Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

Parkinson's Detection Using Machine Learning

Surekha Tadse
Dept. of Electronics & Telecommunication,
G H Raisoni College of Engineering,
Nagpur, India
surekha.tadse@raisoni.net

Muskan Jain
Dept. of Electronics & Telecommunication,
G H Raisoni College of Engineering,
Nagpur, India
jain_muskan.et@ghrce.raisoni.net

Pankaj Chandankhede
Dept. of Electronics & Telecommunication,
G H Raisoni College of Engineering,
Nagpur, India
pankaj.chandankhede@raisoni.net

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432340

Abstract*— Advance technology such as Data Science can be used to find solutions to medical science problems, by using its data and implementing Machine Learning Algorithms on to it, to draw the insights and patterns from the data and spot out the possibilities. This is our approach to find out a way to detect Parkinson's disorder at an early stage; to provide necessary treatment early using Machine Learning with Data Science. Data Science processes and its methods for extracting knowledge and insights from large volumes of data would be witnessed in the project. Machine Learning algorithms are applied onto patient's data-set and accuracies of the algorithms are compared. The model with the highest accuracy fits the best to predict target values for unknown data values. In this way, integrating Medical Science and Data Science with Machine Learning, PD could be detected earlier and necessary treatment would suffice a patient to recover at a good rate.

Keywords*— Data Science, Machine Learning (ML), Machine Learning Algorithms, Parkinson's Disorder (PD), Data-set, Algorithm's Accuracy.

I. INTRODUCTION

A. Background

Parkinson Disorder, the second most common disease after Alzheimer's and a major public health problem (Worldwide- 5M), is a progressive neurological disorder in the central nervous system whose cause is still unknown, it could be due to environmental factors and/or genetic factors which increases with time.

Our body movements and coordination functions well due to a chemical substance Dopamine present in brain. Dopamine can be called as a chemical messenger in brain which has many functions. It completes the pathways of signals from brain to motor cells and vice-versa. Dopamine is produced in a part of brain, Substantia Nigra. In PD, Substantia Nigra cells tend to die out, which leads deficiency of Dopamine, as soon as Dopamine level falls from 60% to 80%, symptoms of PD start to occur.

Symptoms of Parkinson's Disorder can be Motor or Non-Motor. Motor symptoms are the symptoms which could be visually perceived, also called as cardinal symptoms. Primary motor symptoms include, tremors, slowness in motion, rigidity and problems with balance. Non-motor symptoms include rapid movement of eye, disarrayed behavior in sleep, loss of smell, cognitive impairment, constipation, speech and swallowing problems, unexplainable pains and many more. Initially, symptoms of Parkinson's and other disease tend to make no specific difference and hence, PD remains undiscovered to patients. PD has no definite medical cure till now and could be only controlled by oneself by healthy diet and exercise. This makes it important to diagnose PD at an early stage.

B. Technological Approach to Biomedical Problems

By definition, Data Science is the scrutinization of huge amounts of big and raw data, which possess potential to reveal insights that help organizations grow by making strategic choices. Technologically, Data Science is changing the way industries work, the way it uses its data and approaches to its problems. Data Science is helping organizations understand their environment, analyze their existing issues and reveal previously hidden opportunities. The rate at which the data is growing is hugely powerful and is available from numerous sources such as, log files, emails, social media, sales data, patient's information data, sports performance data, sensors data, security alarms and many more. Data can be available as structured data or unstructured data that leads way to multiple patterns which could be approached to develop solutions, using powerful visualization tools to understand the nature of results and recommend actions to be taken next.

Machine Learning, can be visualized as a subset of Artificial Intelligence that analyses data and takes intelligent decisions using computer algorithms based on its learnings without to be programmed explicitly. Machine Learning is what enables machines to solve problems on their own and make accurate predictions using the provided data. ML basically uses Big data. Big Data refers to datasets that are massive, so quickly built and varied that they defy traditional analysis methods.

Big Data is driving digital transformations as digital transformation affects business operations, updating existing processes and operation & creating new ones to harness the benefits of new technologies. This digital change knows how to operate and deliver values to its customers. The availability of vast amount of data and the competitive advantage that analyzing it brings, has triggered digital transformation throughout many industries. Digital transformation is not simply duplicating existing process in digital form, the in-depth analysis of how the business operates, help organizations discover how to improve their processes and operations, and harness the benefits of integrating data science into their workflow.

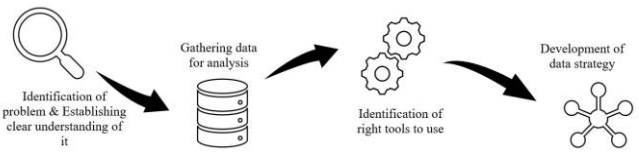


Figure 1. How Data Science works

II. Related research Work

Different kinds of features and data have been used and experimented by researchers to predict Parkinson's Disorder. Models which could automatically predict PD in people by simply analyzing their voice samples have been developed by different authors. Machine learning techniques such as fuzzy c-means (FCM) clustering and pattern recognition methods have been used on big data with accuracy 68.04%, sensitivity 75.34% and specificity 45.83%.

In paper [1], data was modelled for the non-motor symptoms and the biomarkers for example, dopamine transporter imaging and cerebrospinal fluid measurements. Researchers of this paper implemented 8 Machine Learning Algorithms, namely Multilayer Perception, Bayes Net, Random Forest, Boosted Logistic Regression, Boosted Trees, Naïve Bayes, Support Vector Machine and Logistic Regression. According to their work, Random Forest and Boosted Trees prove to the best model with 100% Accuracy. And Naïve Bayes with least accuracy of 94.67%.

Performance Measures	Multilayer Perceptron		BayesNet		Random Forest		Boosted Logistic Regression	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Accuracy(%)	96.09	95.4545	96.5854	96.027	100	96.59	95.8537	97.1591
Recall	0.961	0.955	0.966	0.960	1	0.966	0.959	0.972
Precision	0.962	0.955	0.967	0.965	1	0.970	0.959	0.974
F-Measure	0.961	0.955	0.966	0.961	1	0.967	0.959	0.972
AUC	0.989	0.986	0.994	0.994	1	0.997	0.995	0.989

Table 1. ML algorithms with their performance measures (Paper [1])

Paper [2], primarily focuses on Speech Articulation Difficulty Symptoms to build a model. It also emphasizes on Age factor (65+); varies from individual to individual.

India accounts for 11,747,102/1.06B PD patients; out of which 90% suffer from Speech difficulty i.e. Dysphonia, Dysarthria.

In order to completely understand the market scenario, awareness amongst people, number of PD patients and their approach towards the treatment, neurologists were telephonically approached.

Dr. Dhruv Batra (Neurologist, Government Medical College and Hospital, Nagpur, India) and Dr Pratik Uttarwar (Neurologist, Seven Star Hospital, Nagpur, India) led us to the following:

1. Breaking or dying of some nerve cells in the brain leads to impairment in movements of the body. The cause of the break or death of nerve cell is not certain but suspected to be the lack of dopamine, which is produced in substantia nigra part in the brain. This disorder could be genetic or developed over time.
2. Four primary motor symptoms in Parkinson's patient are tremor, rigidity, slow movement and postural instability. Some symptoms which are taken normally as aging, tremor in palms while eating, writing, movements of hands, walking etc. directly point towards Parkinson's disorder.
3. Symptoms of Parkinson's disorder are often confused with other disorders. Treatment comes with a series of checkup which varies from person to person and response to drug treatment help to distinguish them from Parkinson.
4. Frequency of visits of people having PD is high, roughly around 500 per week, out of which hardly 10% are aware of the disorder. Awareness regarding PD is poor amongst people.
5. To collect data of PD patients, study and contact with different clinics is necessary.

III. METHODOLOGY

Machine Learning is branch of Computer Science that provides, "computers the ability to learn without being explicitly programmed" said by Arthur Samuel, an American pioneer in the field of computer gaming and AI, coined the term Machine Learning in 1959 at IBM.

Data Science Methodology aims to answer the following 10 questions in the prescribed order:

1. What is the problem we are trying to solve?
2. How can we use data to answer the questions?
3. What data do we need to answer the questions?
4. Where is the data coming from and will we get it?
5. Is the data that we collected representative of the problem to be solved?
6. What additional work is required to manipulate & work with the data?
7. In what way can data be visualized to get the answers?
8. Does the model used really answer the initial question or does it need to be adjusted?
9. Can we put the model into practice?
10. Can we get the constructive feedback into answering the questions?

Business Understanding is an important stage in Data Science as it helps us clarify the goal of the entity asking the question. It gives us the exact root to hold onto to, to adopt a certain approach towards the solution. Analytical approach helps to what type of patterns would be required to address the problem statement most effectively. Data Science Methodology being highly iterative in natures, makes the process a never ending one and makes optimization of the process very handy and easy.

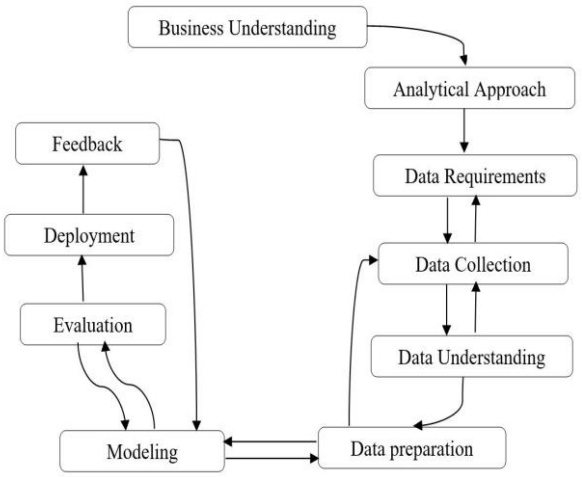


Figure 2. Data Science Methodology

A. Dataset

1) Source and information

The dataset used in the study was created by Max Little of the University of Oxford, which was in collaboration with the National Centre for Voice and Speech, Denver, Colorado, which recorded speech signals. The feature extraction methods for general voice disorders were published in the original study.

A range of biomedical voice measurements were recorded of 31 people, in which 23 having Parkinson's disease (PD) is contained in the dataset. The prime objective of the collected dataset is to distinguish Parkinson disorder affected people from healthy people. In the dataset table, the column "status" contains values '0' and '1', in which '0' stands for healthy people and '1' stands for PD affected people. The data is stored in the ASCII CSV format. Each row in the table corresponds to a voice recording of a person. Each person holds six recordings, in order to study the variations

2) Feature descriptuon

Feature/Attribute	Description
name	*ASCII subject name and recording number*
MDVP: Fo(Hz)	*Average vocal fundamental frequency*
MDVP: Fhi(Hz)	*Maximum vocal fundamental frequency*
MDVP: Flo(Hz)	*Minimum vocal fundamental frequency*
MDVP:Jitter(%) MDVP:Jitter(Abs) MDVP:RAP MDVP:PPQ Jitter: DDP	*Several measures of variation in fundamental frequency*
MDVP:Shimmer MDVP:Shimmer(dB) Shimmer: APQ3 Shimmer:APQ5 MDVP:APQ Shimmer: DDA	*Several measures of variation in amplitude*
NHR, HNR	*Two measures of ratio of noise to tonal components in the voice*
status	*Health status of the subject (one) - Parkinson's, (zero) – healthy*
RPDE, D2	*Two nonlinear dynamical complexity measures*
DFA	*Signal fractal scaling exponent*
spread1, spread2, PPE	*Three nonlinear measures of fundamental frequency variation*

B. Prediction models

In this study, four different machine learning algorithms, Decision Tree, Logistic Regression, K-Nearest Neighbors and Support Vector Machine are used. A brief description of each of them is provided below.

1) Decision Tree

Belonging to the family of supervised learning algorithms, a Decision tree has the main goal of developing a training model that can be used to predicted target variable by simply learning the decision rules inferred from training data. Packages *pydotplus* and *python-graphviz* were essentially used to visualize the decision tree.

A recursive partition of the dataset can be called as Tree classifier. This classifier primarily comprises of a set of nodes: root node, internal nodes and leaf nodes. There exists one root node. Internal nodes are the nodes which have exactly one incoming and one outgoing. Terminal nodes or leaf nodes are nodes with no outgoing edges.

Using IF-THEN rules, a decision tree could be easily generated which could be used as a tool to extract valuable information and insights from datasets [2].

2) Logistic Regression

In this paper, Logistic Regression study is observed to provide the best results as Logistic regression works by fitting to a special s-shaped curve and taking the linear regression in process and converting the numeric insights into a probability format as follows, which is also called sigmoid function σ:

$$h\theta(x)=\sigma(\theta TX)=e(\theta 0+\theta 1x1+\theta 2x2+...)1+e(\theta 0+\theta 1x1+\theta 2x2+\cdots)h\theta(x)=\sigma(\theta TX)=e(\theta 0+\theta 1x1+\theta 2x2+...)1+e(\theta 0+\theta 1x1+\theta 2x2+\cdots)$$ or,

$$Probability\ Of\ a\ Class = P(Y=1|X) = \sigma(\theta TX) = e\theta TX1+e\theta TX$$

- θTX represents the regression result (the sum of all the variables weighted by the coefficients)
- exp represents the exponential function ( ).
- $\sigma(\theta TX)$ represents a sigmoid/ logistic function ( ).
- It is typically a common sigmoid curve ("S" shape).

Briefly, in Logistic Regression an input is passed to the sigmoid function which results in a 'probability'. Logistic Regression algorithm serves the purpose of finding most workable parameter θ, for $h\theta(x) = \sigma(\theta TX)$, in a way that the algorithm can successfully predict the class of each case.

3) K-Nearest Neighbors

An algorithm used for supervised learning, i.e., with a known target variable is KNN (K-Nearest Neighbors). Initially the data is 'trained' with respect to its corresponding data points. To determine class of an unknown data point, the 'K' nearest points are considered as the scale for classification. For example, consider the plot in figure.3, where each yellow dot represents Class A and purple dot represents class B. If we wish to predict Class of an unknown variable i.e., a star, how many nearest neighbors should be taken into consideration to classify it. This is the value of value, number of neighbors, which holds great importance in the model.

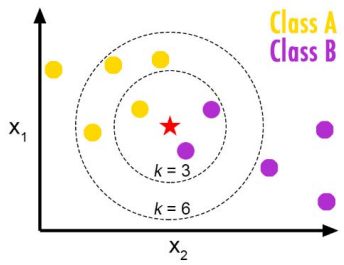


Figure 3. KNN Concept

4) Support Vector Machine

The mapping or tracing data on to a high-dimensional feature (3-D) space is how Support Vector Machine works, by categorizing the data points into related categories

(irrespective of linear separability of data). A separator separates data points into the categories, followed by transformation of data on to a hyperplane. Consequently, characteristics of new data could be used to predict the category to which it belongs.

Machine Learning Algorithm	Accuracy (100%)
Decision Tree	94.87
Logistic Regression	89.00
K-Nearest Neighbors	87.17
Support Vector Machine	82.05

Table 2. (a) Performance measure for various algorithms used in the study

IV. Results and Discussion

Table 2 shows performance of the four machine learning algorithms used in the study. Figure 4 shows visualization of Decision tree. For Logistic regression and Support vector machine classifiers results are studied through Confusion Matrix. Results of classification are represented in the form a matrix, commonly known as confusion matrix [2].

Figure 5 shows Confusion Matrix for Logistic Regression. Out of total 39 (23+4+4+8) cases, in the first row, real 27 (23+4) cases are PD Patient, and according to prediction, 23 are PD Patients. This explains 23 correct predictions out of 27 predictions with 4 incorrect predictions. Similarly, in the second row, out of total 39 cases, real 12 (4+8) cases are healthy and according to prediction, 8 are healthy. This explains 8 correct predictions out of 12 predictions with 4 incorrect predictions.

In the study of K-Nearest Neighbors, the number of neighbors is experimented from k=1 to k=10. The best result is obtained at 7 neighbors. Figure 6 shows the graph of accuracy for each number of neighbors.

Figure 7 shows Confusion Matrix for Support vector machine. Out of total, 39 (5+7+0+27) cases, in the first row, real12 (5+7) cases are PD Patient, and according to prediction, 5 are PD Patients. This explains 5 correct predictions out of 12 predictions with 7 incorrect predictions. Similarly, in the second row, out of total 39 cases, real 27 (27+0) cases are healthy and according to prediction, 27 are healthy. This explains all correct predictions out of 27 predictions with zero incorrect predictions. Figure 8 shows visualization of Support vector machine wherein, initially a hyperplane is created which holds data points of different classes with a class distinguisher collecting indices (support vector indices) of all the nearest data points to the separator. Classes of nearest data points are identified and collected. Hence a final class distinguisher is obtained.

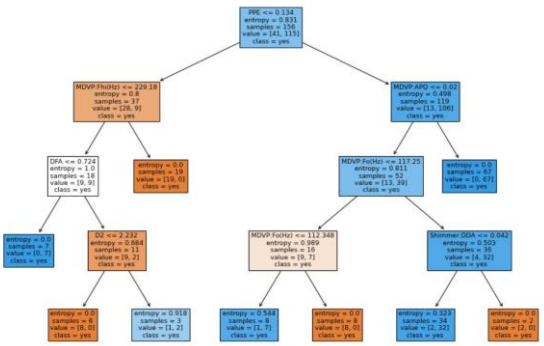
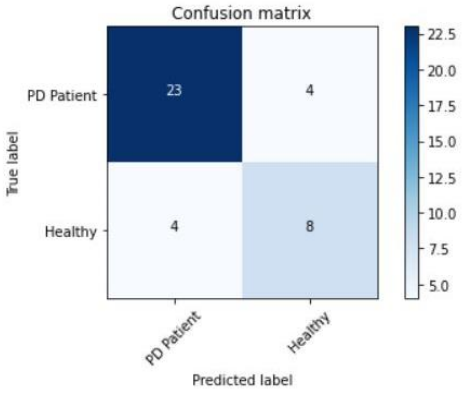


Figure 4. Decision Tree Visualization



	precision	recall	f1-score	support
0	0.67	0.67	0.67	12
1	0.85	0.85	0.85	27
accuracy			0.79	39
macro avg	0.76	0.76	0.76	39
weighted avg	0.79	0.79	0.79	39

Figure 5. Confusion Matrix and classification report- Logistic Regression

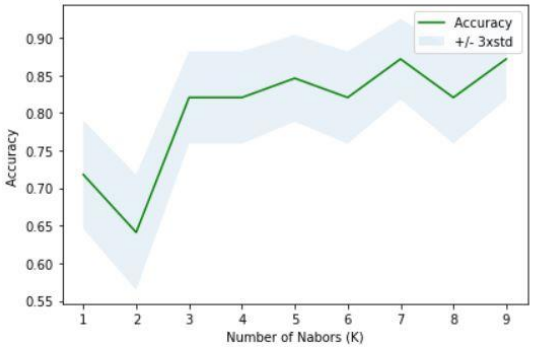
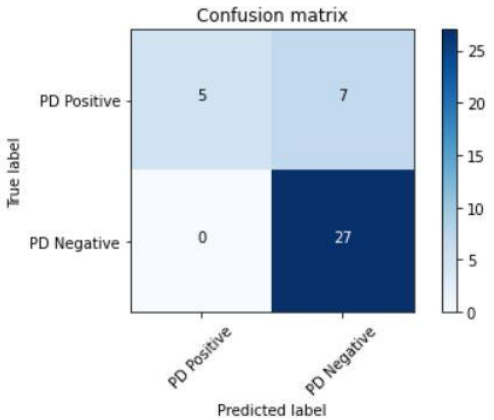


Figure 6. KNN Accuracy graph



	precision	recall	f1-score	support
0	1.00	0.42	0.59	12
1	0.79	1.00	0.89	27
accuracy			0.82	39
macro avg	0.90	0.71	0.74	39
weighted avg	0.86	0.82	0.79	39

Figure 7. Confusion Matrix and classification report-Support Vector Machine

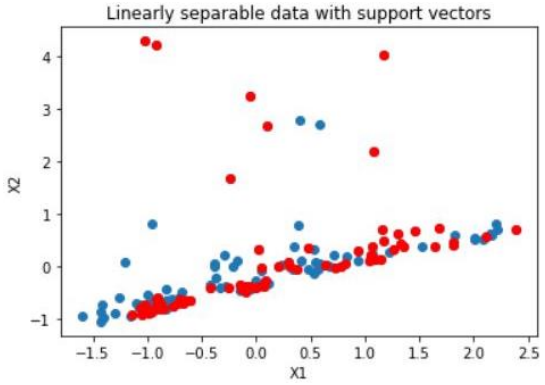


Figure 8. Support Vectors Visualization

V. CONCLUSION

Parkinson Disorder is a disorder whose diagnosis is complex because of its symptoms similar to other disorders. Moreover, the lack of awareness increases the vulnerability of the patient's health. This often leads to misdiagnosis of the disorder. The diagnosis of Parkinson's Disease is not a straight-away-process which implies, a single test like ECG or blood test alone cannot determine PD in a person. Doctors need to study the patient's medical history followed by some neurological tests. With the high rate of misdiagnosis of PD, due to indefinite tests, leads to a crisis. Technology such as Data Science and Machine Learning tend to utilize this crisis as an opportunity to make diagnosis and treatment of PD patients easy.

In the study, it was observed that Decision Tree gives the best results with accuracy of 94.87% which is the highest. Logistic Regression with 89% followed by K- Nearest Neighbors and SVM with accuracies 87.17% and 82.05 % respectively.

Thus, medical history of people holding values of Central Nervous System related features can be used to predict Parkinson's Disorder at early stages. Since PD has not been found with any cure till date, its early detection makes early diagnosis possible.

REFERENCES

[1] An Improved Approach for Prediction of Parkinson's Disease using Machine Learning Techniques, Kamal Nayan Reddy Challa, Venkata Sasank Pagolu, Ganapati Panda, Babita Majhi, arXiv:1610.08250v1 [cs.LG] 26 Oct 2016

[2] Predication Of Parkinson's Disease Using Data Mining Methods: A Comparative Analysis Of Tree, Statistical, And Support Vector Machine Classifiers [Article in Indian Journal of Medical Sciences · June 2011 DOI: 10.4103/0019-5359.107023 · Source: PubMed]

[3] Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings Betul Erdogdu Sakar, M. Erdem Isenkul, C. Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun, Article in IEEE Journal of Biomedical and Health Informatics · July 2013

[4] Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers, June 2011, Indian Journal of Medical Sciences 65(6):231-42, DOI: 10.4103/0019-5359.107023

[5] Detecting Parkinson's Disease with Machine Learning; medium.com/analytics-vidhya/detecting-parkinsons-disease-with-machine-learning-44c17208afce

[6] Supervised Learning with scikit-learn from DataCamp: datascience103579984.wordpre-ss.com/2019/08/25/21-supervised-learning-with-scikit-learn-from-datacamp/

[7] Decision Tree reference- https://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/

[8] The 7 Key Steps To Build Your Machine Learning Model; analyticsindiamag.com/the-7-key-steps-to-build-your-machine-learning-model/

[9] Datasource: archive.ics.uci.edu/ml/datasets/Parkinsons/about.html#:~:text=MDVP%3AFo(Hz)%20%2D,Hz)%20%2D%20Maximum%20vocal%20fundamental%20frequency

[10] Visualize a Decision Tree in 4 Ways with Scikit-Learn and Python: https://mljar.com/blog/visualize-decision-tree/

[11] Confusion Matrix in Machine Learning: https://www.guru99.com/confusion-matrix-machine-learning-example

[12] Logistic Regression in Python: https://realpython.com/logistic-regression-python/

[13] Shakya, Subarna, and Lalitpur Nepal. "Computational Enhancements of Wearable Healthcare Devices on Pervasive Computing System." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 02 (2020): 98-108.

[14] Smys, S. "Survey on accuracy of predictive big data analytics in healthcare." Journal of Information Technology 1, no. 02 (2019): 77-86.

[15] How to design yor data science project- https://medium.com/ml-research-lab/data-science-methodology-101-2fa9b7cf2ffe

[16] A deep desertion of data science: related issues and its applications- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8701415