Design and Analysis of Data Mining Based Prediction Model forParkinson's disease

Chandrashekhar Azad[1], Sanjay Jain[2], Vijay Kumar Jha[3]
[1]Research Scholar, Department of CSE, Birla Institute of Technology, Mesra (Ranchi)
[2]Director, MICA Educational Company, Opposite Ranchi Club Gate, (Ranchi)
[3]Associate Professor, Department of CSE, Birla Institute of Technology, Mesra (Ranchi)

Abstract:

Purpose: The purpose of this research paper is to develop a prediction model for Parkinson's disease. There are many symptoms that lead to Parkinson's disease such age- environmentalfactor, trembling in the legs, arms, hands, impaired speech articulation and production difficulties. In this research paper speech articulation of Parkinson's disease affected people is considered for model formation and analyzes the model based on the symptom of disease.

Methods: In proposed prediction model tree based classification model decision tree, ID3 and decision stumps are used for training and testing the effectiveness of proposed prediction model. Here we also applied K –fold cross validation technique for true prediction so that each record is sued for training and testing.

Results : In proposed model decision tree based our prediction model provide accuracy 85.08%, classification error 14.92%, ID3 provide accuracy 75.33% ,classification error 24.67% and decision stumps based model proved accuracy 83.55% and classification error 16.45%.

Conclusion: Proposed model based on Decision tree provide best result in comparison to other in terms of parameters accuracy and classification error.

Keyword: Parkinsons, Data mining, Decision tree, ID3, Decision Stumps.

1. Introduction:

Neurons are the basic building blocks of thenervous system which incorporate the spinal brain and cord. Neurons normally dont replaces or reproduce themselves. When neurons become damaged or die they cannot be swapped by the body. Neurodegenerative diseases are Parkinson's, Alzheimer's, and Huntington's disease [1, 2]. Today a lot of people are affected from neurodegenerative diseases like Parkinson disease, Alzheimer's disease, Arthritic disease, Prion disorders, Corticobasal degeneration, Progressive supranuclear palsy, Dementia with Lewy bodies, Huntington's disease , Motor neurone diseases , Huntington's Disease, Spinal muscular atrophy , Motor neurone diseases , etc. Neurodegenerative diseases are debilitating and incurable conditions that result in progressive degeneration of nerve cells,and it causes problemsinmental functioning and movement. In Scotland 120 to 230 people in per 100,000 are affected with Parkinson's disease, While population of the Scotland remains stable. In the next 25 years Parkinson's disease affected people may increase by 25-30 % [3]. Parkinson's disease is the 2nd most neurodegenerative diseases. It is identified by progressive loss in control of muscle, and it leads to trembling ofhead and limbs, and at restslowness,impaired balance and stiffness. Day by day symptoms worsen and the affected people may have difficulty in walking,talking and also may in complete simple tasks. In USapproximate 1 million people affected with Parkinson's disease and around theworld approximate 5 million people affected with it. Most of the 60 years or olderage people are affected with Parkinson's disease, it is found approximately 1% in 60 age group and approximate 4% in the age group of 80 years. Since overall life expectationrising and maythe number of people with Parkinson's disease will increase in the near future. Adult-onset PD(Parkinson's Disease) is common, early-onset is in 21-40 years, and juvenile-onset PD before age 21. The Parkinson's disease is date back as far as approximate 5000 BC and Indian civilization termed it as the Kampavata and the they use the seeds of a plant that contain therapeutic levels for treatment. Parkinson's disease was discovered by James Parkinsonin 1817 as "shaking palsy"[4]. Near about 1 million adults in USA are affected withPD and over 60,000 people diagnosed every year. In USA, According to Parkinson's disease Foundation, $25 billion annually expend in thePD and average annual medication costs is in the range of $2,500 to $10,000. In United Kingdom 1 in every 500 people is affected with PD and about 10 million people in world. A male have 50% higher risk than a female in developing Parkinson's disease.In the many of the cases, the symptoms are appear after the age 50 and approx 4-5% of cases in younger than 40 years age group[4,5,6].

Figure 1. Parkinson's Diseases Affected

Symptoms of Parkinson's Disease?

- slowness of movement, slowed motion(Bradykinesia)
- Resting tremor
- muscle stiffness
- Posture and balance
- The arms may not swing when walking
- Swallowing difficulties
- Speech problems
- Loss of facial expression

Possible complications of Parkinson's disease

- Depression
- Sleeping problem
- Urinary incontinence or retention
- Constipation
- Thinking difficulties

Risk Factors:

- Age
- Gender
- Family History
- Race and Ethnicity

2. Applications of data mining for classification:

Decision tree

Decision tree is supervised learning technique in data mining[23,24,25]and it is used for classification which maps unlabeled records to a target class based on the learned model in other word we can termed it as classification trees. In decision tree or classification tree leaf nodes represent class labels and branches represent the test outcome of the features. In the decision analysis process it is used to visually represent the decisions. The goal of decision tree classifier is to produce a predictionmodel with the historical records termed as a training set and the learned model predicts the target value of a given input variables set. A decision tree can be trained by splitting the dataset into subsets based on an attribute value test and the training process is repeated on each subset in a recursive manner. The training iscompleted when splitting of attribute remain no longer to addin the predictions.

Data format:$(x, Y) = (x_1, x_2, x_3 \ldots\ldots x_k,\ Y)$

The Y is the target class that we are trying do classify. The vector $\boldsymbol{x_i}$, $\boldsymbol{i = 1, 2}, \ldots\ k$ is set of input variables..

Attribute Selection Measures:

Attribute selection for decision tree is heuristic for selecting the splitting criterion that best separatesthe given data set into individual classes. Splitting of dataset into smaller partitions based on the outcomes of the splitting criteria, each partitions are pure. The best splitting criteria is the one that have the best value set to split. Attribute selection is alsotermed as splitting rules,sincethe splitting rules of criteria determine how the given tuples to be split.The attribute selection criteria give the ranking to each attribute based the given training set. The attributethathave the best score for the splitting,is chosen for splitting the given dataset. If splitting attribute is continuousvalued or if restricted to binary trees, then, either a split point or a splitting subset must be determined as a part of the splitting criteria. Tree node created for partition in training set is labeled with the outcome of the slitting criteria and branches are extended for each condition of the splitting criteria, and the given training tuples are partitioned accordingly. The attribute selection measures are: information gain, gain ratio, and Gini index.

ID3:

In decision tree learning, Iterative Dichotomiser (ID3) algorithm used to generate a classification tree or decision tree from the given training set. ID3 is the successor ofC4.5 algorithm, and is typically used in the machine learning or data mining for classification.

Decision stumps:

A decision stump is machine learningclassification model, it consist of a one-level decision tree. Decision stump is a decision tree with one internal node is called root node and is immediately connected to the leaf nodes or terminal node. It makesthe classification based on the value of a single input. Decision stump is also called 1-rules.For nominal attribute decision stump classifier build a stump which hold a leaf for each possible attribute value or a stump having two leaves, one corresponds to some chosen category, and the other leaf to all the other categories.

Attributes having continuous values, some threshold attribute value is choosen, and the decision stump contains two terminals for values below and above the defined threshold. In this a missing value is treated as another category.

Dataset:

The Parkinson's disease data set is taken from UCI repository. This is built up ofdata of 31 people, 23 with Parkinson's disease (PD) and rest are healthy. This data set contains the 197 instance and each attribute have the real values. The target of the dataset is to distinguish Parkinson's disease affected from those with non parkinsons diseases affected, in the dataset 0 is labeled for healthy and 1 for Parkinson's disease. The Parkinson's disease dataset was created by Max Little, University of Oxford [22].Statistics of the dataset is given below:

- Dataset Characteristics : Multivariate
- Attribute Characteristics : Real
- Number of Instances : 197
- Number of attribuite : 23
- Missing Values : None
- Area : life

Table1: Dataset description

Role	***Index***	**Attribute Name**	**Type**	**Description**
label	23	Status	binominal	Health Status P for parkinsons and H for Healthy
regular	0	Name	polynominal	ASCII subject name and recording number
regular	1	MDVP_Fo_Hz	real	Average vocal fundamental frequency
regular	2	MDVP_Fhi_Hz	real	Maximum vocal fundamental frequency
regular	3	MDVP_Flo_Hz	real	Minimum vocal fundamental frequency
regular	4	MDVP_Jitter	real	Kay Pentax MDVP jitter as percentage
regular	5	MDVP_Jitter_Abs	real	Kay Pentax MDVP absolute jitter in microseconds
regular	6	MDVP_RAP	real	Key Pentax MDVP Relative Amplitude Perturbation
regular	7	MDVP_PPQ	real	Kay Pentax MDVP five-point Period Perturbation Quotient
regular	8	Jitter_DDP	real	Average absolute difference of differences between cycles, divided by the average period
regular	9	MDVP_Shimmer	real	Key Pentax MDVP local shimmer
regular	10	MDVP_Shimmer_dB	real	Key Pentax MDVP local shimmer in decibels
regular	11	Shimmer_APQ3	real	3 Point Amplitude Perturbation Quotient
regular	12	Shimmer_APQ5	real	5 Point Amplitude Perturbation Quotient
regular	13	MDVP_APQ	real	Kay Pentax MDVP eleven-point Amplitude Perturbation Quotient
regular	14	Shimmer_DDA	real	Average absolute difference between consecutive differences between the amplitude of consecutive periods
regular	15	NHR	real	Noise to Harmonic Ratio
regular	16	HNR	real	Harmonics to Noise Ratio
regular	17	RPDE	real	Recurrence Period Density Entropy
regular	18	DFA	real	Detrended Fluctuation Analysis
regular	19	spread1	real	Non Linear measure of fundamental frequency
regular	20	spread2	real	Non Linear measure of fundamental frequency
regular	21	D2	real	Correlation Dimension
regular	22	PPE	real	Pitch Period Entropy

3. Proposed Prediction Model:

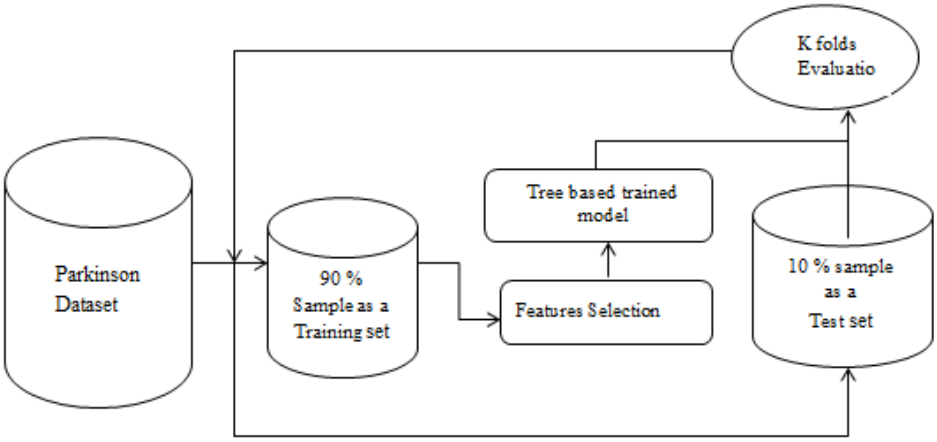

Figure 2. Proposed Prediction Model

Steps in Prediction Process:

Step 0: Start

Step 1: Load data set

Step 2: Model Creation using Training set

Step 3: Testing Model using validation set

Step 4: Performance analysis

Step 5: Selection Of best Model

Step 6: Stop

4. Experiments and Discussion:

This section describe our experiment result, Experiment is carried out using rapid miner on a system having intel i5 3rd generation processor, 4 Gb RAM, 500 GB harddisk, Windows 7 ultimate operating sytem . In the first study, three diffent types of classification algorithms are used to predict a person is either helthy or parkinsons affected. In this experiment used classification algorithms are decision tree, ID3 and decision stump and for parameter we used accuracy and classification error. In the process of classification or prediction of helathey or parkinsons we taken dataset from UCI repository and using the three well known classification method decision tree, ID3 and decision stump to trained the model . Here we used 10 – Fold cross validation technique is used to complete our experiment, reason behind choosing 10 fold cross validation for traing & testing the effectiveness of proposed prediction model is for unbiased predication. In k – fold(Here 10-fold) cross validation technique entire dataset is divided into k parts and K-1 parts are used for training and kth part is taken as testing set , this proess is repeating k times so that each part is taken as testing set.

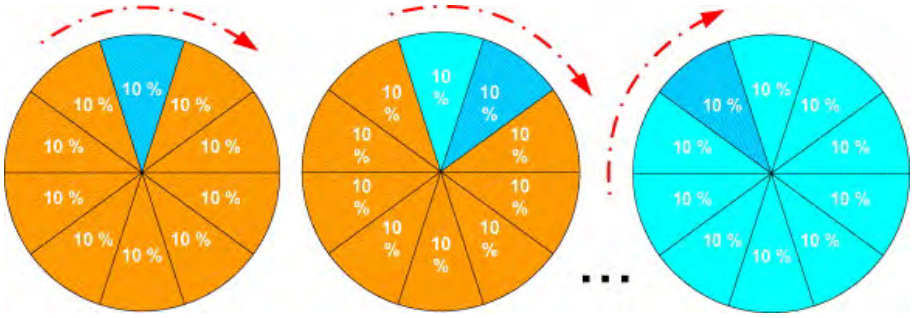

Figure 3. K fold cross validation

Table 2. Classification Results

F	Decision Tree				ID3				Random Tree			
	CM		Acc	CE	CM		Acc	CE	CM		Acc	CE
1	12	0	85.00	15.00	15	5	75.00	25.00	15	3	85.00	15.00
	3	5			0	0			0	2		
2	11	1	78.95	21.05	14	5	73.68	26.32	14	5	73.68	26.32
	3	4			0	0			0	0		
3	15	3	85.00	15.00	15	5	75.00	25.00	15	4	80.00	20.00
	0	2			0	0			0	1		
4	13	2	84.21	15.79	14	5	73.68	26.32	14	3	84.21	15.79
	1	3			0	0			0	2		
5	13	1	85.00	15.00	15	5	75.00	25.00	15	2	90.00	10.00
	2	4			0	0			0	3		
6	14	2	89.47	10.53	14	5	73.68	26.32	14	3	84.21	15.79
	0	3			0	0			0	2		
7	15	2	90.00	10.00	15	5	75.00	25.00	15	3	85.00	15.00
	0	3			0	0			0	2		
8	13	4	73.68	26.32	14	5	73.68	26.32	14	4	78.95	21.05
	1	1			0	0			0	1		
9	15	2	89.47	10.53	15	4	78.95	21.05	15	2	89.47	10.53
	0	2			0	0			0	2		
10	15	1	90.00	10.00	16	4	80.00	20.00	15	2	85.00	15.00
	1	3			0	0			1	2		
Mean			85.08	14.92			75.37	24.63			83.55	16.45

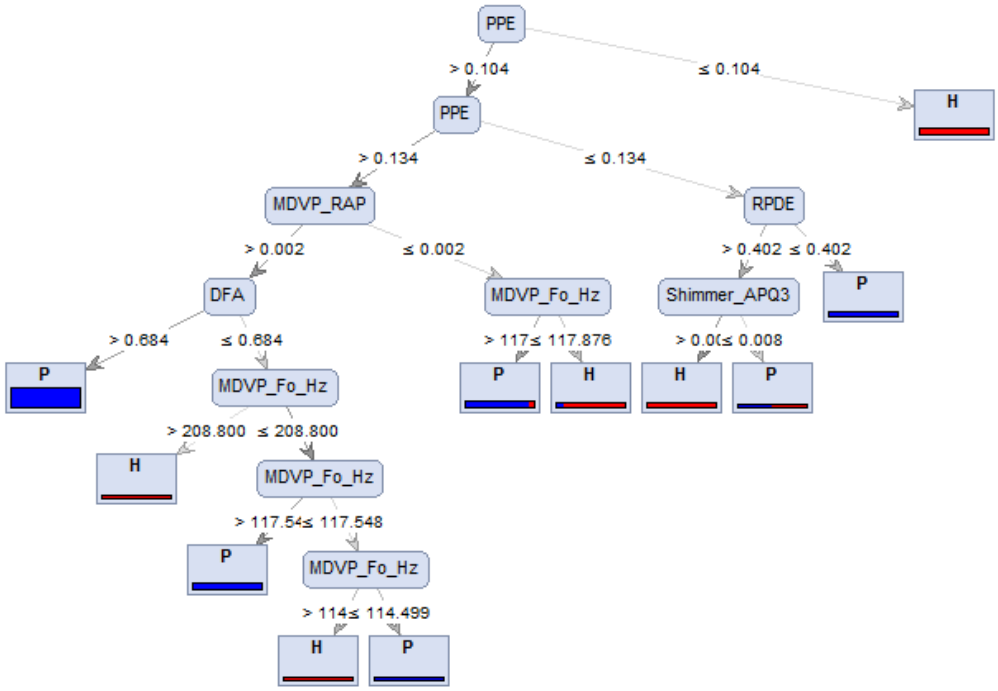

Figure 4. Decision Tree

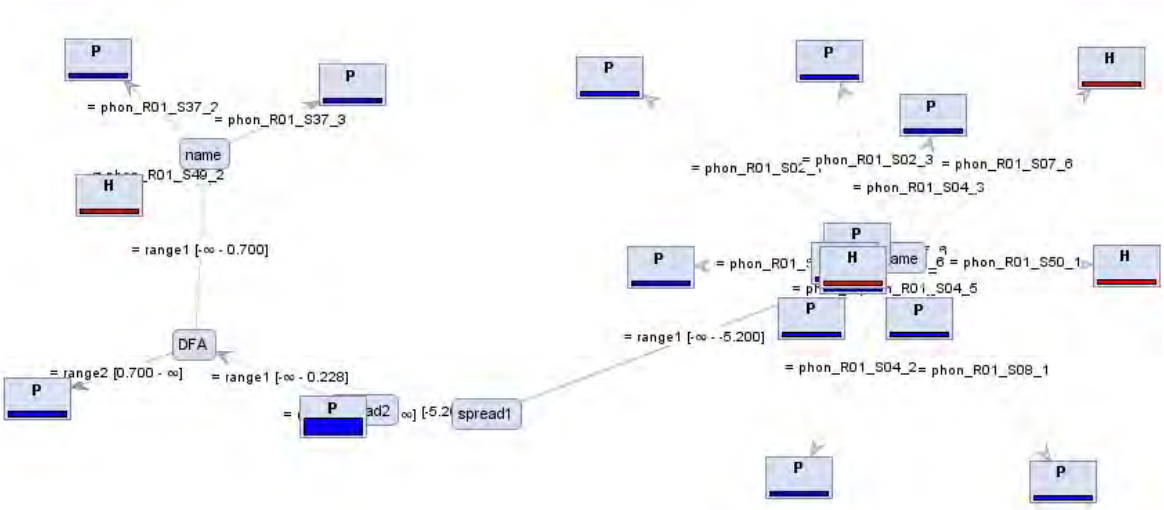

Figure 5. ID3 Tree

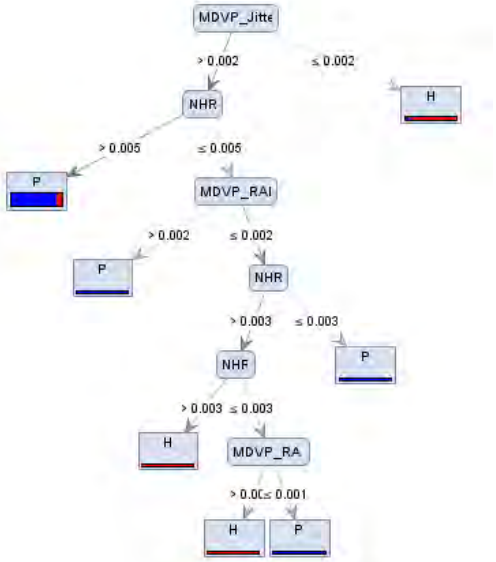

Figure 6. Tree (Decision Stump)

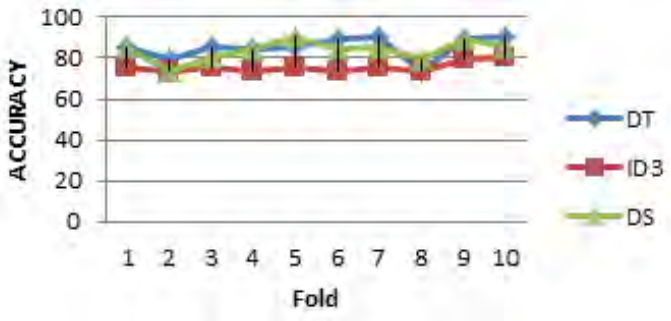

Figure 7. Accuracy by Fold

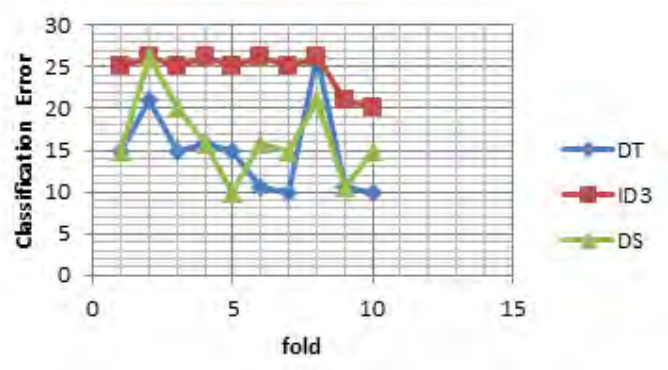

Figure 8. Classification Error by fold

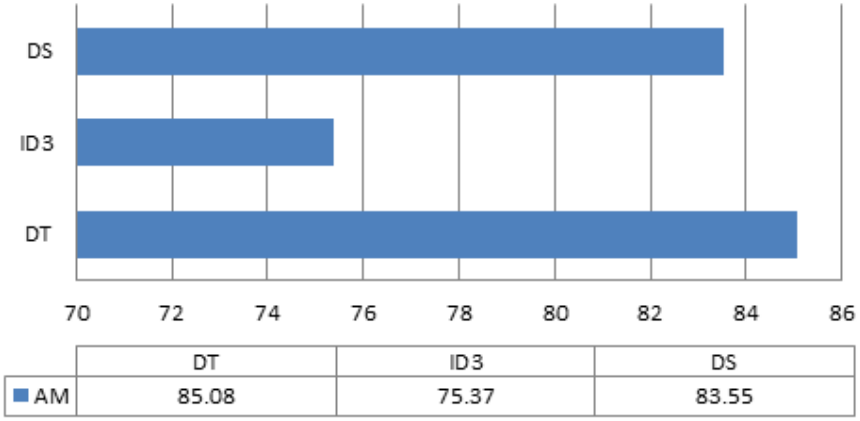

Figure 9. Mean Accuracy

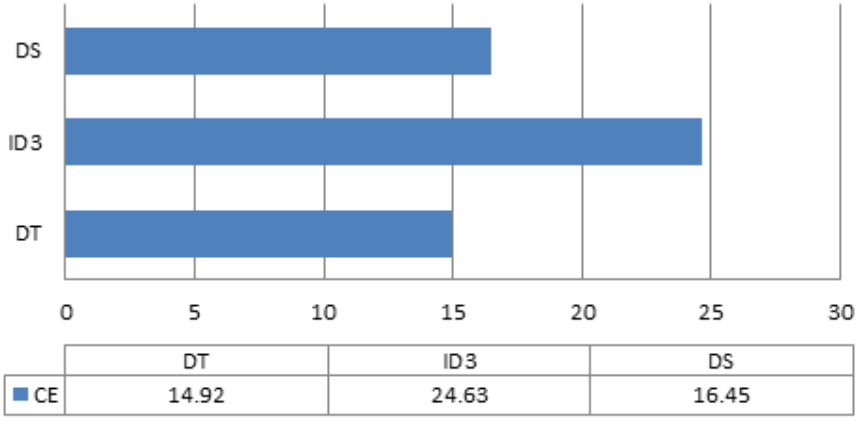

Figure 10. Mean Classification Error

In this paper table 1 provide the description of dataset used in this for training and testing the effectiveness of the proposed model. This table 1 it includes name of attributes or features type of feature, description of the attribute, table index or position of the attribute in the training and testing file. Table 2 shows the results of the experiments carried out using the decision table, ID3 and decision stumps. In the table 2 ACC stands for accuracy and CE stands for classification error. Figure 1 shows the posture a person who is Parkinson diseases affected. In Figure 2 it describe proposed prediction model for parkinsons diseases. In Figure 3 we showed pictorial representation of K fold cross validation. Figure 4, 5,6 describe the Tree generated through classification models decision tree ,ID3 and decision stumps respectively. Figure 7 and figure 8 describe the accuracy and classification error by 10 fold cross validation and the figure 9 and figure 10 shows the mean accuracy and classification error . Easily we can say using the graph, decision tree provide the best result.

5. Conclusion:

In this paper prediction of parkinsons disease paper we proposed a prediction model using data mining method, for predictions we used decision tree , ID3, and decision stumps classification algorithms. Dataset of parkinsons disease is used in this paper is taken from UCI repository. The dataset of Parkinson's disease iscomposed of biomedical voice measurements of 31 people, with 23 PD affected. Each column in thetable is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals. Parkinson's diseaseaffected data is labeled with P and healthy is with H. For experimentsK fold cross validation method isused with different classifiers as well as classification accuracy and error. The meanresults show in terms of classification accuracyerror, Decision tree has here performed very well i.e., accuracy 85.08 and classification error 14.92 and model ID3 performed worst, it gave accuracy 75.33 and classification error 24.63. There are different symptoms that lead to the Parkinson's disease are age and environmental factor, trembling in the legs, arms, hands, impaired speecharticulation and production difficulties. In this research paper speech articulation of Parkinson's disease affected people is considered for model formation and analyzes the modelbased on the symptom of disease.

References:

[1] Alexis Elbaz,James H. Bower, Brett J. Peterson, Demetrius M. Maraganore, Shannon K. McDonnell, J. Eric Ahlskog, Daniel J. Schaid, Walter A. Rocca, "Survival Study of Parkinson Disease in Olmsted County, Minnesota ", Arch Neurol. Vol. 60:91-96, 2003

[2] Parkinson, James (1817)," An essay on the shaking palsy"

[3] Newman EJ, Grosset KA, Grosset DG. Geographical difference in Parkinson's disease prevalence within West Scotland. Mov Disord 2009;24(3):401-6

[4] http://www.medicinenet.com/parkinsons_disease/article.htm

[5] http://www.medicalnewstoday.com/info/parkinsons-disease/

[6] http://www.umm.edu/patiented/articles/who_gets_parkinsons_disease_000051_3.htm

[7] Rajesh Pahwa, Kelly E. Lyons, William C. Roller,"Handbook of Parkinson's Disease" ,Third Edition

[8] William Dauer,Serge Przedborski,"Parkinson's Disease: Mechanisms and Models ",Neuron, Vol. 39, 889–909, September 11, 2003, Copyright 2003 by Cell Press

[9] Sanjay Pandey,"Parkinson`s Disease : Recent Advances",JAPI • june 2012 • VOL. 60

[10] James Parkinson,"An Essay on the Shaking Palsy",J Neuropsychiatry Clin Neurosci 14:2, Spring 2002.

[11] Stanley Fahn and the Parkinson Study Group,"Does levodopa slow or hasten the rate of progression of Parkinson's disease?",Neurol (2005) 252 [Suppl 4]: IV/37–IV/42

[12] Oliver Riedel et al., "Cognitive impairment in 873 patients with idiopathic Parkinson's disease Results from the German Study on Epidemiology of Parkinson's Disease with Dementia (GEPAD)",J Neurol (2008) 255:255–264

[13] Nathan Pankratz et al.,"Genomewide association study for susceptibility genes contributing to familial Parkinson disease ",Hum Genet (2009) 124:593–605,Springer-Verlag 2008.

[14] Keiko Tanaka et al. , " Occupational risk factors for Parkinson's disease:a case-control study in Japan",BMC Neurology 2011

[15] Calvin Yu-Chian Chen, "Mechanism of BAG1 repair on Parkinson's diseaselinked DJ1 mutation ",Journal of Biomolecular Structure and Dynamics , Taylor & fransis,Vol. 30, No. 1, 2012, 1–12

[16] Antonio Del Sol Mesa , "151 Network inference and analysis of Parkinson's disease",Journal of Biomolecular Structure and Dynamics Vol. 31, Supplement, 2013

[17] Audrey McKinlay,Randolph C. Grace "Characteristic of Cognitive Decline in Parkinson's Disease: A 1-Year Follow-Up",APPLIED NEUROPSYCHOLOGY, 18: 269–277, 2011

[18] C. W. Olanow and W. G. Tatton,"Etiology and pathogenesis of Parkinson's disease",Annu. Rev. Neurosci. 1999. 22:123–44

[19] Marco Aurélio M. Freire1, and José Ronaldo Santos,"Parkinson's disease: general features, effects of, levodopa treatment and future directions",Frontiers in Neuroanatomy,November 2010,Volume 4.

[20] Yadav G, Kumar Y, Sahoo G. "Predicationof Parkinson's disease using data mining methods: A comparative analysis of tree, statistical, and support vector machine classifiers."Indian J Med Sci 2011;65:231-42

[21] Chrish zarow etal, "Neuronal Loss Is Greater in the Locus Coeruleus Than Nucleus Basalis and Substantia nigra in Alzheimer and Parkinsons Diseases.", American medical association, 2003

[22] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering.