# SEMINAR REPORT ON PARKINSON'S DISEASES CLASSIFICATION USING DATA MINING ALGORITHMS

**BY**

**CHAITRALI BAKSHI**

**T.Y.B.Tech Computer Engineering**

**Roll No.- 3305, Division – A**

**2020-2021**

**SEMINAR GUIDE**

**NILOFER ATTAR**

**MKSSS's**

**Cummins College of Engineering for Women,
Pune (An Autonomous Institute Affiliated to
Savitribai Phule Pune University)**

# CERTIFICATE

This is to certify that **Ms. CHAITRALI BAKSHI** has satisfactorily completed the seminar on

***PARKINSON’S DIDEASE CLASSIFICATION USING DATA MINING ALGORITHMS***

in the partial fulfillment of her term-work (Seminar) as a part of syllabus for T.Y.B.Tech. Computer Engineering for the Academic Year 2020-2021 as prescribed by MKSSS’s Cummins College of Engineering for Women, Pune (An Autonomous Institute Affiliated to Savitribai Phule Pune University)

**Prof. NILOFER ATTAR**
**Dr. Mrs. M. B. Khambete Principal**
**MKSSS’s**
**Cummins College of Engineering for Women, Pune**
**(An Autonomous Institute Affiliated to Savitribai Phule Pune University)**

# PLAGIARISM CERTIFICATE

This is to certify that **Ms. CHAITRALI BAKSHI** has satisfactorily completed the seminar report on

***PARKINSONS DISEASE CLASSIFICATION USING DATA MINING ALGORITHM***

and the data present in this report is found to be _______ % plagiarism free.
This seminar report is towards the partial fulfillment of her term-work (Seminar) as a part of the syllabus for T.Y.B.Tech. Computer Engineering for the Academic Year 2020-2021 as prescribed by MKSSS's Cummins College of Engineering for Women, Pune (An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Name of the Student and sign Prof. Nilofer Attar

Seminar Guide

# TABLE OF CONTENTS

# ABSTRACT

Data mining is evolving day by day. With the world moving towards digital era, streams like data mining, machine learning etc have been emerging and growing. Data mining algorithm helps organize problems help find the most accurate, appropriate and efficient solutions. Data mining has application in many fields like healthcare, robotics etc. It has major application in Parkinson's disease classification. Parkinson's disease has been a growing concern all around the world. Last year 1 million cases were registered in USA.

The paper also involves feature selection which can be applied on the dataset. Feature selection is a important pre-processing step which reduces complexity, by removing unnecessary features.

This paper helps identify the data mining algorithm which can help in improvement of early prediction of the disease. The process of building an entire algorithm is given. The main purpose of the paper is to find the most accurate, efficient algorithm which can help build a model which will try to achieve almost 100 % accuracy. We will get a quick review about algorithms and outline for building this application.

# CHAPTER 1

# INTODUCTION

## 1.1INTRODUCTION

Data mining is surely an upcoming field and has high demand because it turns huge amounts of data into useful information and knowledge. The applications of data mining are gigantic. One of the most important application is in healthcare. Data mining is used from prediction of medicines, measuring how effective are certain treatments, early-stage detection of various diseases and in many areas. One application of data mining which can be extremely useful is for classification of Parkinson's disease. It is a neurodegenerative disease which can last for years or can be lifelong. Treatment can help reduce the symptoms but the disease is incurable.

### 1.1.1 NEED

After Alzheimer's disease Parkinson's disease is the second most common neurodegenerative disease. Cause of the disease is unknown and almost one million people in USA are living with Parkinson's disease. It has no cure and therefore using data mining algorithms for its classification can help us predicting the disease at an early stage. Cure may not be possible now but early detection can decrease the advancement of the disease and bring its effect under control. When the PD is in early stage it is quite mild and can be controlled by medicines but as the severity increases surgical therapy comes into action.

Therefore, differentiating stages of pd is very important. Usually a team of neurologist, neurosurgeon, psychiatrist, neuropsychologist, rehabilitation specialist to decide whether the person needs surgical help or not. There a data mining system can help solve the above-mentioned problems.

### 1.1.2 DISADVANTAGES

1.Can be expensive to collect all the data and process it.

2.There is no scope for any sort of mistake. The algorithm should be hundred percentage accurate because any inaccuracy can lead to measure loss in the patient's treatment.

3.If the app is not user friendly it will be difficult to learn and use and would lead to more mistakes.

### 1.1.3 HOW TO OVERCOME

Accuracy can be increased by training the model.

The app can be made in user friendly way.

# CHAPTER 2

# LITERATURE

1. A Novel Model for Classification of Parkinson's Disease: Accurately Identifying Patients for Surgical Therapy

   Parkinson's disease is an incurable neurodegenerative disorder which has become quite common amongst people around the age of the 50. The cause of the disease is unknown but surgery can help patients reduce the symptoms of the disease to a great level. The aim of this particular paper is to find patients who are in need of a surgical therapy. The data was of the patients was collected from the Parkinson's Progressive Markers Initiative. It had 1080 subjects and was quite resourceful.

   Decision tree, SVM, Naïve bayes and Multilayer Perceptron (MLP) were the classification algorithms applied on the dataset. Only 60% of the features were selected as they gave equal accuracy as compared to all attributes. MLP algorithm gave the most accurate results.

   The model was able to reach 98.13% accuracy rate.

2. PREDICATION OF PARKINSON'S DISEASE USING DATA MINING METHODS: A COMPARATIVE ANALYSIS OF TREE, STATISTICAL, AND SUPPORT VECTOR MACHINE CLASSIFIERS GEETA YADAV, YUGAL KUMAR, GADADHAR SAHOO

   The aim of the article is to predict Parkinson's disease based on speech articulation difficulty symptoms. Three data mining algorithms are used: tree classifies, SVM and Logistic regression (LR). LR gave the most promising results.

# CHAPTER 3

# Description

## 3.1 WHAT IS PARKINSONS DISEASE

It is a progressive movement disorder. It is also chronic in nature which means it gets worse with time. Normal body functions like breathing, balance, movement weaken and heart also becomes weak. It is caused to death of neurons in brain. Neurons that produce dopamine asre also a part of these dying neurons Dopamine acts like chemical messenger between neurons. It sends message to part of brain which is responsible for movement and therefore a person suffering with pd experiences tremors and lack body control. It is usually observed in people around the age of 50.

**The advancement of pd takes place in 5 stages:**

**First stage:**

The person experiences small tremors on one side of the body only. Other symptoms might be stage in body posture or change in walk, etc.

**Second stage**:

The symptoms start getting worse. It starts to affect both sides of the body. The person can live by themselves but the speed doing daily chores reduces.

**Third stage**:

It is called as the mid stage, the symptoms are quite evident thorough their body walk, expressions etc.

**Fourth stage:**

The person may able to stand by themselves but cannot live independently.

**Fifth stage:**

This is the most advanced stage; the person cannot walk or do any chores by themselves. They experience leg stiffness, hallucinations etc.

## 3.2 WHAT IS CLASSIFICATION

Classification is the process of dividing the given data into classes based on certain attributes so that the new data can be grouped with the predefined class. There various classification algorithms which can help in classification of Parkinson's disease.

## 3.3 STAGES FOR CLASSIFICATION

### 3.3.1 PARKINSONs DATASET:

A dataset is collection of information which is treated as a single unit. Parkinson's dataset should contain records of patients suffering through the disease. Records of their voices, symptoms etc can be collected and the dataset can be divided into two parts: testing data and training data.

Example: **Table 1. Characteristic Features of Parkinson Dataset**

| Feature Number | Feature Name | Description |
|---|---|---|
| 1 | MDVP:<br>Fo (Hz) | Average vocal fundamental Frequency |

| | | |
|---|---|---|
| 2 | MDVP:<br>Fhi (Hz) | Maximum vocal fundamental frequency |
| 3 | MDVP:<br>Flo (Hz) | Minimum vocal fundamental frequency |
| 4 | MDVP:<br>Jitter (%) | Kay Pentax MDVP jitter as percentage |
| 5 | MDVP: Ji tier (Abs) | Kay Pentax MDVP absolute jitter in microseconds |

## 3.3.2: Pre-processing:

The dataset has to be analysed properly to see if it contains any missing values or any redundant values. WEKA filter can be used to overcome the problem of missing issues. For the problem of redundant data new records can be replaced with old records by erasing old records.

## 3.3.3 FEATURE SELECTION:

If we reduce the attributes to identify the most important feature that can contribute to classification it will increase the efficiency . Characterization of data is more efficient with less features. Selecting the most relevant feature amongst the other features can increase time efficiency and performance. If the irrelevant features are not removed, it will have negative effects on task. Information gain technique can be used as it calculates information gain of a particular feature with respect to class.

### 3.3.4 Classification process:

Various data mining algorithms are applied after pre-processing the data. The most efficient ones are

1. **SVM (Support Vector Machine):**

It is a supervised learning algorithm and is closely related to neural networks. In SVM the data is optically divided into two separate categories.

2. **Random Tree (Rnd Tree):**

   It is a machine learning algorithm and it consists of a lot of decision trees. It does ensemble classification. Ensemble methods averages predictions of other independent base models and makes prediction on the average.

3. **Naïve Bayes:**

   It is known to be a probabilistic classifier (gives probability of prediction) . It is based on probability models that incorporate strong independence assumptions.

4. **Logistic regression:**

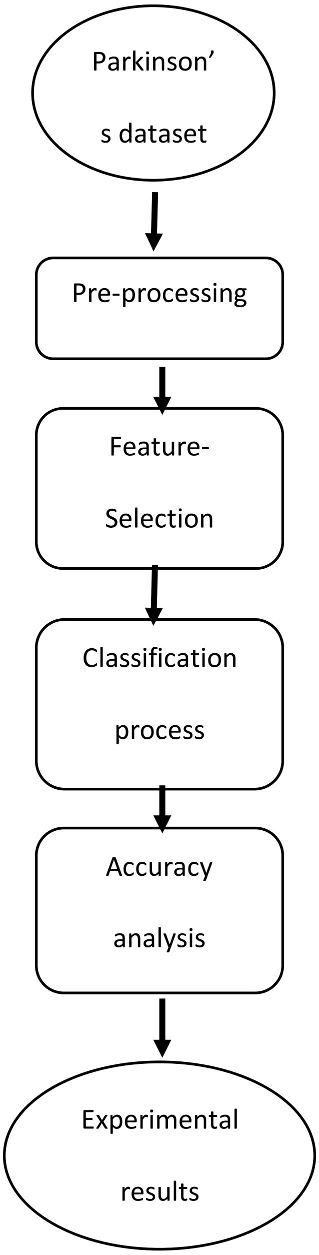
   It is a statistical analysis method which predicts a dependent data variable by analysing the relationship between existing independent variable.

### 3.3.4 Accuracy analysis:

After applying all the above-mentioned algorithm, we now need to check which algorithm gives the most accurate results. Hundred percentage accuracy is not possible but the algorithm which is the most accurate and be used and further work can be done on it to increase accuracy. To summarize the performance of classification algorithms confusion matrix is used. They give correct comparisons of values like True positives, False positives, True negatives , False negatives .

## 3.3.5 Experimental results:

Several experiments can be performed to test which algorithm is provides almost 100% accuracy. We can implement each algorithm induvially and test which one gives the most accurate results. The algorithm should also be user friendly and easy to implement. After finding out the most suitable algorithm we can also train it again to improve accuracy.



# CHAPTER 4

# CLASSIFICATION ALGORITHM

## 1. SVM (Support Vector Machine):

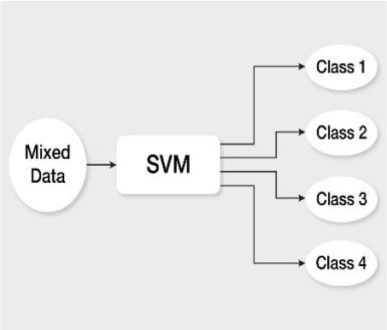
SVM is a linear separator that is it divides the data into classes using a hyperplane. SVM uses less computational power and provides high accuracy. Its job is to find a hyperplane in N-dimensional space that distinctly classifies data points.it uses vector support points which (collection of points with the lowest distance from hyperplane). SVM models are extremely close to neural networks. Thus, SVM is highly used for various healthcare related algorithms.

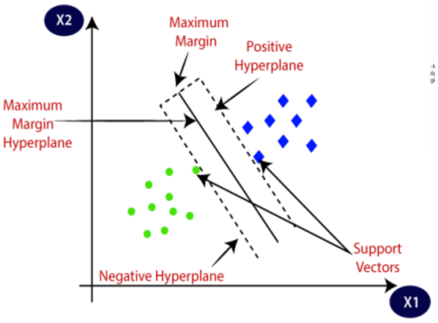
Advantages:

1. Its uses the kernel trick: To make the data easy to classify non-linear data is projected into higher dimension where it can be easily divided using hyperplane.
2. The model is known to be stable as any small changes in data does not affect hyperplane.
3. As it uses support vector it is quite memory efficient.

Disadvantages:

1. Choosing the appropriate kernel for svm is a very tedious and complex task.
2. Using a high dimension kernel can lead to formation of many support vectors which can drastically reduce the training speed.
3. Requires feature scaling.





2. Random forest:

They are as easy to create as single tree models, however their accuracy level is quite high compared to a large single tree model. They can be created using a setting a control button and giving DTREG a direct order to create a decision tree forest model. Out of bag data rows are being used in random tree for validation of model. It uses both bagging and random variable selection for tree building. After the forest is formed test instances are filtered down each tree. After this prediction of respective classes is made. Its error rate is dependent on strength of each tree and co-relation between any two trees.

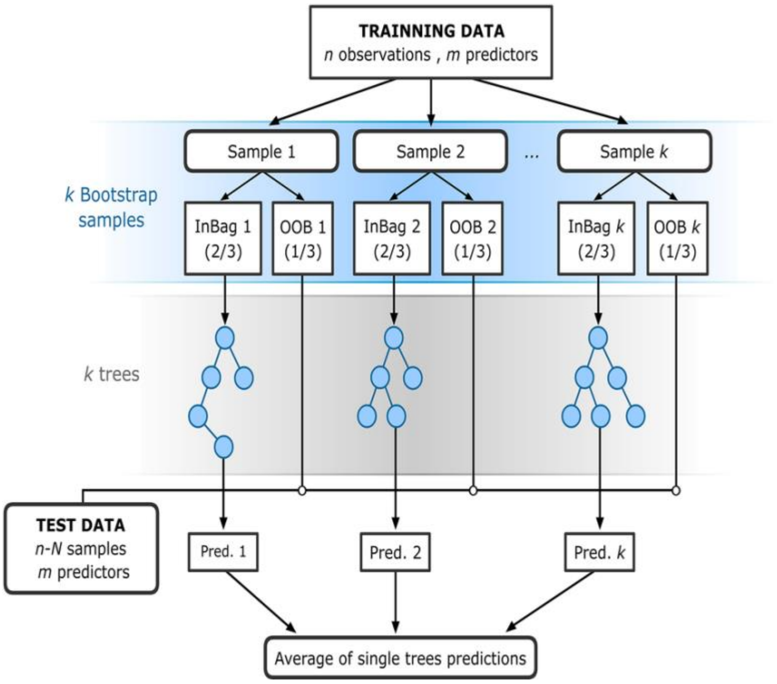
ADVANTAGES:

1.It improves accuracy by reducing overfitting in decision trees.

2. It uses a rule-based approach therefore normalization is not allowed,

DISADVANTAGES:

1. They are not very interpretable.
2. The size of the tress can take up a lot of memory especially for large datasets.



### 3. Naïve bayes:

It is a supervised learning algorithm that follows class condition independence. That is effect of an attribute value of a given class is independent of the values of other attributes.

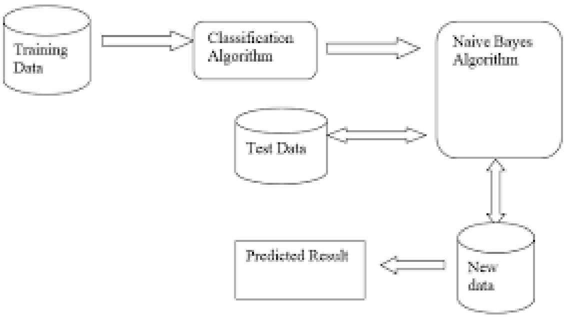
It is useful in reducing computational costs. It is easy to build model and is useful in large datasets. It classifies data by maximising.

ADVANTAGES:

1. It is extremely useful for solving multiclass prediction problems.
2. The build process of naïve is paralysed.

DISADVANTAGES:

1. It assumes all the attributes are mutually independent which is not possible in real life cases.
2. Zero frequency occurs a lot of times. for example if categorical variable has a category in test data set, which cannot be found in training dataset, then model is given 0 probability and hence the prediction is hampered.



**4. Logistic regression:**

It is a classification model where the class label or target is categorical. For example, the algorithm will provide two categories here: whether the person has disease or doesn't.

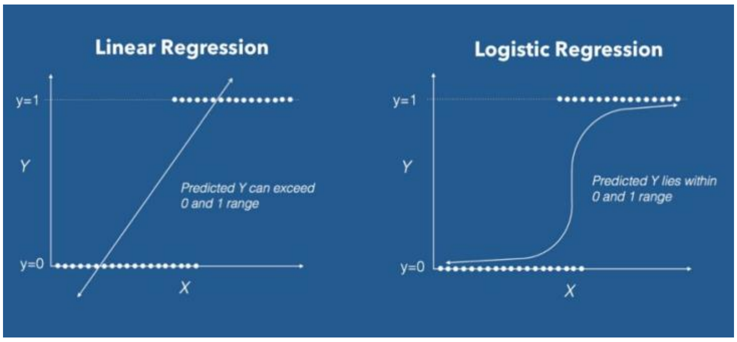
It is known to produce good results in determining chronic diseases low incidence and simple clinical predictors. It is very similar to linear regression, only it doesn't provide a yes (1) and no(0) answer, it provides the most probable answer between 0 and 1. It can work on both continuous and discrete datasets.

**ADVANTAGES:**

1. It is quite easy to interpret and implement. It is also very efficient algorithm to train.
2. It tells how relevant a predictor is and also its direction of association.

## DISADVANTAGES:

1. It assumes there's linearity between dependent and independent variables but it is not possible for real life data to be always linearly separable.
2. It leads to overfit if number of observations are less than number of features.



# CHAPTER 5

# CONCLUSION:

Parkinson's disease is a neurodegenerative disorder which is incurable. It is extremely difficult to identify Parkinson's in the very first stage as symptoms only become observable in the middle and end stage. The classification algorithms can be used in various applications. With
accurate dataset and feature selection we can use all the above-mentioned algorithms and test which has the highest accuracy rate. We can then perform experiments and determine which algorithm gives the highest accuracy. Implementation of this application can be useful to the patients suffering from Parkinson's disease and help reduce symptoms by identifying it in early stage.

FREQUENTLY ASKED QUESTION?

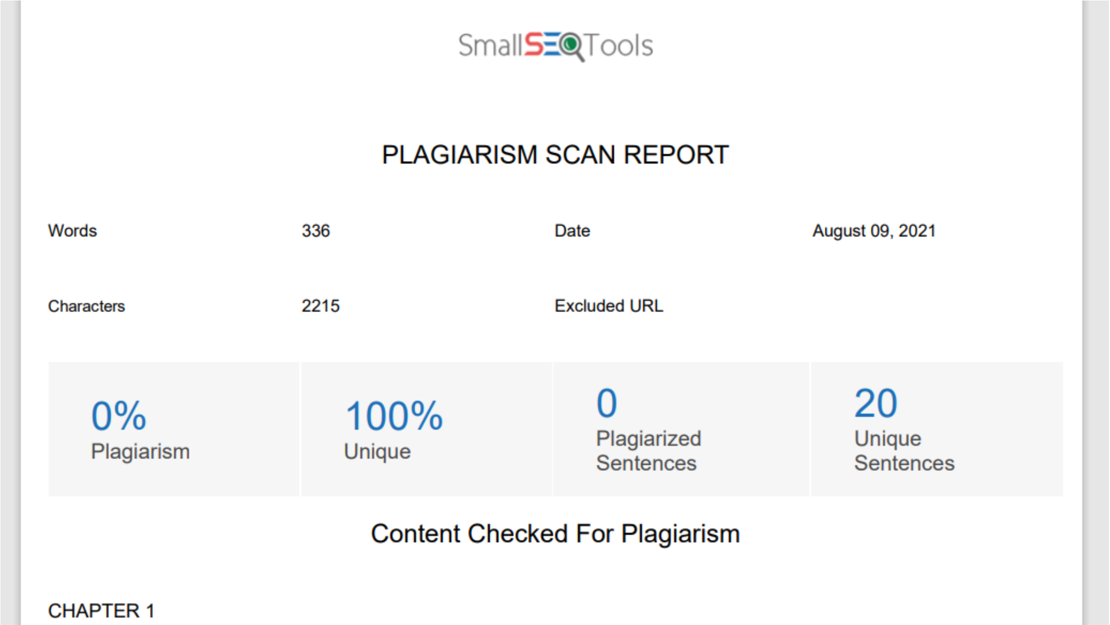
1. What are the challenges faced ?

   The collection of data , softwares used are quite expensive and therefore needs a lot of investment, Also even one wrong predicton can lead to dangerous effects on the person so it has to be accurate.

2. How does it impact healthcare?

   As a huge team of doctors is required , this algorithm can help can give quick decision . If the disease is identified at early stage , it can help reduce symtoms therefore it has a good impact on healthcare.

# PLAGARISM REPORT



SmallSEOTools

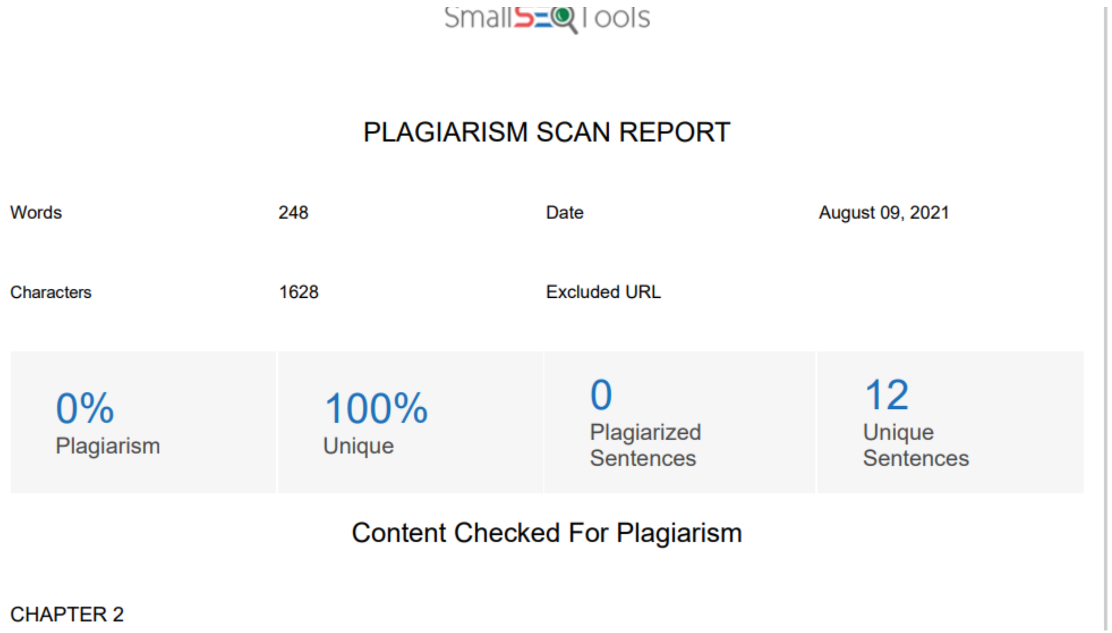
PLAGIARISM SCAN REPORT

| | | | |
|---|---|---|---|
| Words | 336 | Date | August 09, 2021 |
| Characters | 2215 | Excluded URL | |

| | | | |
|---|---|---|---|
| 0%<br>Plagiarism | 100%<br>Unique | 0<br>Plagiarized Sentences | 20<br>Unique Sentences |

Content Checked For Plagiarism

CHAPTER 1



PLAGIARISM SCAN REPORT

| | | | |
|---|---|---|---|
| Words | 248 | Date | August 09, 2021 |
| Characters | 1628 | Excluded URL | |

0% Plagiarism

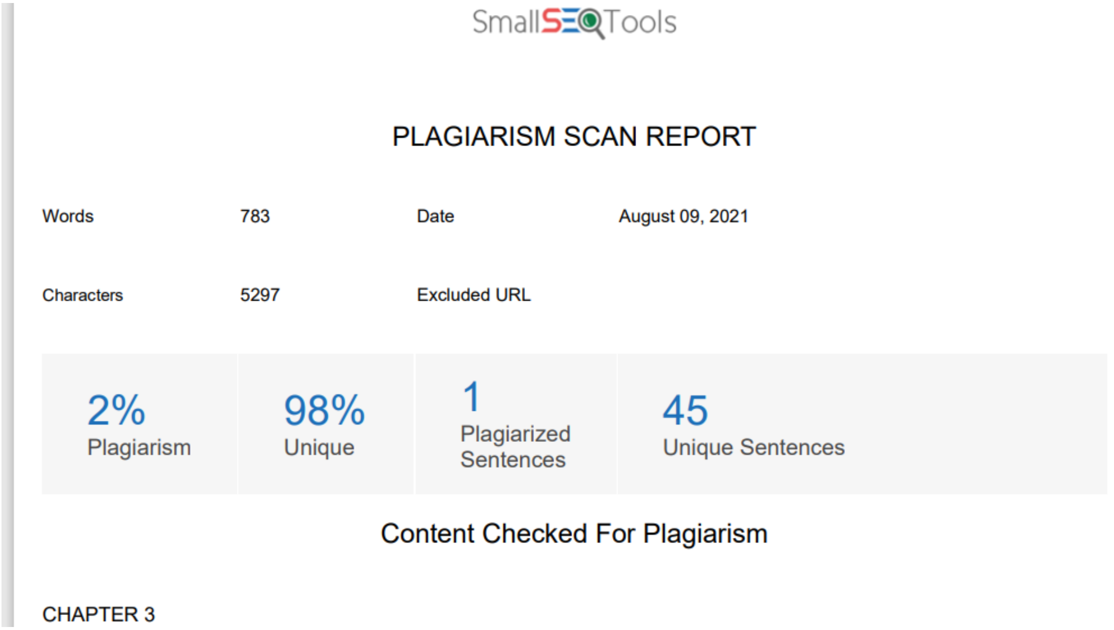
100% Unique

0 Plagiarized Sentences

12 Unique Sentences

Content Checked For Plagiarism

CHAPTER 2



PLAGIARISM SCAN REPORT

| | | | |
|---|---|---|---|
| Words | 783 | Date | August 09, 2021 |
| Characters | 5297 | Excluded URL | |

2% Plagiarism

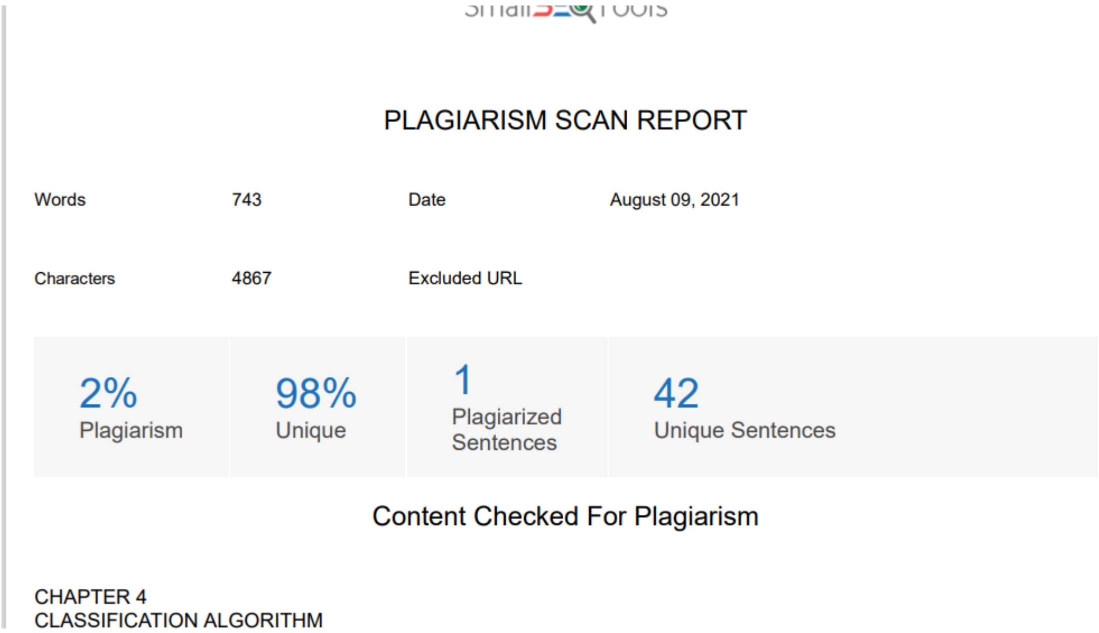
98% Unique

1 Plagiarized Sentences

45 Unique Sentences

Content Checked For Plagiarism

CHAPTER 3



PLAGIARISM SCAN REPORT

| | | | |
|---|---|---|---|
| Words | 743 | Date | August 09, 2021 |
| Characters | 4867 | Excluded URL | |

| | | | |
|---|---|---|---|
| 2%<br>Plagiarism | 98%<br>Unique | 1<br>Plagiarized Sentences | 42<br>Unique Sentences |

Content Checked For Plagiarism

CHAPTER 4
CLASSIFICATION ALGORITHM

# REFERENCES

| | |
|---|---|
| Researchgate | Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers |
| www.hindawi.com | Parkinson's Disease Diagnosis in Cepstral Domain Using MFCC and Dimensionality Reduction with SVM Classifier |
| https://www.frontiersin.org/ | Parkinson's Disease Detection Using Isosurfaces-Based Features and Convolutional Neural Networks |
| scholarspace.manoa.hawaii.edu | A Novel Model for Classification of Parkinson's Disease: Accurately Identifying Patients for Surgical Therapy |
| scirp.org | Parkinson's Disease Diagnosis: Detecting the Effect of Attributes Selection and Discretization of Parkinson's Disease Dataset on the Performance of Classifier Algorithms |

| | |
|---|---|
| https://www.ijcaonline.org/ | ANN based Data Mining Analysis of the Parkinson's Disease |