets. The training set is used to train a machine learning model, while the testing set is used to evaluate the performance of the model.

3.2 Upload Image

User uploads the leaf image of the infected plant in proper file format. The website also checks whether the image is clear or not.

3.3 Process Image

The OpenCV library is used to process the image. Model needs the matrix form of data to read the image. So, all the pre-processing like reshaping the image, resizing the image, converting image to matrix form is done.

Fig. 2 Home page

3.4 Prediction the Disease

In this phase, model actually starts working on prediction of the disease. If model is successful in prediction of the disease, it will return the disease name, possible causes, and the solution. If model failed to identify disease, then it simply returns a message to user.

3.5 Dashboard of the Website

Home page of website is shown in Fig. 2. In the second screen, the user gets the access to upload the leaf image. All the validation on images are implemented in upload button. After successful upload of image they get store in firebase database for further processing as shown in Fig. 3.

4 Discussion on Findings

Labeled data is used in Crop Disease Prediction model preparation. Number of images in each image class are kept same; it prevents overfitting. Totally, 70,243 images were used to prepare the model and 17,557 images for testing purpose.

Figure 4 has the details of the submodule which needs to be imported to create the model. OpenCV dependency is required to check the valid image format in the dataset. TensorFlow keras dependency is required for creation of model. The NumPy dependency for handling the matrix from of images and more dependency as par model require (Fig. 5).

Fig. 3 Upload leaf image and store

```
import tensorflow as tf
from tensorflow import keras
import numpy as np
import pickle
import cv2
from os import listdir
from sklearn.preprocessing import LabelBinarizer
import matplotlib.pyplot as plt
from tensorflow.keras.preprocessing.image import img_to_array
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Conv2D, Flatten, Dropo
from tensorflow.keras.layers import MaxPooling2D, BatchNormalizat
from tensorflow.keras import backend as K
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.models import load_model
```

Fig. 4 Details of the submodule for model

```
valid_image_list, valid_image_label= [], []
for disease_folder in valid_folder:
    print(f"processing {disease_folder} ...")
    disease_img_folder= listdir(f"{valid_dir}/{disease_folder}")

    for disease_img in disease_img_folder:
        image_directory = f"{valid_dir}/{disease_folder}/{disease_img}"
        if image_directory.endswith(".jpg") == True or image_directory.endswith(".J
            valid_image_list.append(convert_image_to_array(image_directory))
            valid_image_label.append(disease_folder)
print("[INFO] Image loading completed")
```

Fig. 5 Code for data filtration

```
EPOCHS = 250
INIT_LR = 1e-3
BS = 32
default_image_size = tuple((256, 256))
image_size = 0
width=256
height=256
depth=3
train_dir=r"dataset\train"
valid_dir=r"dataset\valid"
train_folder=listdir(train_dir)
valid_folder=listdir(valid_dir)
```

Fig. 6 Parameters of model

Check whether the extension of images is correct or not.

Some key parameters of model are defined in Fig. 6 such as epochs, BS, the location of dataset, depth, and width.

Figure 7 depicts the code snippet of TensorFlow model creation. TensorFlow keras sequential model is created. The Zip file of dataset is upload to google drive.

Figure 8 shows the summary of the model. The total params is 454,566 and trainable params is total params is 454,566 and trainable params is 51,686.

Figure 9 shows the accuracy of the model on testing dataset. This model achieves the accuracy of 97.80%.

Accuracy graphs are shown in Fig. 10. In the initial epochs the accuracy is less and the loss is at the top but at the ending the accuracy in on top and we achieve the low loss.

Convolutional Neural Networks (CNNs) are a type of deep learning neural network that is commonly used for image classification. They are often implemented using popular machine learning frameworks such as TensorFlow and keras. CNNs use a process called convolution to analyze images and identify important features such as edges, shapes, and patterns. These features are then used to classify the input image. CNNs are well-suited to this type of task because they are able to automatically learn and extract the most important features from the input data, making them highly effective at image classification. In addition, CNNs are able to process large amounts of data efficiently which is 3GB, making them well-suited for crop diseases prediction.

Image classification is the process of assigning a label or class to an input image based on its content. This is a common task in the field of computer vision, where the goal is to develop algorithms that can automatically identify objects, scenes, and other visual elements in images. Image classification algorithms typically use machine learning techniques, such as convolution neural networks, to learn from a training dataset of labeled images. Once trained, the algorithm can then be applied to new, unseen images and predict their class or label.

```
inputShape = (height, width, depth)
chanDim = -1
if K.image_data_format() == "channels_first":
    inputShape = (depth, height, width)
    chanDim = 1
model=Sequential([
                Conv2D(32,(3,3),padding="same",
                input_shape=inputShape,activation="relu"),
                BatchNormalization(axis=chanDim),
                MaxPooling2D(pool_size=(3,3)),
                Dropout(0.25),

                Conv2D(64,(3,3),padding="same",activation="relu"),
                BatchNormalization(axis=chanDim),
                MaxPooling2D(pool_size=(3,3)),

                Conv2D(64,(3,3),padding="same",activation="relu"),
                BatchNormalization(axis=chanDim),
                MaxPooling2D(pool_size=(3,3)),
                Dropout(0.25),

                Conv2D(128,(3,3),padding="same",activation="relu"),
                BatchNormalization(axis=chanDim),
                MaxPooling2D(pool_size=(3,3)),

                Conv2D(128,(3,3),padding="same",activation="relu"),
                BatchNormalization(axis=chanDim),
                MaxPooling2D(pool_size=(3,3)),
                Dropout(0.25),

                Flatten(),
                Dense(1024, activation="relu"),
                BatchNormalization(),
                Dropout(0.5),

                Dense(n_classes,activation="softmax")

])
```

Fig. 7 TensorFlow keras sequential model

TensorFlow and keras are two popular open-source libraries for machine learning and deep learning. TensorFlow is a low-level library that provides a flexible platform for building and training machine learning models. It provides a wide range of tools and algorithms for building, training, and deploying machine learning models, including support for neural networks and deep learning. Keras, on the other hand, is a high-level library that provides a simple, user-friendly interface for building and training machine learning models. It is built on top of TensorFlow and is designed to make it easy to quickly build and train deep learning models. Together, TensorFlow and keras provide powerful tools for implementing and training machine learning model for crop diseases prediction. Then throw api call the model which gets the image, after pridiction the model returns the result to the user as depicted in Fig. 3.

Fig. 8 Summary of the model

Output exceeds the size limit. Open the full output data in a text editor

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 256, 256, 32)	896
batch_normalization (BatchNo	(None, 256, 256, 32)	128
max_pooling2d (MaxPooling2D)	(None, 85, 85, 32)	0
dropout (Dropout)	(None, 85, 85, 32)	0
conv2d_1 (Conv2D)	(None, 85, 85, 64)	18496
batch_normalization_1 (Batch	(None, 85, 85, 64)	256
max_pooling2d_1 (MaxPooling2	(None, 28, 28, 64)	0
conv2d_2 (Conv2D)	(None, 28, 28, 64)	36928
batch_normalization_2 (Batch	(None, 28, 28, 64)	256
max_pooling2d_2 (MaxPooling2	(None, 9, 9, 64)	0
dropout_1 (Dropout)	(None, 9, 9, 64)	0

...

Total params: 454,566

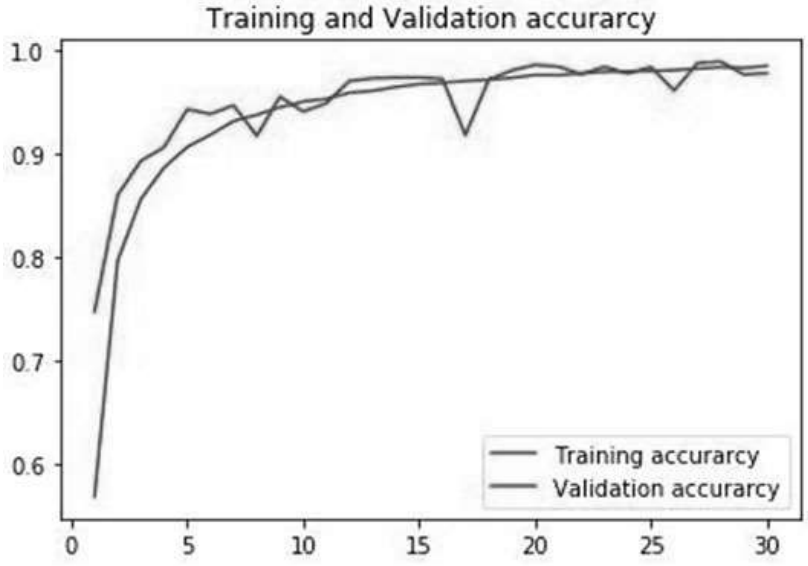
Trainable params: 451,686

Non-trainable params: 2,880

```
print("[INFO] Calculating model accuracy")
scores = model.evaluate(np_valid_image_list, bin_valid_image_labels)
print(f"Test Accuracy: {scores[1]*100}")
```

```
[INFO] Calculating model accuracy
17572/1
[==============================================================================
- 10s 557us/sample - loss: 0.1140 - accuracy: 0.9780
Test Accuracy: 97.80332446098328
```

Fig. 9 Accuracy of the model



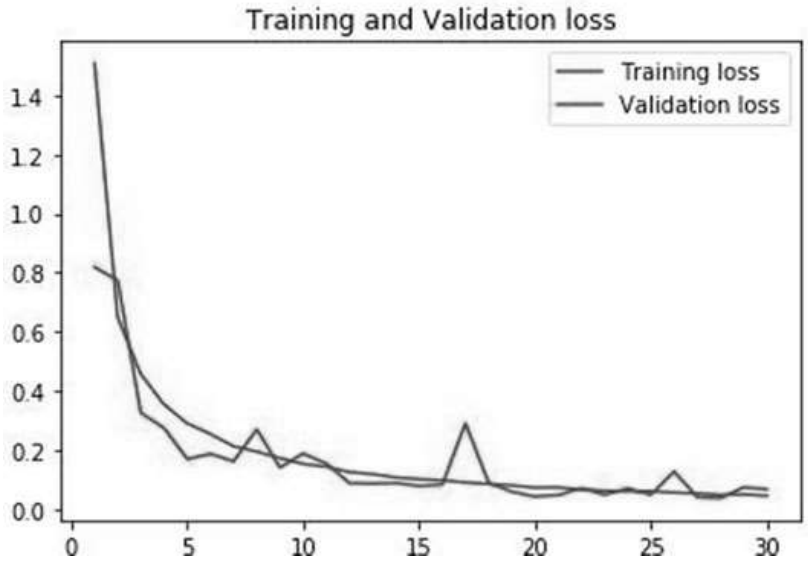


Fig. 10 Accuracy graphs

The main objective of predicting crop diseases is to identify potential diseases in crops and take preventive measures to protect the crops from these diseases. By predicting diseases in crops, farmers can take timely action to treat the crops and prevent the spread of the disease to other plants. This can help to reduce the overall impact of the disease on the crop and ultimately help to improve the productivity and profitability of the farm.

5 Conclusion and Recommendations

After examining the findings, there is a much better understanding of how this deep learning technique works and how it can be used in the context of categorizing plant diseases. The objective now is to classify which type of disease we have. We've used technologies like keras, OpenCV, and TensorFlow with the machine learning language Python. And to efficiently manage the large number of images, we're going to use Firebase to store them. When performing leaf disease prediction, these will be imported from the local folder and saved at firebase.

References

1. Savary S, Ficke A, Aubertot J, Hollier C (2012) Crop losses due to diseases and their implications for global food production losses and food security. Food Secur 4:519–537. https://doi.org/10.1007/s12571-012-0200-5
2. Savary S, Willocquet L (2014) Simulation modeling in botanical epidemiology and crop loss analysis. In: The plant health instructor, 173 p
3. Avelino J, Cristancho M, Georgiou S, Imbach P, Aguilar L, Bornemann G et al (2015) The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions. Food Secur 7(2):303–321
4. Gebbers R, Adamchuk VI (2010) Precision agriculture and food security. Science 327:828–831
5. Gewali UB, Monteiro ST, Saber E (2018) Machine learning based hyperspectral image analysis: a survey, pp 1–42. arXiv, arXiv:1802.08701
6. Yao C, Zhang Y, Zhang Y, Liu H (2017) Application of convolutional neural network in classification of high resolution agricultural remote sensing images. Int Arch Photogramm Remote Sens Spat Inf Sci XLII-2/W7:989–992
7. Rahnemoonfar M, Sheppard C (2017) Deep count: fruit counting based on deep simulated learning. Sensors 17:905
8. Liu B, Zhang Y, He D, Li Y (2018) Identification of apple leaf diseases based on deep convolutional neural networks. Symmetry 10:11
9. Ferentinos KP (2018) Deep learning models for plant disease detection and diagnosis. Comput Electron Agric 145:311–318
10. Lee H, Kwon H (2017) Going deeper with contextual CNN for hyperspectral image classification. IEEE Trans Image Process 26:4843–4855
11. Steen K, Christiansen P, Karstoft H, Jørgensen R (2016) Using deep learning to challenge safety standard for highly autonomous machines in agriculture. J Imag 2:6
12. Suttapakti U, Bunpeng A (2019) Potato leaf disease classification based on distinct color and texture feature extraction. In: Proceedings of 2019 19th international symposium on communications and information technologies (ISCIT), no. Mcd, pp 82–85

13. Bouguettaya A, Zarzour H, Kechida A, Mohammed Taberkit A (2021) Recent advances on UAV and deep learning for early crop diseases identification: a short review. In: 2021 International conference on information technology (ICIT), pp 334–339. https://doi.org/10.1109/ICIT52682.2021.9491661
14. Yang C (2020) Remote sensing and precision agriculture technologies for crop disease detection and management with a practical application example‖. Engineering 6(5):528–532
15. Liu J et al (2021) Plant diseases and pests detection based on deep learning: a review. Springer Nature
16. Gupta R, Sharma A, Gupta S, Garg M, Kaur G (2022) Automatic identification of paddy crop diseases using deep learning approach. In: 2022 3rd international conference on electronics and sustainable communication systems (ICESC), pp 915–920. https://doi.org/10.1109/ICESC54411.2022.9885537
17. Udutalapally V, Mohanty SP, Pallagani V, Khandelwal V (2021) SCrop: a novel device for sustainable automatic disease prediction crop selection and irrigation in Internet-of-Agro-Things for smart agriculture. IEEE Sensors J 21(16):17525–17538
18. Chen W-L, Lin Y-B, Ng F-L, Liu C-Y, Lin Y-W (2020) RiceTalk: rice blast detection using Internet of Things and artificial intelligence technologies. IEEE Internet Things J 7(2):1001–1010
19. Kulkarni O (2018)Crop disease detection using deep learning. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA), pp 1–4. https://doi.org/10.1109/ICCUBEA.2018.8697390
20. Zhang X et al (2019) A deep learning-based approach for automated yellow rust disease detection from high-resolution hyperspectral UAV images. Remote Sens 11(13)
21. Kerkech M, Hafiane A, Canals R (2018) Deep leaning approach with colorimetric spaces and vegetation indices for vine diseases detection in UAV images. Comput Electron Agric 155:237–243
22. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Biochem Biophys Res Commun 86(11):2278–2324
23. Yang S, Gu L, Li X, Jiang T, Ren R (2020) Crop classification method based on optimal feature selection and hybrid CNN-RF networks for multi-temporal remote sensing imagery. Remote Sens 12(19):1–23
24. Sharma P, Hans P, Gupta SC (2020) Classification of plant leaf diseases using machine learning and image pre-processing techniques. In: 2020 10th international conference on cloud computing data science & engineering (confluence), pp 480–484

THE COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS MULTIPLE REGRESSION, XG BOOST AND SVM WITH RESPECT TO RESIDENTIAL ASSET PRICE

Mrs Nidhi, Assistant Professor
Bharati Vidyapeeth's Institute of Management and Information Technology,
Navi Mumbai (India)
mca.nidhipoonia@gmail.com

Saurabh Dnyaneshwar Kathe, Student,
Bharati Vidyapeeth's Institute of Management and Information Technology,
Navi Mumbai (India)

Swapnil Sunil Patil, Student,
Bharati Vidyapeeth's Institute of Management and Information Technology,
Navi Mumbai (India)

ABSTRACT
Tourism is a vital cog in the growth of any country as it is increasing day by day in countries like India. Due to this hotel and Residential asset, stays demand is also increasing rapidly and the best affordable price decision-making for owners plays a very important role. Price prediction and analysis of residential assets are important topics to research in the Indian economy. Existing research papers mostly focus on macroeconomics affecting the prices of residential assets. Here we will focus on micro factors and detailed information about the asset. It can be helpful for an organization in two ways, on one side, it enables space owners to list their space and earn rental money and on the other side, it helps tourists for accessing rented private homes. This paper helps with the increasing competition which reduces prices for customers with better services. It promotes tourism in a region. The prices are predicted based on various factors such as location, neighbourhood, etc. We used various algorithms (multiple regression, XG boost, support vector machine) to predict prices and compare the best one according to error rate.
Keywords: Residential asset price prediction, multiple regression, XG boost, support vector machine modal

Introduction

Tourism is the backbone for the health of any economy, it boosts job creation, foreign currency earnings, infra development, culture and regional development. Technological innovations are improving every sphere of human life, tourism is no exception. To make optimum utilization of resources technology has an important role to play. Technological advancements are upgrading the experience for tourists as well as hotel owners. As tourism is increasing in India, many organizations have promoted tourism in such a way that people can easily rent their residential assets online which can be rented by tourists. These are cost-efficient and even a middle-class family can afford such apartments for a night. In the past, people relied on simple data analysis for calculating the budget but nowadays data science has unlocked the potential for studying complex businesses. Machine learning algorithms work to filter out data or live-streamed which helped to increase the result for the decision-making process. Data collection is very important as we rely on the data for more accurate predictions. Various techniques can be applied to the vast chunk of data to churn own the best possibilities for all stakeholders and parties. One could use traditional regression algorithms or even the latest machine learning models which are better at drawing predictions. Various factors affect the housing prices like the neighbourhood, location, ratings, nearby destinations to visit, cost per night, minimum night's availability, and the number of reviews. The other important factors include the characteristics of the house like the room type (private, public), facilities available, etc. For better decisions and predictions and deciding prices for these assets, we should use different machine learning algorithms. Here we will apply different approaches and try to find out which ML algorithm is best suitable for predicting the residential prices of a property for one day or more days for renting purposes. In our research, we will use multiple regression, XG boost and support vector machine modal and test which is giving the best result for predicting their prices by finding the error rate. For our research first will review the literature on residential prices in tourism and machine learning algorithms then will do a visualisation of data collected from secondary sources according to features and properties of data and after that will apply ML algorithms for prediction and analysis according to error rate. In the end, will compare the error rate and draw results which is a suitable algorithm for studying the research problem of residential asset prices data. In the last section, will draw a conclusion according to the result analysed.

Literature Review

Luo (2019) residential prices are analyzed and predicted according to the micro area by using SVM and random forest modal and prices are predicted according to the pool, area, etc. not by using traditional methods. Studied residential assets and understood how residential asset prices vary from region to region. Studied regression algorithm and how it can be implemented.

Kalehbasti et al. (2019) prices of rental property prediction using a model using deep learning, machine learning modal support-vector regression (SVR), neural networks (NNs) and others. The paper comparison between different algorithms of machine learning and is analysed for property prices of rental using different models.

Alfiyatin et al. (2017) predict housing values for the future according to the concept, physical conditions and location. And predict error for proved combination regression. Shehhi et al. (2020) predict hotel prices by four modal SARIMA, adaptive network fuzzy interference SVM machine model and Boltzmann machine for GCC cities.

Shamim (2022) Machine learning algorithm are used to predict stock prices with lower error rates paper aim to use various techniques for prediction and analyse which algorithm is best for stock price prediction.

Tziridis et al. (2017) predict airfare charges compared modals on eight states' air flight data and decide the prices according to factors that affect prices. In this research paper, we get a better understanding of the factors affecting the prices of plane tickets and how those factors can help in predicting the prices.

Erguven et al. (2012) Evaluation of the effectiveness of primary multiple linear regression in addition to constituent analysis for the Education and Science Ministry of Georgia.

He et al. (2005) support vector machines (SVMs) inverse problem is investigated and the inverse problem is divided into two clusters such that the margin between the two clusters for a given dataset.

Ogunleye et al. (2019) where modal XGBoost for the subject of high-performance chronic kidney disease the optimized and XGBoost by using its decision tree approach gives the best result using GPU.

Hu et al. (2017), Predict prices according to product review by using the frequency, regency and monetary (RFM) model and in the paper prices are predicted by using reviews given by users.

Mitchell (2017), Paper talks about the XGBoost algorithm working, it works using the graphics processing unit and forms a decision tree and its performance is very high on different datasets and its speed could be increased by using a higher version processor.

Li (2019), In the paper XGBoost gives the best prediction having the lowest error rate as compared to linear regression, D-Gex and KNN for Gene profiling and RNA-seq. It is a model of multiple trees hence it predicts gene expression perfectly for the given datasets of health problems.

Dong (2020), here in the paper talks about the prediction of electrical resistivity measurement using XGBoost algorithm and the algorithm gives a satisfactory result according to the fitting line as compared to values of RMSE.

Mo (2019), here in the paper XGboost algorithm acts best in predicting the window behaviour of a residential building which requires Heating Ventilation and Air Conditioning best performance.

Research objective

The chief objective of the paper for research is mentioned below:

1) To visualize also analyze residential price data with respect to various factors like neighbourhood, price, and location.
2) To study multiple regression approaches with respect to residential asset prices.
3) To study boosting technique (XG Boost) approach with respect to residential asset prices.
4) To study Support Vector Machine in order to predict better output and better prices on residential assets.
5) Compare the Multiple Regression, Support Vector Machine, XGBoost with respect to identification of better residential asset prices with better accuracy rate and minimum error rate.

Problem statement

As the tourism industry is growing day by day, the prices of hotels have been increasing. Due to which the middle-class people are facing renting issues when travelling. In order to understand the tourism prices and the factors affecting those prices we are analyzing the data of residential asset pricing. It can help us to understand major factors affecting the prices for example if the asset is located near a good Neighbourhood or bad neighbour or a beach etc. This research can benefit in promoting tourism as well as to set the best prices for a rented residential asset for a night.

Research Methodology

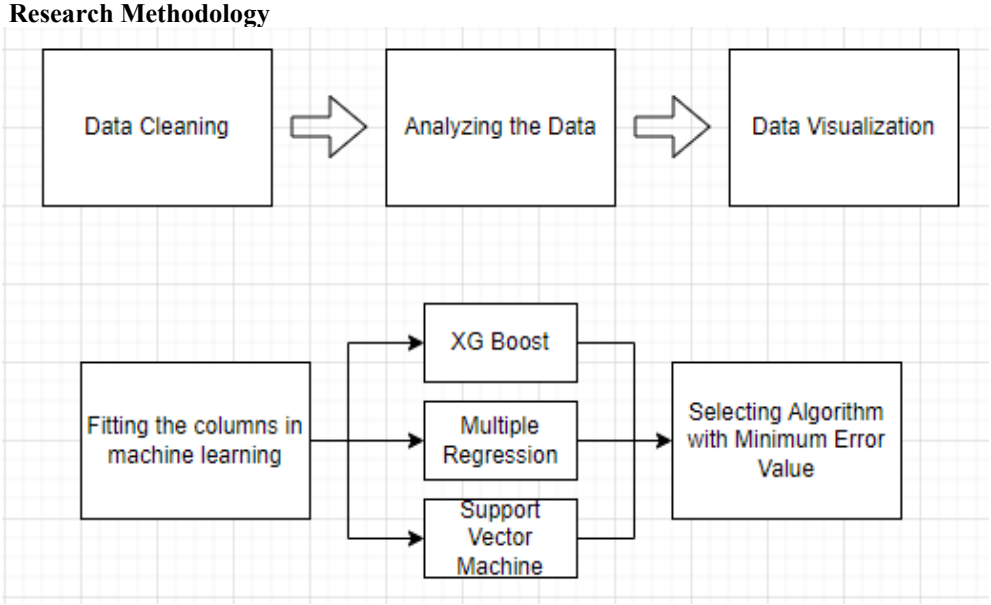


Figure 1: System Flow (source: https://ieeexplore.ieee.org/)

In Figure 1, Nguyen (2017), explains, the diagram represents a workflow of the research work. This diagram shows the separate steps of a process in sequential order. It helps others how a process is done. First, we remove unfitting, matching and inadequate data inside the dataset. After merging several data sources, there are likelihoods for data to be repeated or mislabelled and then thoroughly smearing logical and statistical techniques to define and estimate data. We can show information trends, graphically and weight patterns. It benefits the reader to attain rapid understanding. Then Multiple Regression, XG Boost and SVM models are used to train on the dataset.

The present study consists of four distinctive stages: (1) the assortment of the residential asset features that impact the prices, (2) to train and test the applied ML models on the group of adequate residential asset data (3) the choice of the regression ML models being compared and (4) investigational assessment of the ML models. Each phase of processing is discussed as follows:

Phase 1: - In this phase, the utmost enlightening structures of an asset that regulate the prices are fixed and it defines the problem which we are solving in the paper under. For all assets, the subsequent structures were measured:

Here is the list of Features we define as F1, F2 so on:
F1: Room type
F2: Ratings
F3: Nearby destinations to visit.
F4: Cost per night
F5: Facilities available
F6: Longitude
F7: Latitude
F8: Neighbourhood
F9: Neighbourhood Group
F10: Minimum nights
F11: Availability

F12: Number of reviews

Phase 2 (Collection of Data) – It is phase where we focused on the prediction of a single asset price. For trials, a set of asset data (From the Airbnp dataset from Kaggle) for every asset the features are (F1 to F12) were composed from the Web manually.

Phase 3 (ML Models Selection) – For current study ML models were selected and applied to the same data. i.e. Multiple Regression, Xgboost, and Support Vector Machine ML models

Phase 4 (Evaluation) – The residential asset data collected in phase 2, were used to train the mentioned ML models. The prediction accuracy indices are used here (error rate between the desired and predicted prices)

Data Analysis And Visualization

```
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x207902583d0>
```

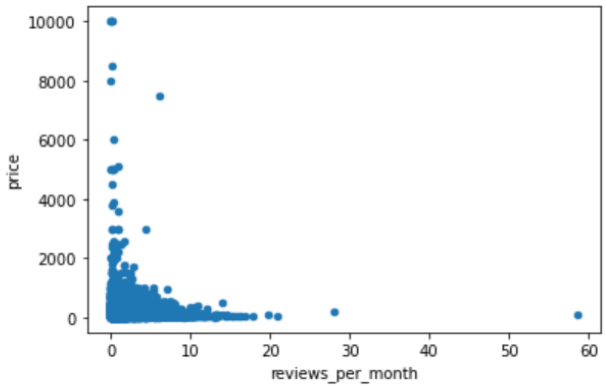


Figure 2: Price by Reviews

In Figure 2 We see that as the reviews are less the prices are more. Most tourists cannot afford costly homes. That's why they prefer affordable houses. Reviews are more where the price is less because Most of tourist thinks about why they spend a lot of money on a rental house where they only go to sleep at night.

```
<matplotlib.axes._subplots.AxesSubplot at 0x207908e5d30>
```

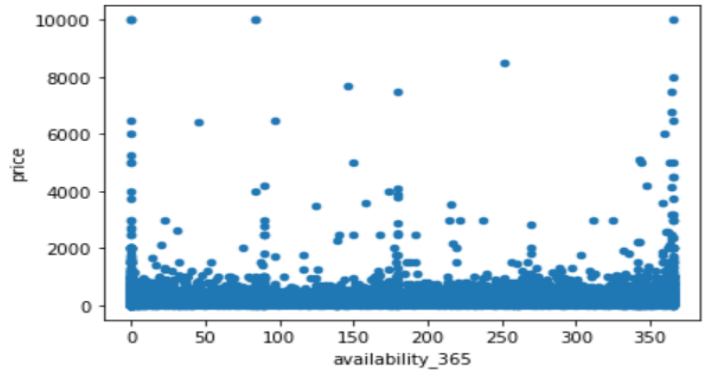


Figure 3: Price by Availability

In Figure 3, both the data columns are independent. There is no relation between price and availability but when one or more dimensions are added there might be a pattern occurring and this dimension can be added through ML models.

Dataset is a collection of data. This data helps to analyze the trends and hidden patterns and make decisions based on the dataset. These records in the dataset are organized in a way how we plan to access the information. Every column relates to a particular variable and every row relates to a given member of the data set.


```
In [68]: ml_data['price'].hist()
Out[68]: <matplotlib.axes._subplots.AxesSubplot at 0x2bb4ae26b80>
```

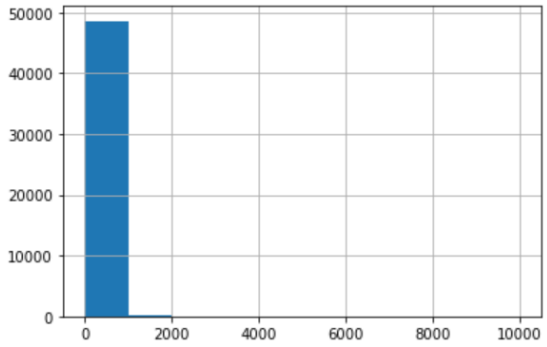


Figure 4: Price Range

Figure 4 represents the price range. This graph indicates which category has the maximum house rent.

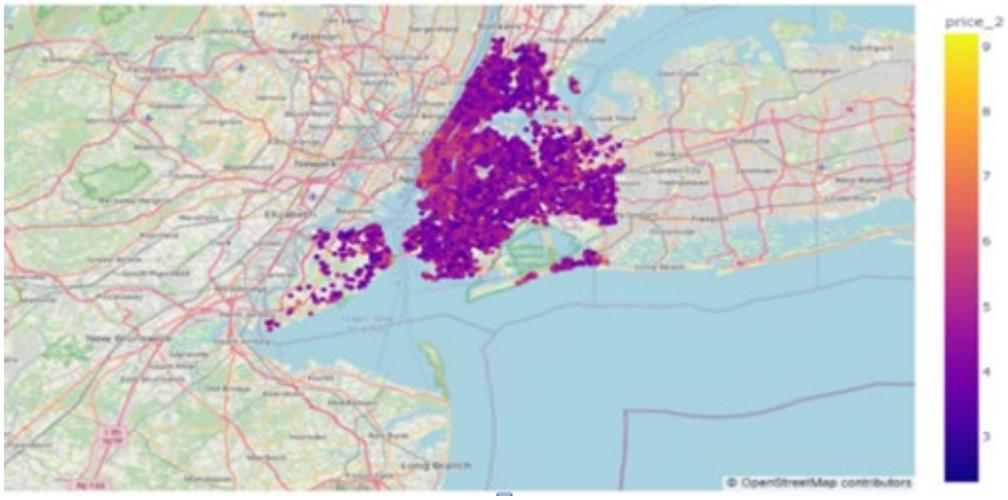


Figure: 5 Locations and Price

In Figure 5, the Visual Representation shows the location and price of homes. In this map, the dark blue colour shows the lower price of house availability and the yellow colour shows the higher price of house availability.

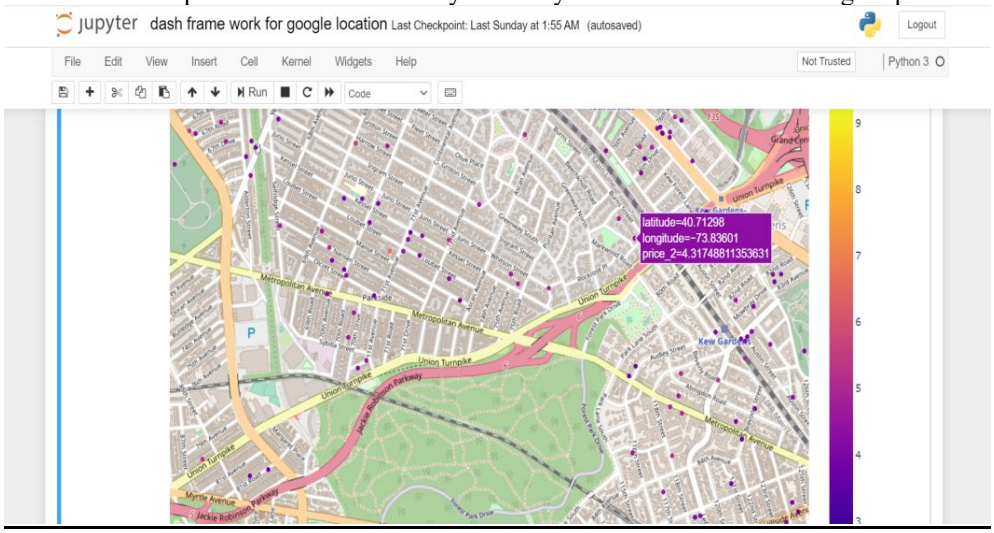


Figure 6: Location

The above Figure 6 is to get a better idea about the location which is preferable.

Result And Findings

Multiple Regressions: The relation between a single dependent variable and several independent variables can be analysed by using the technique of multiple regressions. Multiple regression analysis is a technique by using the independent variables whose values are known to predict the value of the single dependent value. The value of every predictor is weighed, where the weights denote their relative contribution to the overall prediction. Multiple regressions help to compare various columns such as price, neighbourhood, availability, reviews, etc. at the same time which draws a line in multiple dimensions which can only be done through machine learning.

XG Boost: It is a machine learning algorithm which helps to convert a weak tree into a strong tree. It makes use of boosting techniques. XG Boost is a framework that can run on multiple languages. So you can very well run XG Boost in R, Python, etc. XG Boost is very much a platform free means it is portable. So you can run on Windows, Linux, iOS, etc. XGBoost gain so much popularity because of two main things: It's the speed of processing and the kind of result or output it gives.

SVM: Supervised learning algorithm Support Vector Machine is majorly used for Classification and its aim is to create the finest decision boundary or line which can segregate space into classes of n-dimensional. The hyper plane is the top decision boundary and SVM uses a hyper plane for choosing the extreme points/vectors. The following tables give the error value of each machine learning algorithm used on the residential asset data:

Comparative analysis of Multiple regression, Support Vector Machine, XG Boost algorithms error rate with respect to identification of better residential asset prices	
Algorithm	**Error Rate**
Multiple Regression	60
XG Boost	37
SVM	26

Table 1 Comparative analysis of multiple regressions, SVM, XG Boost algorithms error rate with respect to identification of better residential asset prices

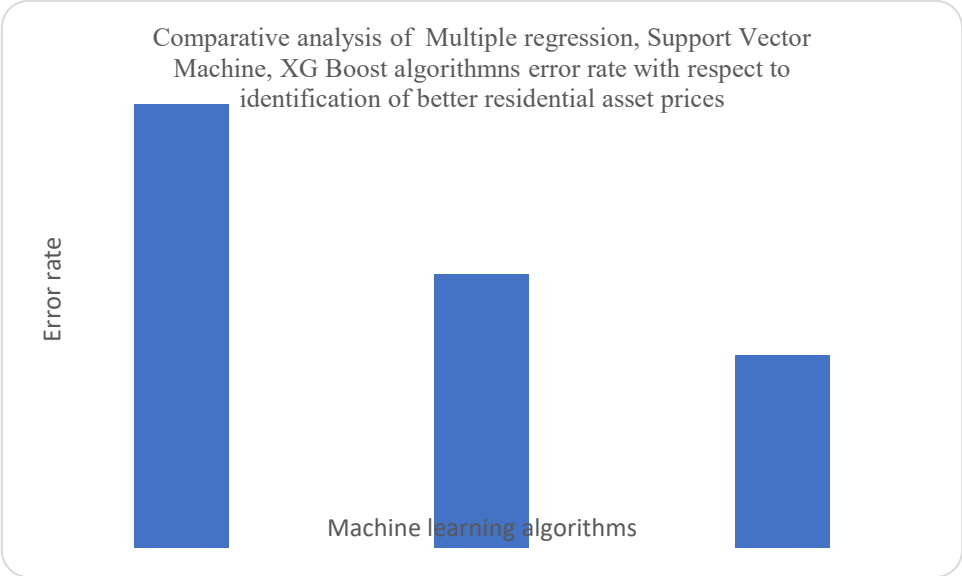


Figure 7: Error Rate Graph for Machine Learning Algorithms for Given Dataset

In Figure 7 the graph shows the error rate of machine learning algorithms regression, XG Boost and SVM using the Residential price dataset.

Conclusion:

The paper predicts charges of Residential assets using regression, XG Boost and SVM machine learning approaches. We conclude that Residential asset prices vary based on various factors like the type of rooms as well as the location. The investigational outcomes demonstrate that for predicting asset rents and prices ML models are suitable tools. The locality and the neighbourhood affect the most in-house pricing. The Comparative Analysis of Machine learning modal multiple regression, Support Vector Machine, and XG Boost algorithms error rate concerning the identification of better residential asset prices shows that SVM is the best method as the error rate is 26. This research can help Residential owners to set the house rent for tourism. From this paper, they get an idea about the range of asset rent according to its location. We conclude that this research paper can help an organization to improve tourism. This research can help organizations to predict prices and add competition between organizations. It can help the tourist to buy places at the best prices and can afford to stay near tourist spots. We hope this can be helpful in the future.

References

Alfiyatin N., Febrita R., Taufiq H., & Mahmudy W. (2017), "Modelling house price prediction using regression analysis and particle swarm optimization", Malang, East Java, Indonesia. International Journal of Advanced Computer Science and Applications, 8(10).

AlShehhi, M., & Karathanasopoulos A. (2020), "Forecasting hotel room prices in selected GCC cities using deep learning", Journal of Hospitality and Tourism Management.

DongW., Huang Y., LehaneB., & Ma, G. (2020) , "XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring", Automation in Construction, 114, 103155.

Erguven M. (2012), "Comparison of the efficiency of principal component analysis and multiple linear regressions to determine students' academic achievement", In 2012 6th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-5). IEEE.

He Q., & Chen F (2005), "The inverse problem of support vector machines and its solution", In 2005 International Conference on Machine Learning and Cybernetics (Vol. 7, pp. 4322-4327). IEEE.7)

Hu H., Chen K, & Lee J. (2017), "The effect of user-controllable filters on the prediction of online hotel reviews" Information & Management, 54(6), 728-7

Luo Y. (2019, December), "Residential Asset Pricing Prediction using Machine Learning", In 2019 International Conference on Economic Management and Model Engineering (ICEMME) (pp. 193-198). IEEE.

Li, W., Yin Y., QuanX.& Zhang, H. (2019),"Gene expression value prediction based on XGBoost algorithm", Frontiers in genetics, 10, 1077.

Mitchell, R. & Frank, E. (2017), "Accelerating the XGBoost algorithm using GPU computing", PeerJ Computer Science, 3, e127.

Mo, H., Sun, H., Liu, J., & Wei, S. (2019), "Developing window behavior models for residential buildings using XGBoost algorithm", Energy and Buildings, 205, 109564.

Nguyen& Dong(2017),"Joint network coding and machine learning for error-prone wireless broadcast", 10.1109/CCWC.2017.7868415

Kalehbasti P., Nikolenko L., & Rezaei, H. (2019), "Airbnb price prediction using machine learning and sentiment analysis", arXiv preprint arXiv:1907.12665.

Shamim R. (2022), "Machine learning's algorithm profoundly impacts predicting the share market stock's price", IJFMR-International Journal For Multidisciplinary Research, 4(5)..

Tziridis K., Kalampokas T., Papakostas, G., & Diamantaras, K. (2017, August), " Airfare prices prediction using machine learning techniques.",In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 1036-1039). IEEE.

Ogunleye A., & Wang Q. (2019)., "XGBoost model for chronic kidney disease diagnosis", IEEE/ACM transactions on computational biology and bioinformatics, 17(6), 2131-2140.44.

PREDICTIVE ANALYSIS OF FOREIGN DIRECT INVESTMENT IN INDIA USING BUSINESS INTELLIGENCE (BI) TOOL- TABLEAU

Ms. Nidhi, Assistant Professor
Bharati Vidyapeeth's Institute of Management and InformationTechnology
Navi Mumbai (India)
mca.nidhipoonia@gmail.com

Jayesh Kishor Patil, Student
Bharati Vidyapeeth's Institute of Management and InformationTechnology
Navi Mumbai (India)
jayeshkpatil11@gmail.com

Satish Shivaji Pachakar, Student
Bharati Vidyapeeth's Institute of Management and Information Technology
Navi Mumbai (India)
satish.pachakar148@gmail.com

ABSTRACT
Foreign investment is vital for every country, mostly in the financial progress of developing countries worldwide. Countries try to attract FDI inflows with their policies and it becomes a key battleground in the markets. Foreign Direct Investment (FDI) has become a crucial source of investment for many countries, including India, due to its potential to bring in capital, create jobs, transfer technology, and super economic growth. This paper highlights the trend of FDI in India in state-wise and countrywide shares of FDI. Business Intelligence (BI) Tool- Tableau is used to get analyzed data in the form of different visualizations like tables, and line graphs which are easy to understand for investors. In this paper, secondary data is being used which is collected from the official website of RBI. This data will cover FY. 2000-2022 and using this data we can predict and visualize the upcoming year 2025 inflow. This analysis will help to make decisions for the growth of Indian states.
Keywords: Indian Economy, FDI inflows Trend, FDI Prediction, Analysis of Foreign Direct Investment in India, Business Intelligence (BI) Tool- Tableau

Introduction

FDI was announced in India in 1991 by former finance minister Dr. Manmohan Singh under Foreign Exchange Management Act. The Indian government makes rules and laws on equity croft for foreign investors in different sectors and the Indian government has taken several measures to attract Foreign Investors, such as streamlining the FDI policy and eliminating sectoral capson the foreign project in numerous industries. As a result, FDI inflows into India have increased significantly in recent years. In the Study of UNCTAD, only China is ahead of India in foreign investments and because of globalization and liberalization, FDI is considered a machine for raising lucrative progress and advancement. If we observe the frugality of India, it'll be moving widely due to this India will uprising from a developing country to an advanced country.Transmogrification needs an enormous sign of wealth in the form of both fiscal and directorial and in this process, FDIremnants the most reachable and active option for fiscal capital in India. Hence we can say that Foreign Direct Investment is defined as a fiscal method of incoming capital from a country outside the boundaries of a nation and because of this capital flux which rises the creation capacity of several sectors of miserliness is nominated as Foreign Direct Investment.Planning Commission plays an important roles the government of a country must frame, apply and run the FDI plans. To a great level, the size and significance of FDI in any country can be dependent upon its macroeconomic programsIf we study the data of (FDI) in India has been gradually growing in recent years, although the COVID-19 pandemic had an impact on FDI invasions in 2020. For making better decisions FDI data is to be analyzed but for analyzing such a huge quantity of data some Business Intelligence tools are essential to help investors to understand the pattern of data which is available in the market or published by the government on the RBI website. In this paper, we are using the Business Intelligence (BI) Tool- Tableau data which will help us to understand the total FDI till the current date and we can analyze the top sectors, states, countries of inflow, etc.that have higher FDI inflow.

Literature Review

Singh (2019), had studied the various policies of FDI India and given proposals for policies also and according to her, the Indian economy is one of the topmost developing marketplaces all over the world. Kumar (2021) talks about the topmost foreign inflow places in the world, studied data from 2000-2021, and investigated the existing FDI rules and tendencies.In the paper, we get a detailed overview of the trend of FDI and its policy

framework. Ramasamy (2017), discusses the properties of FDI spillover on local efficiency and shows that the technology, human capital, and various provisions of FDI have a substantial imprint on local output. This paper research and development shows that FDI has a significant impact on regional productivity in India.

Rani (2020), learning says that the Indian economy is the fastest-emerging economy and talks about statistical tools like CAGR. It also talks about the GDP and uses the secondary data set for study in India. Merajothu (2020), surveyed the FDI impact on GDP and tested hypotheses using simple linear regression and took 19 years of data from 2000-19, where talked about India its availability of large amounts of natural resources, well market environments, and extremely proficient and knowledgeable capitals, which provide a better platform for reserves. Patil (2014), effects of FDI on Sectors of the economy of India such as software, pharmaceuticals, IT services, automobile, industry, and e-commerce which have received the prime sum of FDI in India.

Anitha (2012), talks about Foreign Direct ventures in India where there is the accessibility of large amounts of natural capital as well as better marketplace environments and highly proficient and practiced resources, which offer a better stage for reserves and Economic Growth in India. Bhargava (2018), Premeditated on the Visualization of data sets and different methods, also the process of making significant data by the visual framework and data analysis where it's executed after the data correction. Sharma (2011), explores the effectiveness of BI in strategies for an organization for employees also it explains business intelligence outcomes and categories distribution. It talks about performance management and steps to create a balanced scorecard.

Tvrdíková (2007), talks about business intelligence applications and tools in operational data. It says BI tools are necessary for data mining and they help to get information from big datasets. The main focus is on data warehouses and on new tools. Duggal study says domestic and foreign investments play an important role in the growth of India. Foreign investment can decrease the domestic savings gap. It covers a 13-year period to analyze and examine the trend and pattern of FDI. Miyamoto (2003), says that max human capital is a key ingredient to attracting FDI, and it's beneficial to host countries in the form of their activities. Rahate (2021) assured that BI tools can compress large datasets into their insights. The motto behind the created paper is to demonstrate an idea regarding data analytics. It is focused on data exploration and visualization, and model development. Implementation is done on the tableau tool.

Research Objective

The key objective of a research paper are stated below

- To study the trends of FDI inflows in India using the Tableau BI tool.
- To identify State-wise inflow of FDI in India using the Tableau BI tool.
- To detect the country-wise flow of FDI into India by using the Tableau BI tool.
- To determine the sector-wise distribution of FDI inflows in India using tableau.
- To Predict/Forecast FDI inflows in India whether increasing/ decreasing by using Tableau BI tools.

Problem Statement

As FDI data is an enormous amount of data, formerly analysis was done manually and it takes time to recognize lots of data but by using analytics tools we can easily predict future analyses or forecasts. Critical tools like tableau will be also helpful for FDI data analysis. There is one benefit: if we have lots of data, we can get more accurate predictions with the help of that data. Using the Business Intelligence (BI) Tool- Tableau analysis like the current inflow of any particular country or state and according to that we also get the predicted inflow for upcoming years using these tools we analyze and investing is very helpful for investors.

Research Methodology

Data Collection and Filtration: In the research, we collected secondary data. The database is created by gathering info and data from several consistent sources like the Department of Industrial Policy and Promotion (DIPP), the Reserve Bank of India, and other resources like Online databases of the Indian economy, articles, journals, etc.

Data Analysis by visualization tool: In figure 1Tvrdíková (2007), the process of the system flow is shown. Data were analyzed by a visualization tool tableau where we get the statistical data and created many visualizations like a trend line that tells the trend of the current situation of the FDI in India and created the forecasted line and graphs chart which shows the predicted inflow of upcoming years based on the previous data.

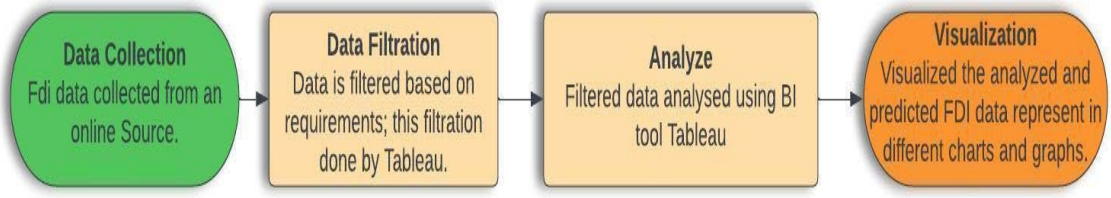


Figure 1: System Flow of research problem

Design Of Experiment

- Select a sample of FDI data in India from the online source.
- Visualize and predict the FDI inflow in India using the BI tool Tableau.

Selection Of Samples

Data on FDI is collected from the Reserve bank of India's official website and other online sources. In this research paper, we analyzed the last 22 years' data which is the year 2000 – 2022. The total amount of inflow in the last 22 years is 8, 87,762 USD millions The used data set is a secondary data set and data is filtered according to requirement, here we use only those data which is related to FDI This research paper only focuses on FDI. Using the given last 22 years of data set we predicted the upcoming trend of FDI in India which is increasing or decreasing.

Conduction Of Experiment

Data collected from Reserve Bank of India official site and online sources undergo the experiment as mentioned below:

Collection of Data: In this step data will be collected from different sources and data will be secondary.

Data Filter: Based on research requirements to filter the data only useful data will be used and others will be filtered.

Analyze: Once we get the required data then the data was used for analysis purposes for BI tools. In analysis, we get the final output.

Visualization: In visualization, we can present the analyzed data in the form of a table, map, or any other chart. We can use an understandable chart for specific output visualization. Figure 1, shows the dashboard of the BI tool Tableau Used for FDI data visualization.

Figure 2: FDI Tableau Dashboard

Figure 2 shows the tableau dashboard for analyzing collected data.

Data Analysis And Result

Trends of FDI Inflows in India Using Tableau BI Tool: The Foreign Direct Investment (FDI) trend in India has been encouraging in current years. India has arisen as one of the most striking termini for FDI internationally, due to its large customer bazaar, trained labor force, and favorable government policies.

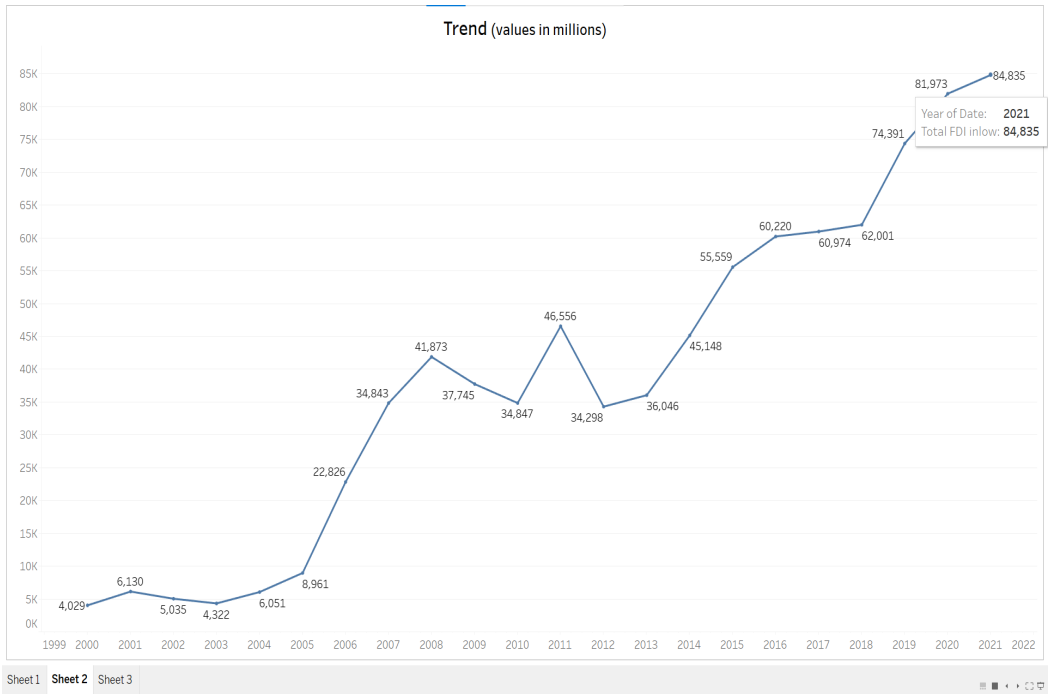


Figure 3: Analysis using Tableau (BI) tool on Current Trend From 1999 – 2022 Values in US Million

Figure 3, created using **the BI** tool with the help of data which is collected from (dpiit.gov.in)shows the trend of FDI in India which is positive and is expected to continue growing in the near term, driven by India's huge consumer market and FDI inflow in each as shown in the above figure. As we see in 1999 the inflow of FDI was 4,029 million it increased from 1999 to 2005 and is not consistent. As we see in Figure 3, there is massive growth after FY. 2005 - 2008 it will reach 41,873 million from 8,961. There is $ 32,912 million growth in 3 years. The major cause behind the increase in FDI was the introduction of automatic routes.

Impact Of Foreign Direct Investment On States Of India

Foreign Direct Investment (FDI) has a significant impact on the states of India. It can enhance economic growth, create jobs, and improve infrastructure. Some of the key benefits of FDI for the states include increasedInvestmentInvasion: FDI brings in fresh capital that can be used to finance development projects and support the growth of local businesses.

Statement On State-Wise FDI inflow from October 2019 to September 2022	
State Name	**FDI inflow (values in US million)**
Maharashtra	47,165
Karnataka	39,361
Gujarat	30,660
Delhi	22,197
Tamil Nadu	7,896

Table 1: Top 5 States Value in US Million (Source: FDI Statistics from FDI fact sheet released by dpiit.gov.in)

Every State of India makes different FDI inflows as per the Reserve Bank of India. Table 1 shows state-wise data for the top 5 States by FDI inflow in US millions these top 5 states are according to FY. October 2019 to September 2022. In order to top states. Maharashtra received the highest inflow in USD millions. The state of Maharashtra is the highest receiver of FDI which is $47,165 million.

State Name	FDI inflow (values in US million)
Meghalaya	1.097
Jammu And Kashmir	1.055
Tripura	0.562
Ladakh	0.188
Nagaland	0.014

Table 2: Bottom 5 States Values in Us Millions (Source: FDI Statistics from FDI fact sheet released by dpiit.gov.in)

Table 2 gives the bottom 5 states with their inflow this state received the least inflow. The state of Nagaland is the lowest receiver of FDI which is $ 0.014 million.

Country-Wise Flow of FDI into India by Using Tableau BI Tool

Country-wise FDI in India data is stated below which shows the top 10 countries by FDI inflow in US millions these 10 countries belong to the last 22 years.

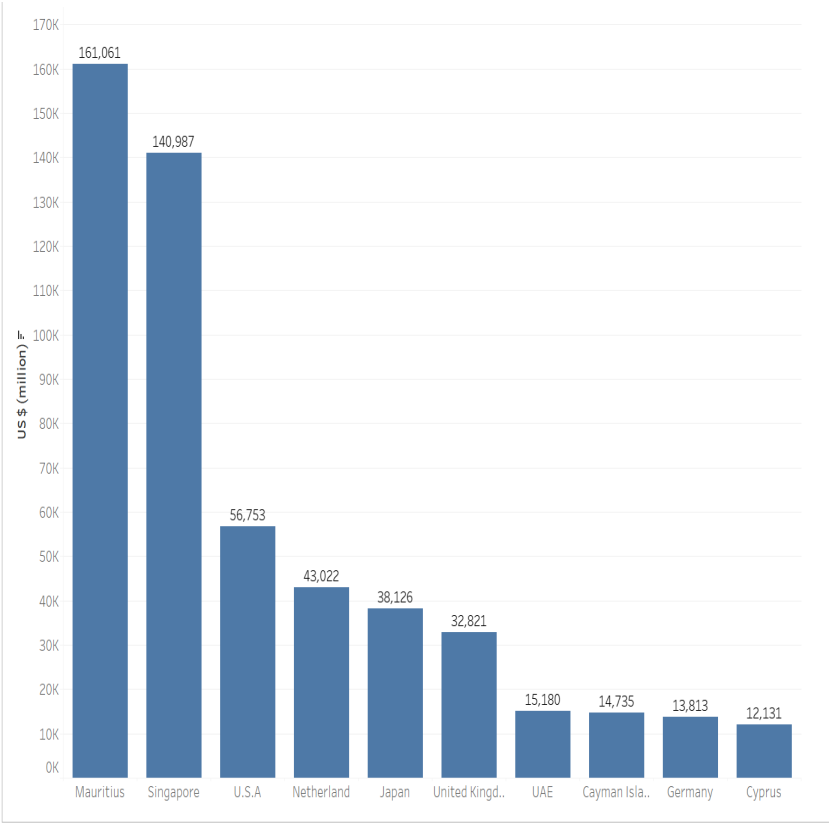


Figure 4: TOP 10 Countries By Inflow Values in US Million (Source: created using BI tools with the help of data which is collected from (dpiit.gov.in)

Statement On Country-Wise FDI Inflow From April2000 to September 2022	
Country Name	**FDI inflow (values in US million)**
Mauritius	1,61,061
Singapore	1,40,987
U.S.A	56,753
Netherland	43,022
Japan	38,126

Table 3: Top 5 Country Values in US Million (Source: Analyzed FDI Statistics from FDI fact sheet released by (dpiit.gov.in)

Table 3 shows that in the last 22 years of country investment top investors such as Mauritius have valued 1,61,061 US million.

Sector-Wise Distribution of FDI Inflows in India

The inflow of FDI in India is mostly diversified into different sectors. The top 5 most attractive sectors for investment in India according to analyses are mentioned below the top 5 sectors according to the last 22 years.

STATEMENT ON SECTOR-WISE FDI INFLOW FROM APRIL 2000 TO September 2022	
Sectors	**FDI inflow (Values in us million)**
SERVICES SECTOR	98,356
COMPUTER SOFTWARE & HARDWARE	91,799
TELECOMMUNICATIONS	39,025
TRADING	38,021
AUTOMOBILE INDUSTRY	33,774

Table 4: Top 5 sector Values in million (Source: Analyzed FDI Statistics from FDI fact sheet released by dpiit.gov.in)

Table 4 shows that the top sector that received FDI in India is SERVICES SECTOR $98,356 million and Computer Software & Hardware sector is having a good number $91,799 million.

Predict/Forecast FDI Inflows in India

Here we can Predict / forecast the FDI inflow in India using tableauup to FY. 2025 Figure 5, below shows that total inflow is increasing year by year according and it will increase by a value of 98925 million in the year 2025.

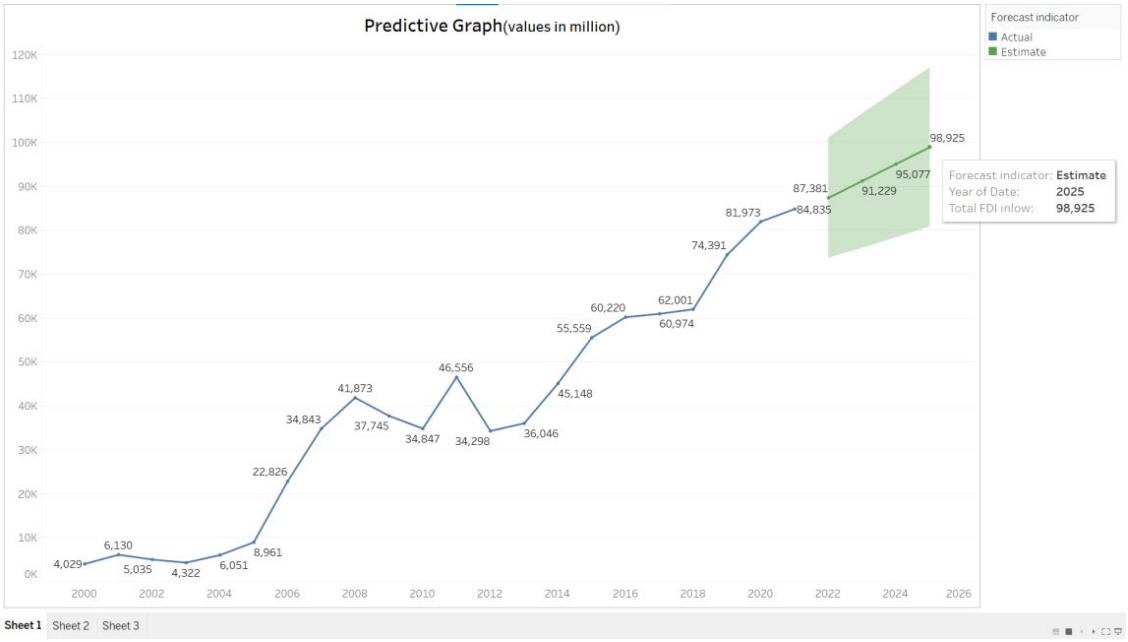


Figure 5: Predictive FDI Inflow from FY.2022-2025 (source: Researcher)

Figure 5 is created using BI tool with the help of data which is collected from the department for the promotion of industry and internal trade(dpiit.gov.in)

Conclusion

In the research, we can conclude that the appreciation rate or trend in FDI has not been very stable but has positively gone up in spite of the predominant ups and downs in the economy. Visualization and analysis of the

states-wise FDI data in India show that FDI not only attracts richer states but also goes to poor states; there is a difference in both states like Maharashtra, Delhi, and TamilNaduare developed states that are the reason they received more. At the same time, FDI has proved very much helpful in the growth of poor states like Bihar and Jharkhand. Indian state governments like Madhya Pradesh, Orissa, Rajasthan, Bihar, Jharkhand, and some north Eastern States should alter norms for FDI towards giving a boost to sales, acquiring resources, improving infrastructure, increasing the supply in the market, and making it less risk-oriented. The topmostSectors that attracted foreign investment are services,computers, telecommunication, construction,and hardware. According to data analysis Mauritius, Singapore, the U.S.A, the Netherlands, and Japan, this country are leading sources of FDI. This Top 5 country covers 71% of FDI inflow in India. The decisions governing FDI have been spread over many areas that have to be streamlined or a practical policy has to be developed because of the impact of the reforms in India on the policy environment for Foreign Direct Investment. With all the measures the government has taken we can predict/forecast from the study that it will grow more and will achieve 98925 million dollars in 2025 FDI helps the country in removing infrastructure bottlenecks, increasing exports, providing skilled and trained manpower, removing local disparity in the states and helping in accomplishing an all-round development of each and every part of the states in India.

References

Anitha R. (2012),"Foreign direct investment and economic growth in India.", International Journal of Marketing, Financial Services & Management Research, 1(8), 108-125

Bhargava M., Kiran K. & Rao D. (2018), "Analysis and design of visualization of educational institution database using power bi tool.", Global Journal of Computer Science and Technology, 18(C4), 1-8.

Kumar V. (2021), "An Analysis of Foreign Direct Investment in India.", Asian Basic and Applied Research Journal, 95-102.

Merajothu D. (2020), "An empirical study on foreign direct investments impact on economic growth of India", Available at SSRN 3598037.

Patil J., Salunkhe S. & Kadam B. (2014), "Effects of FDI on Indian Economy: A Critical Appraisal", Journal of Economics and Sustainable Development, 19.

Ramasamy M., Dominic D. & Poovendhan M., (2017). "Effects of FDI spillover on regional productivity: Evidence from panel data analysis using stochastic frontier analysis", International Journal of Emerging Markets.

Rani S., Ghosh S.(2020), "A study on foreign direct investment in India.", PalArch's Journal of Archaeology of Egypt/Egyptology, 17(7), 15968-15977.

Sharma S.,Djiaw V., (2011). "Realizing the strategic impact of business intelligence tools", Vine.

Singh S., (2019), "Foreign direct investment (FDI) inflows in India", Journal of General Management Research, 6(1), 41-53.

Tvrdíková M., (2007), "Support of decision-making by business intelligence tools" In 6th International Conference on Computer Information Systems and Industrial Management Applications (CISIM'07) (pp. 364-368). IEEE.

Duggal A., "Foreign Direct Investment in India.", Journal of Internet Banking and Commerce

https://www.icommercecentral.com/open-access/foreign-direct-investment-in-india. php?aid=86435

Miyamoto K.(2003), "Human Capital Formation And Foreign Direct Investment In Developing Countries", Paper No. 211 OECD Development Centre

Rahate V., Yadav M., Chouhan B., Narvare S. & Sawant K.(2021), "Data Analytics for Betelnut's Selling Dataset Using Tableau", (2021) International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)

"FDI Statistics", available at https://dpiit.gov.in/publications/fdi-statisticssite accessed on 15th February 2023.

https://www.tableau.com/node/62770site accessed on 26th February, 2023.

ISSUES AND CHALLENGES OF WEB SCRAPING: HEALTHCARE INDUSTRY CASE STUDY APPROACH

Prof. Shravani Pawar Assistant. Professor,
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai,
pawarshravani81@gmail.com

Dr. Priya Chandran Assistant. Professor,
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai,
priyaci2005@gmail.com

Mr. Pawan Salvi, Student,
Master of Computer Application,
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai
, pawanpsalvi2610@gmail.com

ABSTRACT
The Internet is the greatest source of information mankind has ever created. There are various distinct materials in many formats including audio,video, text, etc. However, the data that makes up much of the internet is disorganised, making it challenging to extract and use in automated procedures. Web scraping avoided this labour-intensive manual process of organizing and extracting information, providing an easy way to collect facts and figures from web pages, transform it to a format of your choice, and store it locally. Web scraping has a wide range of uses, which includes brand monitoring, sentiment analysis and data augmentation. Many organizations use different methods to extract useful information. This research paper focuses on various web scraping tools and libraries that have been developed in recent years and are widely utilised to gather data, transform it into structured data, and utilise this organised data in word processing applications. We also discussed the issues and challenges of implementing web scraping techniques in the healthcare industry.
Keywords: Web Scraping, Web data, Data extraction, Data analysis, Extraction error handling

Introduction

Most computer users use the internet and browse the websites using multiple browsers, where data and multimedia are displayed in a way which is easy to understand. In spite of the fact that doing so is totally up to the site owner, many websites provide APIs that can be used to swiftly access much of this data. It is straightforward for them to decide not to grant API access to this data. On the other hand, web scraping and web crawling are faster and more viable methods that can be used to collect information from thousands or even millions of web pages. This technique is pretty beneficial for a variety of applications, however it excels in commercial enterprise intelligence. Since it facilitates them to make decisions, information is critical for groups and organisations, specifically for the reason that the majority of data is now available online. Data collection from many sources, including both public and private ones, is the primary step in any data science research or development. Company sales data and financial reports are examples of private sources. Open data, websites, and journals are examples of public sources. Website analysis, website crawling, and data organisation are the three key, related stages in online scraping. Web scraping and data mining are distinct from one another since the latter includes data analysis while the former is not relevant in this situation. For data mining, sophisticated statistical techniques must also be applied. Because there are so many widely available tools and libraries that offer productive executions of a substantial chunk of the required functionality, web scraping is frequently a very simple process. The capacity to send distinct HTTP requests with varying headers and payloads is a unique feature of the majority of web scraping programmes. This study examines online scraping, including what it is, working, technologies used and how it connects to business intelligence and artificial intelligence. We have also discussed the issues and challenges of implementing web scraping techniques in the healthcare industry.

Literature Review

Ferrara (2014) discussed a number of technological challenges relating to the amount, diversity, velocity, and genuineness of data on the web has to be resolved before web data may be utilised. Quantitative and subjective information are exchanged in a variety of organised, semi-structured, and unstructured formats on the web, including web pages, HTML tables, web databases, emails, tweets, blog posts, photographs, and videos.

Glez (2014) studied web scraping as the automated extraction and structuring of data from the web using technological tools with the intention of further analysing this data. Individual researchers or even big study teams would find it difficult to physically collect and compose Big Data that is readily available on the Web because of its amount, diversity, velocity, and authenticity.

Baumgartner (2005) studied website analysis, website crawling, and data organising are the three main, interwoven stages of online scraping. Website analysis is looking into a website or Web repository's fundamental structure in order to comprehend how the necessary data is kept. This calls for a fundamental comprehension of the World Wide Web architecture, mark-up languages (such as HTML, CSS, XML, and XBRL), and various Web databases (e.g. MySQL). A script which automatically browses a website and obtains the necessary information is created and run as part of website crawling.

Fernandez (2011) discussed the availability of tools for the automated crawling and parsing of web data has something to do with the common request of these languages in Data Science. The techno-logical outline of business analytics accordingly supports analytics implanted in decision support activities of businesses. As a holistic approach, business analytics encompasses all disciplines of business administration.

Hillen (2019) It is vital to clean, pre-process, and arrange the required data after it has been extracted from the chosen web source. This will allow for further analysis of the data. Given the amount of data involved, a programmed approach could also be required to reduce the amount of time spent. Natural Language Processing (NLP) libraries and data manipulation methods are available in several computer languages, including R and Python, and are helpful for cleaning and organising data.

Zhou (2014) in his study the authors have focused on extracting web contents which are less structured, such as new articles. They also have proposed a method to automatically cluster and extract based on the relevance identified.

Michalakidis (2016) proposed different techniques to extract data from different structured and unstructured sources. The authors have proposed an error reporting mechanism associated with collection of data and metaphysical based data dictionaries for improving the quality.

Kumar (2020) stated that data mining techniques are extensively used in healthcare research. Most of the data collected are unstructured and it is very difficult to collect that data manually. The authors have discussed web scraping techniques to collect data from unstructured HTML documents and store it in a format capable of doing data analysis.

Singrodia, Khder & Krotov (2022) various aspects of web scraping and its tools and techniques are discussed the legal and ethical issues related to web scraping techniques used.

Hillen (2019) discusses how web scraping can be used in food price research. He also discusses data collection methods on different real time data. The authors have also discussed the limitations in terms of non-availability dataset in food price research.

Tools & Techniques used in Web Scraping

Python:

Python is an object-oriented, high-level programming language. Code clarity is prioritised in its layout philosophy, which makes heavy use of indentation. It supports quite a few programming paradigms, which include procedural, object-oriented, and practical programming in addition to established programming.

Beautiful Soup:

A Python module alluded to as Beautiful Soup is utilised to parse HTML and XML texts. For processed pages, it produces a parse tree that can be utilised to extract HTML facts for net scraping. The application is likewise funded with the aid of using Tide lift, a paid open-supply preservation subscription.

Requests:

One essential aspect of Python for sending HTTP requests to a given URL is the requests module. REST APIs and internet scraping each want requests, which needs to be learnt earlier than the use of those technologies further. A URL responds to requests by returning a response. Python requests have integrated control equipment for each request and the response. Python customers may also put up HTTP/1.1 requests and the use of the HTTP library requests, which is licenced beneath the Apache 2 licence. Python requests are essential to test with the internet. One needs to ship a request to the URL so that you can perform moves including hitting APIs, downloading complete Facebook pages, and performing many different things

JavaScript:
The lightweight object-oriented programming language JavaScript (is) is used by many websites to script their webpages. It is a full-fledged, interpreted programming language that, when combined with an HTML content, enables dynamic interactivity in websites.

Discussion and Analysis
Broadly speaking, web data scraping is a method of systematically collecting and combining information from different web sources. A software agent known as a web robot simulates the browsing interaction between web servers and the users. The robot systematically scans a required number of websites, analyses their contents to identify and extract relevant information, and then organizes that content as needed. Web scraping APIs and frameworks are commonly used by the organizations to extract the useful information.

We identify the tools that are now available on the market, explain the terms "web scraping" and "web crawling," and instruct readers on how to build their own web scrapers using one of these tools. The Web data scraper connects to the target Web site via the HTTP protocol, a stateless text-based Internet protocol that governs request-response interactions between a client, frequently a Web browser, and a Web server. Web data scrapers must carefully arrange their retrieval tasks to avoid overloading the server and must abide by the site's terms of usage. After obtaining the HTML file, the Web data scraper can extract the desired contents. For this reason, regular expression matching is widely employed, either independently or in conjunction with other justifications.

Web scraping process is depicted in Figure1. It involves amassing all of the records that have been retrieved from a couple of URLs provided, which can be structured, semi-structured or unstructured. The information is extracted and it should be pre-processed, cleaned and transformed to a format which can be used according to the business requirements. After the transformation, the records are stored inside the business database for further usage.

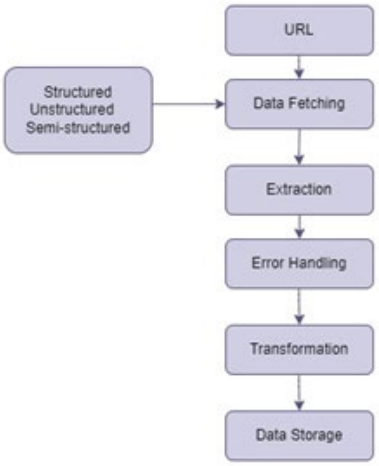


Fig1: Web Scraping Process

Features of web Scraping
1) Pace
The speed that web scraping technology offers is by far its most prominent advantage. Web scraping enables you to quickly rub a few pages at once without having to screen and oversee each individual request. Web scraping's speed is due in part to its ability to scan web pages quickly and extract data from them as well as the ease with which it may be incorporated into daily activities. Because you don't need to worry about creating, downloading, integrating, or installing web scrapers, getting started with them is simple.

2) Profitability
Web scraping offers a complex benefit at a sensible cost, which is one of its best highlights. You won't have to contribute to developing a complex system because a basic scraper can frequently complete the entire task. A professional data extraction project would be impossible without automation because time is money and the web is developing at an accelerated rate.

3) Flexibility and methodical approach
Web scraping programmes and APIs are not pre-programmed fixes. They are therefore very open, adaptable, and interoperable with other scripts. Make a scraper for a single, large work, then modify it to meet a variety of smaller tasks by making just minor changes to the core. For instance, a workflow made of various APIs may be used to scrape every monitor offered by Tata 1mg and compare it to the ones you sell online. The online scraping API gives users the ability to personalise the data gathering and analysis process and make the most of its capabilities to achieve all of their web scraping goals.
4) Performance reliability

Data accuracy is a process in and of itself that web scraping offers. Once properly configured, your scraper will accurately and reliably capture data directly from websites with very little risk of error. It is crucial to have data in a comprehensible and orderly manner in addition to being able to gather it. If the script is properly written, you can virtually remove the possibility of error and guarantee that the information and data you obtain are of higher quality each and every time you gather them.

5) Automatic delivery of structured data
Simple values may frequently be utilised right away in other databases and applications because of the fact that well-scraped data always comes in a machine-readable format by default. Its most appealing feature for both professionals and non-pros is the simple API interaction with other applications. The initial stage in your data analysis pipeline that includes other built-in solutions is web scraping. Web scraping is a difficult operation since it involves many different computer languages and software programmes. The greatest thing is that a web scrape will also do the required account maintenance, such as simple troubleshooting, updates, and backups. I can be confident that my data is secure when I use web scraping, which is the main advantage.

Limitations
1) Needs perpetual maintenance
Maintenance of the product is the actual deal. You have no influence over whether an external website changes its HTML layout or content because your scraper's activity is inextricably linked to it. Developers must therefore respond to those changes in order to prevent scrapers from breaking or becoming out of date and unable to keep up. While some information will be updated automatically, any scraper will often require ongoing maintenance to remain operational.
2) Data Extraction
Setting realistic expectations is crucial when working with complex data extraction and processing. The main purpose of a scraper is to gather the necessary sort of data, package it in the format you require, and upload it without loss into your computer or database. Although the data will be delivered in a structured format, more complicated data will need to be processed before it can be incorporated into other applications.
3) Scrapers can get blocked
Some websites simply dislike being scraped. They might do this because they think scrapers are using up their resources, or it may essentially be that they do not need to create it basic to match businesses to compete. Access can occasionally be denied due to the scraper's place of origin. The usage of proxy servers is a common solution to this form of IP blocking. In many situations, these techniques can let scraping bots work covertly. In any case there are circumstances when even these remedies are inadequate to handle severe blocking, and the final drawback is that a website cannot be scrapped.
4) Learning Curve
It takes practice to master even the simplest scraping tool. Some tools still need you to know how to code. Some tools for non-coders may take weeks to learn. Understanding of XPath, HTML, and AJAX is required in order to efficiently scrape webpages.
5) The structure of websites change frequently
Data that has been scraped is organised in accordance with the website's structure. Sometimes when you visit a site again, the design has changed. Some website designers update their work frequently to improve user interface, while others may do so to prevent scraping. Changes to the website layout might be significant or little, such as moving a button's position. Your data can become corrupt even with a small alteration. You must modify your crawlers every few weeks in order to obtain accurate data because the scrapers were constructed using the old site as a guide.
6) Data extraction on huge scale is difficult
Some tools can only handle small-scale scraping, therefore they cannot extract millions of data. Owners of ecommerce businesses that require millions of lines of frequent data feeds directly into their database are bothered by this. Multiple cloud servers are used to perform tasks. You get blazing speed and colossal storage space.
7) A web scraping tool is not omnipotent

Texts may be extracted from source code and formatted using regular expressions using sophisticated programmes. Pictures can only be scraped for their URLs, which may then be transformed into images. It is significant to highlight that because most web scrapers gather data by parsing via HTML components, they are unable to crawl PDFs.

Case Study Approach: Web scraping in Health Sector

In our research we have conducted a study on the challenges and issues of web scraping in the healthcare industry. The fact that public health records vary in size and type is undeniable. Web scraping too is slowly and steadily gaining importance in healthcare. The truth that, vast amount of information produced by the medical industry is very difficult to analyse using traditional methods. So, web scraping along with data mining can improve decision making by identifying patterns and trends in large volumes of complex data. One can filter some websites with web scraping and use the information for public health analysis, prescription drugs pricing analysis, disease monitoring, insurance database, competitive analysis etc. In this case study first we discuss the data extraction problems in the healthcare industry. Most of the healthcare research depends on the data collected from Electronic Patient record (EPR) and data repositories. Most of the EPR data is unstructured or semi structured (Michalakidis G, 2016). Also these data are collected from heterogeneous sources and each may have its own structures. These problems of collecting data from heterogeneous sources can be concluded as,

- Local autonomy
- Architectural difference
- Representational dissimilarity
- No Precise interpretation

Generic approach for error reporting.

The above problems can be avoided if the data is collected from a single source. Since the data is very crucial for the research we cannot restrict to data from a single source. The solution for this is to the need for a structured or generic approach for data extraction. By using this approach, we can categorize the extraction error into different classifications and can address these issues according to the severity of the error. This paves the way for creating an online generic approach for error reporting.

Conclusion

Online scraping is a well-known term that has gained more prominence as a result of the need for free data gathered from web pages. The data are needed by many professionals and researchers for processing, analysis, and the extraction of important outcomes. In contrast, those working with use cases need to allow data from many sources to be integrated into creative applications that will provide supplemental benefits and originality. We have examined the many facets of web scraping, starting with the web scraping tools and software, and also looked at their advantages and disadvantages. Then discussed the challenges and issues of web scraping in the healthcare industry. We look forward to extend our study in other sectors like the investment sector and also to study novel approaches for data collection using web scraping.

References

Baumgartner, R., Fröhlich, O., Gottlob, G., Harz, P., Herzog, M., & Lehmann, P. (2005). Web data extraction for business intelligence: the lixto approach. Gesellschaft für Informatik eV.

Fernández V, J. I., Blasco Garcia, J., Iglesias Fernandez, C. A., & Garijo Ayestaran, M. (2011). A semantic scraping model for web resources-Applying linked data to web page screen scraping.

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. Knowledge-based systems, 70, 301-323.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797.

Hillen, J. (2019). Web scraping for food price research. British Food Journal, 121(12), 3350-3361.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing & Its Applications, 13(3).

Krotov, V., & Johnson, L. (2022). Big web data: Challenges related to data, technology, legality, and ethics. Business Horizons.

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping.

Kumar, V., Thareja, R., Thareja, R., & Jain, P. R. (2020). Applying Data Science Solutions in the Healthcare Industry. In ICT for Competitive Strategies (pp. 35-42). CRC Press.

Michalakidis, G. (2016). Appreciation of structured and unstructured content to aid decision making-from Web scraping to ontologies and data dictionaries in healthcare. University of Surrey.

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scraping and its applications. In 2019 international conference on computer communication and informatics (ICCCI) (pp. 1-6). IEEE.
Zhou, Z., & Mashuq, M. (2014). Web content extraction through machine learning. Stanford Univ, 1-5.

