CHAPTER -3
CHAPTER -3

THYROID DATASETS

3.1 INTRODUCTION

This chapter briefly reviews the structure of the thyroid, functions of the thyroid, various thyroid function tests, thyroid disease classification and thyroid data sets from UCI Machine Learning Repository (UCITD).

TD is a study of Endochronology, which at most seems to be an element in Science of Medicine. Thyroid is one of the most common diseases that is frequently misunderstood and misdiagnosed. Based on the thyroid conditions the diseases which are obtained by the thyroid gland are classified as 7 Categories i.e., Hyperthyroid, Hypothyroid, Binding protein, Replacement therapy, Anti-thyroid therapy, Non thyroid illness (general fever) & miscellaneous features (Aral et al., 2007). There are 28 pristine attributes which are considered for tracing out the specimens of TDA, but considering this geographical world the thyroid gland effect will be different all around.

Computer professionals are able to provide ES to diagnose different kinds of diseases with high accuracy. These systems are based on AI which helps the physician or consultant to minimize the cost, time and become expert in effective diagnoses. Medical diagnose is implemented for diseases related to cancer, diabetes, heart diseases, lungs, liver and so on. Recently an android mobile is released into the market for testing, which consists of 121 medical tests based on the symptoms which it identifies through persons mass, skin, saliva and so on. Some examples for android apps developed are disease dictionary, drug dictionary, feed baby pro, pregnancy+, muscle triquer pointer.

3.2 KNOWLEDGE BASE

KB is constructed using 190 patient’s data of southern states in INDIA with 257 rules and 390 patients (MORE Indian thyroid datasets of IETD, 2012) data consisting 549 rules obtained from the Intelligent System Laboratory of K.N.Toosi University of Technology (UCI Datasets, 1997) from Imam Khomeini hospital.
Thyroid disease records are supplied by the Garavan Institute and J. Ross Quinlan *(Thyroid Disease Records, 1987)*, New South Wales Institute, Sydney, Australia, 2012. The information collected through inputs is stored as a knowledge base that serves as repository for quick processing and future retrieval.

**General Description of Thyroid Disease Databases and Related Files**

This directory contains 6 databases, corresponding test set, and corresponding documentation. They were left at the University of California at Irvine by Ross Quinlan during his visit in 1987 for the 1987 Machine Learning Workshop.

The documentation files (with file extension "names") are formatted to be read by Quinlan's C4 decision tree program. Though briefer than the other documentation files found in this database repository, they should suffice to describe the database, specifically:

1. Source
2. Number and names of attributes (including class names)
3. Types of values that each attribute takes

**In general, these databases are quite similar and can be characterized somewhat as follows:**

1. Many attributes (29 or so, mostly the same set over all the databases)
2. Mostly numeric or Boolean valued attributes
3. Thyroid disease domains (records provided by the Garavan Institute of Sydney, Australia)
4. Several missing attribute values (signified by "?")
5. Small number of classes (under 10, changes with each database)
6. 2800 instances in each data set
7. 972 instances in each test set (It seems that the test sets' instances are disjoint with respect to the corresponding data sets, but this has not been verified)
**Fig 3.1:** Thyroid Dataset from IETD ML.

**Fig 3.2:** Thyroid Dataset from UCI Database.
**UCI THYROID DATASETS:**

![UCI Database from Machine Learning Repository](image1)

**Fig 3.3: UCI Database from Machine Learning Repository**

**DONATED INDIAN IETD DATASETS:**

![India ETD Database](image2)

**Thyroid Disease Data Set (Indian IETD Dataset)**

![Download Data Folder Data Set Description](image3)

**Data Set Information:**

- # From
  - # Documentation: as given by V Prasad
  - # 6 databases from the VCH, Andhra Pradesh, India
  - # Approximately the following for each database:
    - **”3” classes, 215 instances, 5 attributes**
    - **”2” classes, 510 instances, 5 attributes**
    - No missing values
  - # A Thyroid database suited for training ANNs
  - **”3” classes, 772 training instances, 703 testing instances**
  - Includes cost data

**Source:**

V Prasad, Research Scholar, Dept of CSE, GITAM Institute of Technology

**Relevant Papers:**


**Papers That Cite This Data Set:**

Fig 3.4: IETD data submitted to Machine Learning Repositories

Symptoms
S1=Recognition of Thyroxin    S2=Query on thyroxin    S3=Anti thyroid Medication
S4=Sick /Weakness/Drowsiness   S5=Pregnancy Problems  S6=Thyroid Surgery
S7=I131 Treatment             S8=Hypothyroid Query  S9=Hyperthyroid Query
S10= Lithium                  S11=Goitre          S12=Tumor

Indication in Data Sets      1- Yes          0-No

3.2.1 IF THEN ELSE RULES

Rule 1
If Symptoms
S1=0, S2= 1, S3=1, S4= 1, S5=0, S6= 1, S7= 0, S8= 0, S9= 0, S10= 0, S11= 1, S12= 0
Diagnose: Thyroid is negative
Prevention or curing method: Feel free, but regular checkups on thyroxin is necessary.

Rule 2
If Symptoms
S1=1, S2= 0, S3=0, S4= 0, S5=0, S6= 1, S7= 0, S8= 1, S9= 0, S10= 0, S11= 0, S12= 0
Diagnose: Hyperthyroid
Prevention or curing method: Referral source is SVI (or) consult a near by doctor ,
Since thyroxin levels are more.

Rule 3
If Symptoms
S1=0, S2= 1, S3=1, S4= 0, S5=0, S6= 0, S7= 0, S8= 0, S9= 1, S10= 1, S11= 0, S12= 0
Diagnose: Thyroid primary, antithyroid therapy
Prevention or curing method: Antithyroid medication is required and suggested to use antithyroid drugs.
Rule 4

If Symptoms
S1=0, S2= 0, S3=0, S4= 0, S5=1, S6= 1, S7= 0, S8= 1, S9= 0, S10= 0, S11= 1, S12= 0

Diagnose: Missing regular menstrual cycles, drowsiness and swellings

Prevention or curing method: Eat food that includes proteins in it to reduce swellings. Drink less water and try to have water which contains more minerals in it.

Rule 5

If Symptoms
S1=0, S2= 1, S3=0, S4= 0, S5=0, S6= 0, S7= 1, S8= 1, S9= 1, S10= 1, S11= 1, S12= 0

Diagnose: Goitre

Prevention or curing method: Swellings to be reduced and antic radioactive ionized treatment suggested.

Rule 6

If Symptoms
S1=1, S2= 1, S3=1, S4= 1, S5=1, S6= 1, S7= 1, S8= 0, S9= 0, S10= 1, S11= 0, S12= 1

Diagnose: T3 Toxic

Prevention or curing method: Thyroxin levels are more and requested to compel down by removing the goitre if observed and to reduce the thyroxin levels by daily checkups, radio iodine treatment.

3.3 SIMPLIFIED SAMPLE DATASETS

The datasets mainly consists of 12 symptoms with which we can predict the existence of disease obtained through thyroid gland.

Table 3.1: Simplified sample dataset.

<table>
<thead>
<tr>
<th>SYMPTOMS</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
<th>Diagnose</th>
<th>Prevention or Referral Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule Set-1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>Thyroid is Negative</td>
<td>SVHC (or) But regular checkups on thyroxin necessary.</td>
</tr>
</tbody>
</table>
Table 3.1 shows that after considering the entered symptoms, the disease can be predicted from this sample database which consists of more than 2096 tuples for identification of the TD.
3.4 THYROID INFORMATION SYSTEM

INPUTS:
- Thyroid disease
- Thyroid disease parameters
- Curing methodologies
- Preventions

OUTPUTS:
- Information disease
- Description about the disease
- Diagnosis

3.5 UCI THYROID DATASET

UCI Machine Learning Repository (Center for Machine Learning and Intelligent Systems) currently maintains 221 data sets in various areas life sciences, physical sciences, CS or Engineering, social sciences, business, game and others as a service to the machine learning community. All data sets can be viewed through the searchable interface. These datasets can be used for tasks like classification, clustering, regression and causal discovery. It also accepts the truth datasets from the donors as per the donation policy.

ATTRIBUTES ORDER:

TSH TEST

A test for thyroid-stimulating hormone (TSH) is done to:
- Find out whether the thyroid gland is working properly.
  - An underactive thyroid gland (hypothyroidism) can cause symptoms such as weight gain, tiredness, dry skin, constipation, a feeling of being too cold, or frequent menstrual periods.
An overactive thyroid (hyperthyroidism) can cause symptoms such as weight loss, rapid heart rate, nervousness, diarrhea, a feeling of being too hot, or irregular menstrual periods.

**T4 TEST(S)**

T4 circulates in the blood in two forms:

1) **T4** bound to proteins that prevent the T4 from entering the various tissues that need thyroid hormone.

2) **Free T4**, which does enter the various target tissues to exert its effects. The free T4 fraction is the most important to determine how the thyroid is functioning, and tests to measure this are called the Free T4 (FT4) and the Free T4 Index (FT4I or FTI). Individuals who have hyperthyroidism will have an elevated FT4 or FTI, whereas patients with hypothyroidism will have a low level of FT4 or FTI.

Combining the TSH test with the FT4 or FTI accurately determines how the thyroid gland is functioning. The finding of an elevated TSH and low FT4 or FTI indicates primary hypothyroidism due to disease in the thyroid gland. A low TSH and low FT4 or FTI indicates hypothyroidism due to a problem involving the pituitary gland. A low TSH with an elevated FT4 or FTI is found in individuals who have hyperthyroidism.

**T3 TEST(S)**

T3 tests are often useful to diagnosis hyperthyroidism or to determine the severity of the hyperthyroidism. Patients who are hyperthyroid will have an elevated T3 level. In some individuals with a low TSH, only the T3 is elevated and the FT4 or FTI is normal. T3 testing rarely is helpful in the hypothyroid patient, since it is the last test to become abnormal. Patients can be severely hypothyroid with a high TSH and low FT4 or FTI, but have a normal T3. In some situations, such as during pregnancy or while taking birth control pills, high levels of total T4 and T3 can exist. This is because the estrogens increase the level of the binding proteins. In these situations, it is better to ask both for TSH and free T4 for thyroid evaluation.
THYROXINE BINDING GLOBULIN(TBG) TEST

It is used to identify the issues in thyroid. It is a protein bounded test suggested for checking the levels of T3 & T4 protein bounds.

Table 3.2: UCI thyroid dataset.

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE:</td>
<td>CONTINUOUS.</td>
</tr>
<tr>
<td>SEX:</td>
<td>M (0), F (1).</td>
</tr>
<tr>
<td>ON THYROXINE:</td>
<td>F, T.</td>
</tr>
<tr>
<td>QUERY ON THYROXINE:</td>
<td>F, T.</td>
</tr>
<tr>
<td>ANTITHYROID MEDICATION:</td>
<td>F, T.</td>
</tr>
<tr>
<td>SICK:</td>
<td>F, T.</td>
</tr>
<tr>
<td>PREGNANT:</td>
<td>F, T.</td>
</tr>
<tr>
<td>THYROID SURGERY:</td>
<td>F, T.</td>
</tr>
<tr>
<td>131 TREATMENT</td>
<td>F, T.</td>
</tr>
<tr>
<td>QUERY HYPOTHYROID:</td>
<td>F, T.</td>
</tr>
<tr>
<td>QUERY HYPERHYPOHYROID:</td>
<td>F, T.</td>
</tr>
<tr>
<td>LITHIUM:</td>
<td>F, T.</td>
</tr>
<tr>
<td>GOITRE:</td>
<td>F, T.</td>
</tr>
<tr>
<td>TUMOR:</td>
<td>F, T.</td>
</tr>
<tr>
<td>HYPOPITUITARY:</td>
<td>F, T.</td>
</tr>
<tr>
<td>PSYCH:</td>
<td>F, T.</td>
</tr>
<tr>
<td>TSH MEASURED:</td>
<td>F, T.</td>
</tr>
<tr>
<td>TSH:</td>
<td>CONTINUOUS.</td>
</tr>
<tr>
<td>T3 MEASURED:</td>
<td>F, T.</td>
</tr>
<tr>
<td>T3:</td>
<td>CONTINUOUS.</td>
</tr>
<tr>
<td>TT4 MEASURED:</td>
<td>F, T.</td>
</tr>
<tr>
<td>TT4:</td>
<td>CONTINUOUS.</td>
</tr>
<tr>
<td>T4U MEASURED:</td>
<td>F, T.</td>
</tr>
<tr>
<td>T4U:</td>
<td>CONTINUOUS.</td>
</tr>
<tr>
<td>FTI MEASURED:</td>
<td>F, T.</td>
</tr>
<tr>
<td>FTI:</td>
<td>CONTINUOUS.</td>
</tr>
<tr>
<td>TBG MEASURED:</td>
<td>F, T.</td>
</tr>
<tr>
<td>TBG:</td>
<td>CONTINUOUS</td>
</tr>
<tr>
<td>REFERRAL SOURCE:</td>
<td>WEST, STMW, SVHC, SVI, SVHD and OTHERS</td>
</tr>
</tbody>
</table>

Symptoms

S01-Age  S02-Sex  S03-On-thyroxin  S04-Query-on-thyroxin
S05-On Anti-thyroid-medication  S06-Sick  S07-Pregnant
S08-Thyroid-surgery  S09-I131-treatment  S10-Query-hypothyroid
S11-Query-hyperthyroid  S12- Lithium  S13-Goiter
S14-Tumor  S15-Hypo-pituitary  S16-Psych
S17-Tsh-measured  S18-Tsh  S19-T3-measured
S20-T3  S21-Tt4-measured  S22-Tt4
S23-T4u-measured  S24-T4u  S25-Fti-measured
S26-Fti  S27-Tbg-measured  S28-Tbg
S29-Referral Source  S30-Predicted Disease

Indication in Data Sets  1- Yes / T  0-No / F  Rule set-RS  Symptoms-S

**Table 3.3:** UCI thyroid dataset with 8 tuples out of 2096 tuples.

<table>
<thead>
<tr>
<th>S</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
<th>S13</th>
<th>S14</th>
<th>S15</th>
<th>S16</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS-1</td>
<td>41</td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-2</td>
<td>87</td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-3</td>
<td>44</td>
<td>M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-4</td>
<td>19</td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-5</td>
<td>25</td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-6</td>
<td>60</td>
<td>M</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-7</td>
<td>29</td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RS-8</td>
<td>53</td>
<td>M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

After considering the entered symptoms, the disease can be predicted from this sample database, like wise it consists of more than 2096 tuples for identification of TD as shown in Table 3.3.
3.5.1 IF THEN ELSE RULES

Rule 1: IF
THEN
Disease =Negative.|3733
Referral-Source is SVHC

Rule 2: IF
S1=87, S2=F, S3=F, S4=F, S5=F, S6=F, S7=F, S8=F, S9=F, S10=F, S11=F, S12=F, S13=F, S14=F, S15=F, S16=F, S17=T, S18=0.15, S19=T, S20=1.7, S21=T, S22=162, S23=T, S24=0.87, S25=T, S26=186, S27=F, S28=0
THEN
Disease =Hyperthyroid.|1873
Referral-Source is SVI

Rule 3: IF
THEN
Disease =Negative.|3345
Referral-Source is SVHD

Rule 4: IF
S1=19, S2=F, S3=F, S4=F, S5=F, S6=F, S7=T, S8=F, S9=F, S10=F, S11=F, S12=F, S13=F, S14=T, S15=F, S16=F, S17=T, S18=0.45, S19=T, S20=3.2, S21=T, S22=130, S23=T, S24=1.83, S25=T, S26=71, S27=F, S28=0
THEN
Disease =Goitre.|3523
Referral-Source is STMW

**Rule 5:** IF
S1=25, S2=F, S3=F, S4=F, S5=F, S6=F, S7=F, S8=F, S9=F, S10=F, S11=F, S12=F, S13=F, S14=F, S15=F, S16=F, S17=T, S18=1.6, S19=T, S20=5.4, S21=T, S22=152, S23=T, S24=1.5, S25=T, S26=102, S27=F, S28=0
THEN
Disease = Negative.|1183
Referral-Source is STMW

**Rule 6:** IF
S1=60, S2=M, S3=F, S4=F, S5=T, S6=F, S7=F, S8=F, S9=F, S10=F, S11=F, S12=F, S13=F, S14=F, S15=F, S16=F, S17=T, S18=0.2, S19=T, S20=4, S21=T, S22=68, S23=T, S24=1, S25=T, S26=67, S27=F, S28=0
THEN
Disease = T3 Toxic.|547
Referral-Source is others.

**Rule 7:** IF
THEN
Disease = GoiTre.|2469
Referral-Source is SVH

**Rule 8:** IF
S1=53, S2=M, S3=F, S4=F, S5=F, S6=F, S7=F, S8=F, S9=F, S10=F, S11=F, S12=F, S13=F, S14=F, S15=F, S16=F, S17=T, S18=1.4, S19=T, S20=1.9, S21=T, S22=104, S23=T, S24=0.93, S25=T, S26=112, S27=F, S28=0
THEN
3.6 CLASSIFICATION OF THYROID DATASETS

*Table 3.4: Classification table of thyroid with different classes and representation of elements.*

<table>
<thead>
<tr>
<th>S.NO</th>
<th>CLASS NUMBER</th>
<th>CLASS NAME</th>
<th>REPRESENTATIONAL ELEMENTS</th>
<th>DESIRED DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Class 1</td>
<td>Hyperthyroid</td>
<td>A</td>
<td>Hyperthyroid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td>T3 toxic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>Toxic goitre</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D</td>
<td>Secondary toxic</td>
</tr>
<tr>
<td>2</td>
<td>Class 2</td>
<td>Hypothyroid</td>
<td>E</td>
<td>Hypothyroid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td>Primary hypothyroid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>Compensated hypothyroid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H</td>
<td>Secondary hypothyroid</td>
</tr>
<tr>
<td>3</td>
<td>Class 3</td>
<td>Binding protein</td>
<td>I</td>
<td>Increased binding protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>J</td>
<td>Decreased binding protein</td>
</tr>
<tr>
<td>4</td>
<td>Class 4</td>
<td>General health</td>
<td>K</td>
<td>Concurrent non-thyroidal illness</td>
</tr>
<tr>
<td>5</td>
<td>Class 5</td>
<td>Replacement therapy</td>
<td>L</td>
<td>Consistent with replacement therapy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M</td>
<td>Under replaced</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>Over replaced</td>
</tr>
<tr>
<td>6</td>
<td>Class 6</td>
<td>Antithyroid treatment</td>
<td>O</td>
<td>Antithyroid drugs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>I131 treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>Surgery</td>
</tr>
<tr>
<td>7</td>
<td>Class 7</td>
<td>Miscellaneous</td>
<td>R</td>
<td>Discordant assay results</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>Elevated tbg</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>Elevated thyroid hormones</td>
</tr>
</tbody>
</table>

The thesis had found the solutions in both levels of clinical and classification settings for the Classes 1, 2 & 3 and remaining classes are leftover for future work.
3.7 SUMMARY

The common attributes of both IETD and UCI datasets are considered and carried out for the experimentation to evaluate the performance of the classification variable in Chapter 6. UCITD datasets are individually used in Chapter 4 & 5 for analyzing the performance and prediction capabilities of machine learning and evolutionary algorithms. However, the results with IETD dataset can be attributed to the limited number of samples compared to UCITD dataset, due to lack of relevant features. But the variance calculations and significance constants proved that, there is a variation in the significance to large extent within the group compared with between the groups of ANOVA. The datasets which are theoretically and considered for the proposed work are checked and verified by renowned doctors and thyroid specialists.