

THYROID DISEASE PREDICTION USING FEATURE SELECTION AND MACHINE LEARNING CLASSIFIERS

Dr. Dayanand Jamkhandikar ¹, Neethi Priya²

¹ Professor, Computer Science and Engineering Department, GNDEC Bidar, Karnataka, India ,

² Student, Computer Science and Engineering Department, GNDEC Bidar, Karnataka, India

ABSTRACT- *Hypothyroidism or hyperthyroidism is a major disease in India which arises due to malfunctioning of thyroid hormones. In the traditional way diagnosis includes clinical examination and the many blood tests. Diagnosis of Thyroid Disease is very tedious and difficult tasks at early stages with high accuracy. Medical industry has enormous quantity of data, but the bulk of this data is not processed. For proper diagnosis data must be processed accurately. For accurate processing intelligent Machine learning techniques can be used. Machine learning algorithms have been employed to model the prediction and diagnosis of thyroid patients. In this paper an attempt is made to analyze naïve bayes, k-nearest neighbour and Support Vector Machine (SVM) for multiclass classification of thyroid dataset. With comparative study, different ML techniques will able to achieve better accuracy in disease prediction.*

Keywords: hyperthyroidism, hypothyroidism, machine learning, feature selection, SVM, KNN and naïve bayes

1. INTRODUCTION

The thyroid is a little gland in the neck that produces thyroid hormones. It may produce too much or too small of these hormones. Hypothyroidism is a situation in which thyroid gland is not able to produce sufficient thyroid hormones. These hormones regulate metabolism of the body and further affects how the body uses energy. Lacking the accurate amount of thyroid hormones, body's normal functions start to slow down and body faces changes each day. In human services then clinical science, the applications based concerning Data dig are extremely gainful and significant. Analysis on Thyroid Disease is fairly dark or difficult errands. In medical field, Data mining assumes a quintessential assignment for finding concerning ailment. Information Mining gives several arrangement strategies in conformity with the forecast of ailment precision. The inert perception records gathered from plenty medicinal services association is treasured for the venture factors trial because some infections.

For diagnosis entire medical history and physical tests are used. As these tests produces large amount of data and ML can be used for finding important features from large amount of data. Due to this specialty of ML can be used in combination with medical science for the accurate diagnosis of hypo thyroidism disease. A number of ML techniques have been evolved and in order to achieve best accuracy of a model ensembles are widely used. In this paper we are utilizing 3 calculations according to anticipate thyroid illness at starting period by Utilizing a number highlights ultimately we are waiting for the exactness about the result or looking at it.

1.1 Overview concerning thyroid

The thyroid is an endocrine organ as shown in below figure 1. The capability concerning thyroid member is in accordance with relinquish of thyroid hormones. It compasses in imitation of each single vile part thru the circulatory provision then control digestion or improvement. The extensive elements concerning thyroid part incorporate breath, blood flow, gut developments, temperature control, muscle working, assimilation and working concerning cerebrum. Any colorings action in the thyroid part might also affect the ordinary physiological work ethnic body .The thyroid hormone influences the development and improvement relying on the excuse on immunity .At the point now the creation over thyroid hormone is much less or is recognized so hypo-thyroidism .At the point so the introduction concerning thyroid hormone is high below it type on thyroid contamination is known as much hyper-thyroidism.

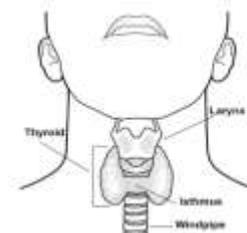


Fig. 1 thyroid organ

1.2 Thyroid and its fitness impacts

Thyroid difficulty are near basic endocrine infection, over the round the world. In an Indian study 42 million folks are experiencing these maladies. The Thyroid ailments are unique in relation in conformity with others as a long way as much their relative deceivability, utterance ease, scientific remedy mode availability .The incorrect advent over thyroid hormone influences wellbeing conditions.

1.2.1 Hyperthyroidism

Increment creation in the thyroid hormones causes hyperthyroidism Graves.' disorder is some of the immune system difficulty up to expectation causes hyperthyroidism .The warning signs are glacial pores and skin , increment affectability to heat ,diminishing regarding cloud ,weight reduction, raise pulse, hypertension, overabundance perspiring, neck extension, anxiety, menstrual intervals abbreviate, visit gut traits then fingers trembling

1.2.2 Hypothyroidism

Reduction introduction among the thyroid hormones motives Hypothyroidism. The medical term hypo implies inadequate or less. The Symptoms incorporates corpulence, low pulse, and increase between tranquil affectability, neck expanding, dead skin, arms deadness, hair issue, violent menstrual intervals and stomach associated issues. What's more, it Symptoms may decompile upstairs period postulate not rewarded.

Thyroid hormones: The thyroid body produces are tri-iodothyronine (T3) or L-thyroxine (T4).The thyroid hormones manage distinct metabolic exercises, for example, age regarding warmth, the utilization over sugars, protein and fats. The pituitary organ controls advent about tri-iodothyronine and L-thyroxine hormones. The Thyrotropin-Stimulating Hormone from pituitary part is discharged now thyroid hormone is required then circles through the habit rule according to enter at thyroid organ. TSH at so much point animates the thyroid organs because the advent concerning T4 then T3 hormones. The advent concerning thyroid hormone is confined by way of the input arrangement on pituitary organ. The TSH creation is less then T3, T4 are greater of the dissemination and TSH advent is more when T3, T4 are less.

2. LITERATURE SURVEY

The author in [1] examined and then recommends the object over rarely any statistics dig tactics for method regarding thyroid sickness. Malady determination assumes a indispensable assignment and it is vital because of somebody top clinician. Thyroid sickness is one sizeable contamination

and augur is the particularly troublesome assignment. Irina Ioniñă yet Liviuloniñă" [2] and [3] suggested that the float discipline alludes after thyroid illness order into twins on the almost general thyroid dysfunctions (hyperthyroidism yet hypothyroidism) among the populace. The creators examined yet seemed at four characterization models [4] [5] Naive Bayes, Decision Tree, Multilayer Perception and Radial Basis Function Network. Ali keles et al. [6] proposed an expert system for predicting of thyroid that is known as Expert System for Thyroid Disease Diagnosis(ESTDD).This expert system diagnose thyroid diseases through neuron fuzzy rules with 95.33% of accuracy. S. B. Patel [7] worked to predict the diagnosis of heart disease patients using classification mining techniques. Three classification function techniques in data mining are compared for predicting heart disease with reduced number of attributes.

3. PROBLEM STATEMENT

3.1 Existing System:

Diagnosis of Thyroid Disease is very tedious and difficult tasks. The diagnosis thyroid disease in the traditional way includes clinical examination and the many blood tests. But then the main task is to diagnosis the disease at early stages with high accurate percentage. In medical field, Data mining plays a crucial role for diagnosis of disease. Data Mining provides many classification techniques for the prediction of disease accuracy. The need of patient data collected from much health care organization is useful for the risk factors analysis for many diseases. The clinical decisions are usually based on the doctor's intuition. Therefore this may lead to disastrous consequences. Due to this there are many errors in the clinical decisions and it results in excessive medical costs.

3.2 Limitations of present fabric

- There is no action because health records yet ML techniques between existing explorations.
- No trial of the previous information.

3.3 Proposed System

In healthcare services data mining technique is mainly used for making decision, disease diagnosing and giving better treatment to the patients at comparatively low cost. Classification of thyroid disease plays is an important task in the prediction of disease. Dimensionality reduction may be done as a future work so that number of blood test the thyroid will be reduced and also time required diagnosing disease. The thyroid Dataset is taken from UCI data repository site. The Database consists of thyroid patient records. The Patients record is having different attributes described in the data set

description and different data mining techniques are applied to get the predication of thyroid disease. Data mining Algorithms such as KNN, Naïve bayes and Support vector machine are considered for the study.

3.4 Advantages of proposed system:

- We can predict the results using best classifier.
- Dynamic nature in prediction.
- We can predict on our own by collecting the readings from clinical test.

4. SYSTEM ARCHITECTURE

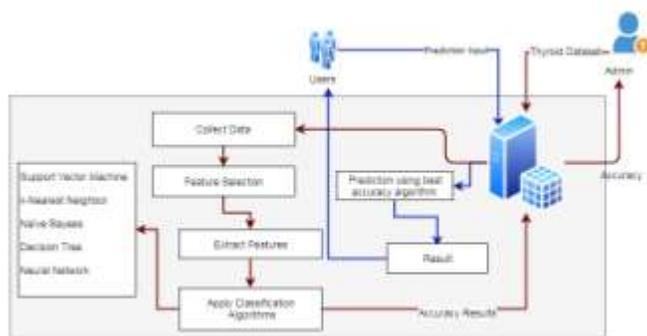


Figure 2: system architecture

The above figure 2 shows the system architecture as the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. First of all we collect the different patient's data. Out of many attributes we only select 15 attributes through feature selection. After extracting the features we apply the classification algorithms SVM, KNN and naïve bayes. Admin will train the data and can predict the best classification algorithm based on the accuracy of the result.

5. IMPLEMENTATION

Machine learning employs three main different algorithms those are explained below:

5.1 Support Vector Machine (SVM)

One of the kind of lesson fabric tab is Support Vector Machine[8], which is utilized in accordance with function characterization among an excellent exactly and utilizations 2 class classifier, alluded as much constrained airplane as much "choice power then choice surface". The atypical plane isolates

fine preparing take a look at along the bad making ready facts check within an arrangement. The points over pastime comprises a simple expand, utilized because of graph rearrangement, quadratic enchantment issue execute remain defined.

5.2 k-Nearest near (KNN)

KNN [9] is some about the best regular AI calculations utilized because order, It companies an information point structured of whether its neighbors are characterized. KNN stores every alone reachable litigation and organizations a density measures. The k-Nearest Neighbors tab (or KNN because short) is an easy tab in accordance with comprehend yet in accordance with execute. The usage choice keep explicit because of association troubles yet wish stand shown utilizing the Iris blossoms characterization issue. The model because of KNN is the whole preparing dataset. At the point then a port end is required because of a concealed facts example, the KNN calculation choice seem via the coaching dataset because of the k-most comparative cases. The augur faith over the almost comparable occasions is summed above then lower back so the hope because of the inconspicuous occurrence. The resemblance measure is reliant on the type about information. For authentic esteemed information, the Euclidean severance be able be utilized. Different kinds about information, because example, complete then doubled information, hamming split may stay utilized. On account over relapse issues, the regular of the predicted tension may stay returned.

5.3 Naive Bayes (NB)

In AI naïve bayes [10] classifiers are a crew regarding straightforward probabilistic classifiers dependent about applying bayes hypothesis with sure autonomy presumptions of the highlights. They are among the least complicated Bayesian provision models. it shares a usual government as the proximity about a precise thing in category lamely after the proximity of partial mean element.

5.4 The dataset description:

Dataset is committed from UCI AI storehouse [11]. Database contains concerning sufferers thyroid records. Every thyroid patient's document is comprises of 15 characteristics files beneath. Characteristic execute remain Boolean (genuine/bogus) then steady esteemed are addicted beneath. Below figure 3 shows the data set Hypothyroid.csv. Figure 4 shows the dialog flow and figure 5 shows the user registration.

	Age	Sex	Onset	Weight	Height	HeartRate	BloodPressure	Cholesterol	Glucose	Thyroid	result
1	24	Male	10	70	170	70	120	150	100	0	0
2	25	Female	12	60	160	65	110	140	90	1	1
3	26	Male	15	80	180	80	130	160	110	0	0
4	27	Female	18	75	175	75	125	155	105	1	1
5	28	Male	20	90	190	90	140	170	120	0	0
6	29	Female	22	85	185	85	135	165	115	1	1
7	30	Male	25	100	200	100	150	180	130	0	0
8	31	Female	28	95	195	95	145	175	125	1	1
9	32	Male	30	110	210	110	160	190	140	0	0
10	33	Female	32	105	205	105	155	185	135	1	1
11	34	Male	35	120	220	120	170	200	150	0	0
12	35	Female	38	115	215	115	165	195	145	1	1
13	36	Male	40	130	230	130	180	210	160	0	0
14	37	Female	42	125	225	125	175	205	155	1	1
15	38	Male	45	140	240	140	190	220	170	0	0
16	39	Female	48	135	235	135	185	215	165	1	1
17	40	Male	50	150	250	150	200	230	180	0	0
18	41	Female	52	145	245	145	195	225	175	1	1
19	42	Male	55	160	260	160	210	240	190	0	0
20	43	Female	58	155	255	155	205	235	185	1	1
21	44	Male	60	170	270	170	220	250	200	0	0
22	45	Female	62	165	265	165	215	245	195	1	1
23	46	Male	65	180	280	180	230	260	210	0	0
24	47	Female	68	175	275	175	225	255	205	1	1
25	48	Male	70	190	290	190	240	270	220	0	0
26	49	Female	72	185	285	185	235	265	215	1	1
27	50	Male	75	200	300	200	250	280	230	0	0
28	51	Female	78	195	295	195	245	275	225	1	1
29	52	Male	80	210	310	210	260	290	240	0	0
30	53	Female	82	205	305	205	255	285	235	1	1
31	54	Male	85	220	320	220	270	300	250	0	0
32	55	Female	88	215	315	215	265	295	245	1	1
33	56	Male	90	230	330	230	280	310	260	0	0
34	57	Female	92	225	325	225	275	305	255	1	1
35	58	Male	95	240	340	240	290	320	270	0	0
36	59	Female	98	235	335	235	285	315	265	1	1
37	60	Male	100	250	350	250	300	330	280	0	0
38	61	Female	102	245	345	245	295	325	275	1	1
39	62	Male	105	260	360	260	310	340	290	0	0
40	63	Female	108	255	355	255	305	335	285	1	1
41	64	Male	110	270	370	270	320	350	300	0	0
42	65	Female	112	265	365	265	315	345	295	1	1
43	66	Male	115	280	380	280	330	360	310	0	0
44	67	Female	118	275	375	275	325	355	305	1	1
45	68	Male	120	290	390	290	340	370	320	0	0
46	69	Female	122	285	385	285	335	365	315	1	1
47	70	Male	125	300	400	300	350	380	330	0	0
48	71	Female	128	295	395	295	345	375	325	1	1
49	72	Male	130	310	410	310	360	390	340	0	0
50	73	Female	132	305	405	305	355	385	335	1	1
51	74	Male	135	320	420	320	370	400	350	0	0
52	75	Female	138	315	415	315	365	395	345	1	1
53	76	Male	140	330	430	330	380	410	360	0	0
54	77	Female	142	325	425	325	375	405	355	1	1
55	78	Male	145	340	440	340	390	420	370	0	0
56	79	Female	148	335	435	335	385	415	365	1	1
57	80	Male	150	350	450	350	400	430	380	0	0
58	81	Female	152	345	445	345	395	425	375	1	1
59	82	Male	155	360	460	360	410	440	390	0	0
60	83	Female	158	355	455	355	405	435	385	1	1
61	84	Male	160	370	470	370	420	450	400	0	0
62	85	Female	162	365	465	365	415	445	395	1	1
63	86	Male	165	380	480	380	430	460	410	0	0
64	87	Female	168	375	475	375	425	455	405	1	1
65	88	Male	170	390	490	390	440	470	420	0	0
66	89	Female	172	385	485	385	435	465	415	1	1
67	90	Male	175	400	500	400	450	480	430	0	0
68	91	Female	178	395	495	395	445	475	425	1	1
69	92	Male	180	410	510	410	460	490	440	0	0
70	93	Female	182	405	505	405	455	485	435	1	1
71	94	Male	185	420	520	420	470	500	450	0	0
72	95	Female	188	415	515	415	465	495	445	1	1
73	96	Male	190	430	530	430	480	510	460	0	0
74	97	Female	192	425	525	425	475	505	455	1	1
75	98	Male	195	440	540	440	490	520	470	0	0
76	99	Female	198	435	535	435	485	515	465	1	1
77	100	Male	200	450	550	450	500	530	480	0	0
78	101	Female	202	445	545	445	495	525	475	1	1
79	102	Male	205	460	560	460	510	540	490	0	0
80	103	Female	208	455	555	455	505	535	485	1	1
81	104	Male	210	470	570	470	520	550	500	0	0
82	105	Female	212	465	565	465	515	545	495	1	1
83	106	Male	215	480	580	480	530	560	510	0	0
84	107	Female	218	475	575	475	525	555	505	1	1
85	108	Male	220	490	590	490	540	570	520	0	0
86	109	Female	222	485	585	485	535	565	515	1	1
87	110	Male	225	500	600	500	550	580	530	0	0
88	111	Female	228	495	595	495	545	575	525	1	1
89	112	Male	230	510	610	510	560	590	540	0	0
90	113	Female	232	505	605	505	555	585	535	1	1
91	114	Male	235	520	620	520	570	600	550	0	0
92	115	Female	238	515	615	515	565	595	545	1	1
93	116	Male	240	530	630	530	580	610	560	0	0
94	117	Female	242	525	625	525	575	605	555	1	1
95	118	Male	245	540	640	540	590	620	570	0	0
96	119	Female	248	535	635	535	585	615	565	1	1
97	120	Male	250	550	650	550	600	630	580	0	0
98	121	Female	252	545	645	545	595	625	575	1	1
99	122	Male	255	560	660	560	610	640	590	0	0
100	123	Female	258	555	655	555	605	635	585	1	1

Figure 3: data set

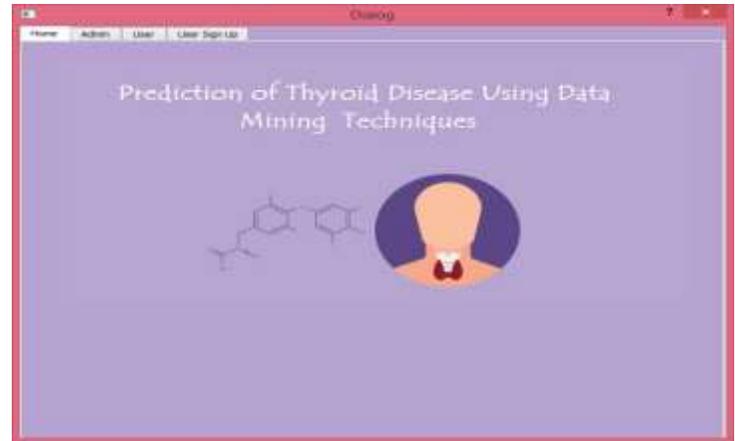


Figure 4: dialog flow

5.5 MODULES

ADMIN: Admin is the one who decides which classification algorithm has to be given to the user. Admin after login; train the dataset with accordant iii calculations. Support Vector Machine (SVM), k-Nearest Neighbor (KNN) and naive Bayes (NB). The classification can be done by the admin. The following functions are performed by the admin:

- After the training admin will test the accuracy by splitting 30 percentages of data from the training file.
- Then admin will find the best classification algorithm, called Naïve bayes.
- Admin also can see the graph of the accuracy of the three algorithms and feature selection score graph.

USER: User is an end consumer of the application; our application intention help in imitation of the consumer by means of hope thyroid contamination by instruct the past patient's dataset together with classification algorithms. User may login with his medical details to known whether he is suffering from thyroid disease or not. The functions of the user are given below:

- User can register with own details and after login user upload single patient record in the csv file.
- User can see the result with through the forecast of calculation.

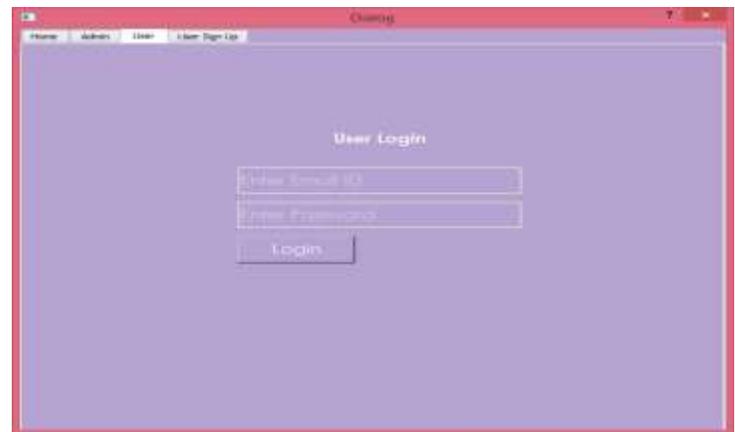


Figure 5:user registration

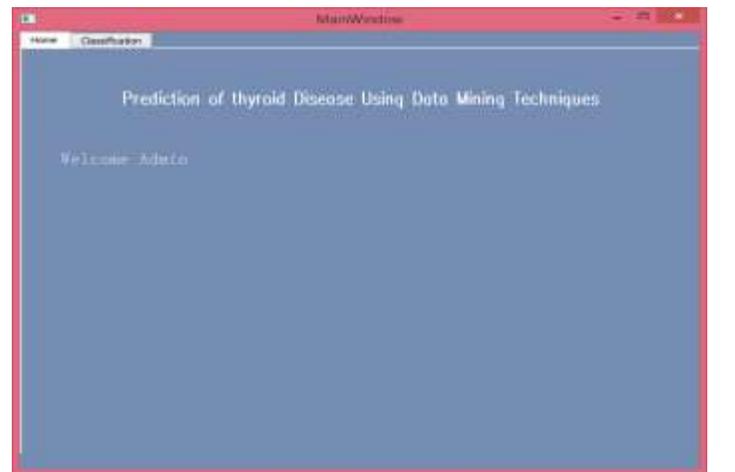


Figure 6: admin window

6. RESULT ANALYSIS

Below figure 7 shows the classification window and figure 8 shows the graph of feature selection. The accuracy prediction for various machine learning approaches is shown in figure 9.

The graph of accuracy analysis using SVM, Naïve Bayes algorithm and KNN is shown in figure 10.

Figure 11 identifies whether a person is having thyroid or not by displaying positive if the person is suffering from thyroid or else negative.



Figure 7 Classification window

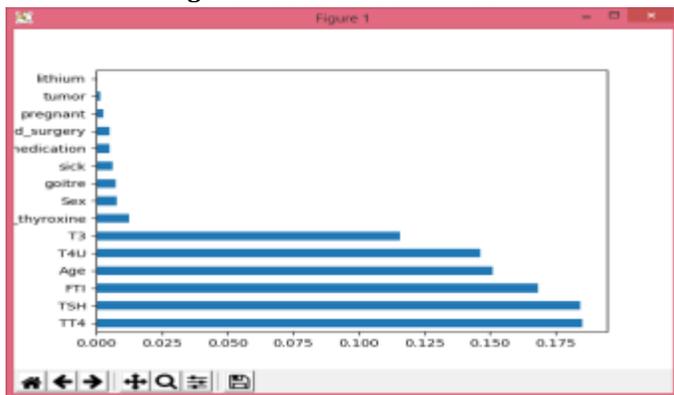


Figure 8 Graph of feature selection

Sample:

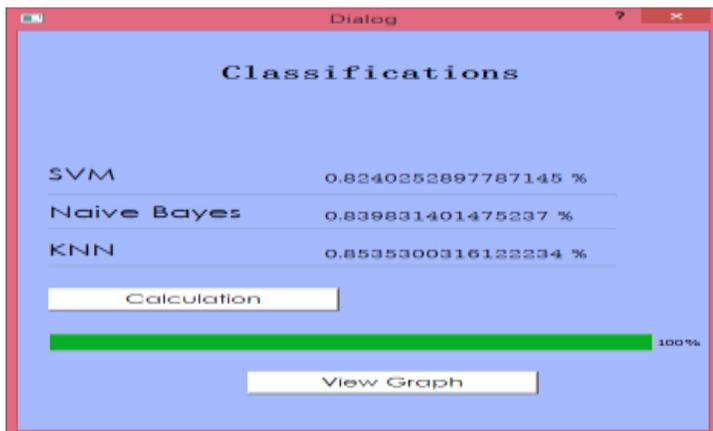


Figure 9: accuracy prediction

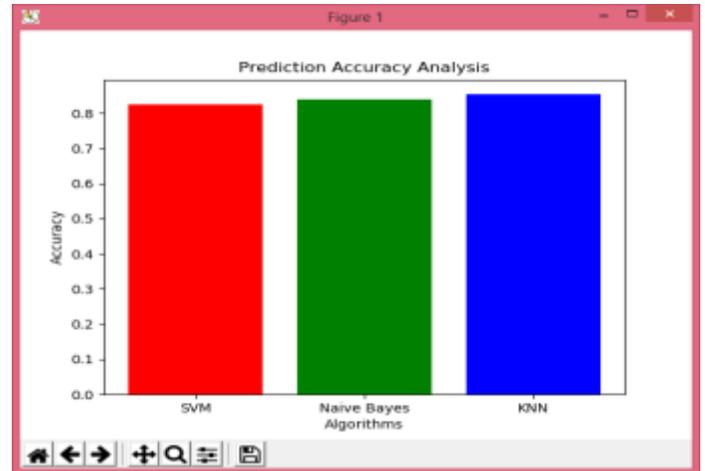


Figure 10: graph of accuracy analysis

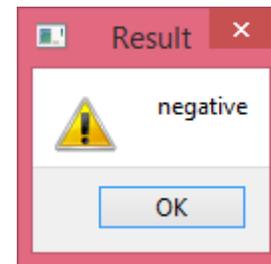


Figure 11: showing end results (Positive or negative)

CONCLUSION

In this paper we have proposed method to predict the thyroid disorder at earlier stage using data mining techniques. Data mining classification algorithms are used to diagnose the thyroid problems. Proposed technique helps to minimize the noisy data of a patient. Data mining Algorithms such as KNN, Naïve bayes, Support vector machine are considered for the study. The results of these classification methods are based on accuracy and performance of the model. The resulting classification of effective data helps to find the treatment to the thyroid patients with better cost and facilitate the management. For the given data set the accuracy using SVM is 0.82, Naïve Bayes is 0.83 and KNN is 0.85.

REFERENCES

- [1] Roshan Banu D, then K.C.Sharmili "A Study of Data Mining Techniques after Detect Thyroid Disease" International Journal concerning Innovative Research into Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2017
- [2] IrinaloniÑă or LiviuloniÑă" Prediction on Thyroid Disease Using Data Mining Techniques" The Classification Technique because of Talent Management the use of SVM, (ICCEET), 978-1-4673-0210-4/12, pp. 959- 963, 2017
- [3] Khushboo Taneja, Parveen Sehgal, Prerana "Predictive Data Mining because of Diagnosis over Thyroid Disease the use of Neural Network" International Journal regarding Research of Management, Science & Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016
- [4] Hanung Adi Nugroho, Noor Akhmad Setiawan, Md. DendiMaysanjaya," A Comparison regarding Classification Methods about Diagnosis of Thyroid Diseases" IEEE International Seminar of Intelligent Technology and Its Applications, 2017
- [5] K. Rajam and R. Jemina Priyadarsini "A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques" ,IJCSMC, Vol. 5, Issue. 5, May 2016, pg.354– 358.
- [6] Ali keles et al. "Expert System for Thyroid Disease Diagnosis(ESTDD)". This expert system diagnose thyroid diseases through neuron fuzzy rules with 95.33% of accuracy.
- [7] S. B. Patel "Review of machine learning techniques "also worked to predict the diagnosis of heart disease patients using classification mining techniques. Three classification function techniques in data mining are compared for predicting heart disease with reduced number of attributes.
- [8] Umadevi S , Dr .Jeen Marseline K.S,"Applying Classification Algorithms to Predict Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 10, September 2017
- [9] H.S., S.K., J.H.J., and A.M."A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning"
- [10] Lewis, D. "Naive Bayes at forty" the independence assumption in information retrieval. It describes about the working of naïve bayes algorithm.
- [11] UCI Machine learning repository (patient's data) (online). Available: <http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/hypothyroid>