CST8390_012 - Assignment 2

Thyroid Disease Dataset Analysis Report by Performing Clustering (K-Means and Farthest First) and Outlier Detection (Local Outlier Factor and Isolation Forest)

Author of the overall report and Workload

Shu Han Han #041-060-762	Modeling and Evaluation of Outlier Detection
Wan-Hsuan Lee #041-060-761	Modeling and Evaluation of Clustering

Computer Programming, Algonquin College

June 16, 2023

Table of Contents

Introduction

Thyroid diseases are complex health conditions that affect the functioning of the thyroid gland, leading to physiological imbalances. Analyzing thyroid disease datasets using data mining techniques provides valuable insights into patterns, subgroups, and outliers, improved understanding, diagnosis, and treatment. This report follows the CRISP-DM methodology and focuses on a comprehensive analysis of the Thyroid Disease dataset, utilizing clustering and outlier detection models.

The analysis begins with clustering techniques to identify natural groupings or clusters within the dataset. Clustering algorithms like K-Means and Farthest First uncover distinct subpopulations of patients with similar attributes and thyroid profiles, aiding in understanding disease progression, and treatment responses. The results of the clustering analysis will be visually presented and interpreted to gain insights into the underlying structure of the dataset.

In addition to clustering, outlier detection methods are then employed to identify potential anomalies. Outliers, which significantly deviate from the majority, may indicate data quality issues, measurement errors, or rare thyroid disease cases. The Local Outlier Factor and Isolation Forest algorithms are used to detect and analyze outliers and understand their impact on the overall analysis and their potential implications for diagnosis or treatment.

The clustering results and identified outliers are thoroughly analyzed and interpreted in the context of thyroid diseases. Patterns and trends within the clusters are examined to understand the relationships between patient attributes and thyroid disorders. The insights gained have the potential to improve diagnostic accuracy, facilitate personalized treatment approaches, and enhance patient outcomes.

In the following sections, we will delve into the details of each step in the analysis, including data understanding, data preparation, modeling, evaluation, and discussion of results. By following a structured approach, we strive to uncover meaningful patterns, identify patient subgroups, and detect potential anomalies. This study aims to enhance our understanding of thyroid diseases and provide valuable insights for medical practitioners, researchers, and stakeholders in the field.

Business Understanding

1. Description

The dataset aims to address the problem of thyroid diagnosis. The objective is to develop a predictive model that can accurately classify patients into different thyroid conditions based on their clinical attributes and lab test results.

The dataset consists of a collection of patient records, where each record contains various attributes and corresponding thyroid condition labels. The attributes may include demographic information (such as age and sex), medical history (such as previous thyroid surgeries or treatments), symptoms, and results of thyroid-related tests (such as TSH, T3, and TT4 levels).

The dataset is likely generated from clinical settings or medical records, where patient information is collected during the diagnostic process. The data collection may involve multiple clinics or healthcare facilities to ensure a diverse representation of patients and conditions.

2. Determine Goals

The main goal of business understanding in the context of the thyroid disease dataset is to leverage the available data to gain insights that can improve diagnosis, treatment, and prevention of thyroid diseases. By analyzing the patterns and relationships within the dataset, the aim is to enhance disease management strategies, optimize treatment plans, identify risk factors, enable personalized medicine approaches, and allocate resources effectively. Ultimately, the goal is to utilize the knowledge extracted from the dataset to make informed decisions that lead to better patient outcomes and advancements in the field of thyroid disease management.

3. Produce Project Plan

By following CRISP-DM methodology, utilizing clustering and outlier detection models, and guidelines in the assignment file (CST8390 Assignment 2), we make a work breakdown list to ensure the collaboration of teamwork.

Data Understanding

1. Collect Initial Data

- ann-train.data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Age	Sex	On Thyroxine	Query On Th	On Antithyro	Sick	Pregnant	Thyroid Surg	I131 Treatm	Query Hypotl	Query Hypert	Lithium	Goitre	Tumor	Hypopituitar	Psych	TSH
2	0.73	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0.0006
3	0.24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00025
4	0.47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0019
5	0.64	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0009
6	0.23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00025
7	0.69	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00025
8	0.85	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00025
9	0.48	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00208
10	0.67	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0013
11	0.76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0001
12	0.62	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0.011
13	0.18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.0001
14	0.59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0008
15	0.49	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0006
16	0.53	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.0023
17	0.39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0001
18	0.39	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.0006
19	0.65	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.0016
20	0.64	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.032

2. Describe Data

- Data Description:
  - Instances: 3772
  - Attributes: 21(15 attributes are binary, 6 attributes are continuous) plus 1 class
  - Class:
    - Clustering: Thyroid Condition (Hyperthyroid, Hypothyroid, and Normal)
    - Outlier Detection: Outlier ("Yes": Hyperthyroid or hypothyroid; "No": Normal)

No	Attribute	Description
1	Age	The age of the patient.
2	Sex	The gender of the patient.
3	On Thyroxine	This indicates whether a patient is currently taking thyroxine medication, which is commonly prescribed for thyroid hormone replacement therapy.
4	Query On Thyroxine	This suggests seeking information specifically about patients who are currently on thyroxine medication.
5	On Antithyroid Medication	This attribute indicates whether a patient is currently taking medications that inhibit the production or release of thyroid hormones. Antithyroid medications are often prescribed for the treatment of hyperthyroidism.
6	Sick	This refers to the state of being unwell or experiencing illness. In the context of thyroid disease, it may be relevant

		to evaluate the impact of illness on thyroid function or treatment.
7	Pregnant	This indicates whether a patient is currently pregnant. Thyroid conditions can sometimes be influenced by pregnancy, and special considerations may be required for pregnant individuals with thyroid disease.
8	Thyroid Surgery	This denotes whether a patient has undergone surgery involving the thyroid gland. Thyroid surgery may be performed for various reasons, such as the removal of thyroid nodules, thyroid cancer, or to treat certain thyroid disorders.
9	I131 Treatment	I131 Treatment involves the use of radioactive iodine (I131) for therapeutic purposes. It is often employed to treat hyperthyroidism or thyroid cancer by destroying the overactive thyroid cells or thyroid cancer cells.
10	Query Hypothyroid	This suggests seeking information specifically related to hypothyroidism, a condition characterized by an underactive thyroid gland and insufficient production of thyroid hormones.
11	Query Hyperthyroid	This indicates a search or request for information specifically related to hyperthyroidism, a condition characterized by an overactive thyroid gland and excess production of thyroid hormones.
12	Lithium	Lithium is a medication primarily used to treat bipolar disorder. It can sometimes affect thyroid function and may be associated with the development of thyroid disorders.
13	Goitre	Goitre refers to the enlargement of the thyroid gland, often resulting from various thyroid conditions, including iodine deficiency, hyperthyroidism, or hypothyroidism.
14	Tumor	In the context of thyroid disease, a tumor refers to an abnormal growth or mass that can develop in the thyroid gland. Thyroid tumors can be benign (non-cancerous) or malignant (cancerous).
15	Hypopituitary	Hypopituitary refers to a condition where the pituitary gland does not produce sufficient amounts of one or more hormones, including thyroid-stimulating hormone (TSH), which can impact thyroid function.
16	Psych	This term refers to the psychological or psychiatric aspects related to thyroid disease. Thyroid disorders can sometimes affect mood, cognition, and overall mental well-being.

17	TSH	Thyroid-Stimulating Hormone. A blood test that measures the level of TSH, which is produced by the pituitary gland to regulate the thyroid gland.
18	T3	Triiodothyronine. A measurement of the level of triiodothyronine hormone in the blood.
19	TT4	Total Thyroxine. A measurement of the total level of thyroxine hormone in the blood.
20	T4U	Thyroxine Uptake. A measurement used to assess the binding capacity of thyroxine-binding proteins in the blood.
21	FTI	Free Thyroxine Index. A calculated index representing the concentration of free thyroxine in the blood.
22.1	Thyroid Condition	Hyperthyroid, Hypothyroid, and Normal.
22.2	Outlier	"Yes": hyperthyroid or hypothyroid; "No": normal

- Data Format:

No.	Attribute	Format
1	Age	Numeric
2	Sex	Numeric
3	On Thyroxine	Numeric
4	Query On Thyroxine	Numeric
5	On Antithyroid Medication	Numeric
6	Sick	Numeric
7	Pregnant	Numeric
8	Thyroid Surgery	Numeric
9	I131 Treatment	Numeric
10	Query Hypothyroid	Numeric
11	Query Hyperthyroid	Numeric
12	Lithium	Numeric
13	Goitre	Numeric
14	Tumor	Numeric
15	Hypopituitary	Numeric
16	Psych	Numeric
17	TSH	Numeric
18	T3	Numeric
19	TT4	Numeric
20	T4U	Numeric
21	FTI	Numeric
22.1	Thyroid Condition	Nominal {Hyperthyroid, Hypothyroid, Normal}
22.2	Outlier	Nominal {No, Yes}

3. Explore Data

1 Age to Thyroid Condition:

1.1 Thyroid Condition Distribution in Equal-width 10 Bins of Age Attribute:

Age	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
B1of10	4	18.18%	3	13.64%	15	68.18%	22
B2of10	10	6.58%	3	1.97%	139	91.45%	152
B3of10	16	4.10%	8	2.05%	366	93.85%	390
B4of10	21	3.91%	13	2.42%	503	93.67%	537
B5of10	20	4.64%	10	2.32%	401	93.04%	431
B6of10	26	4.87%	16	3.00%	492	92.13%	534
B7of10	44	5.97%	19	2.58%	674	91.45%	737
B8of10	27	4.43%	16	2.63%	566	92.94%	609
B9of10	20	6.69%	5	1.67%	274	91.64%	299
B10of10	3	4.92%	0	0.00%	58	95.08%	61
Total	191	5.06%	93	2.47%	3488	92.47%	3772

1.2 **Observation**: Bin 1 of age has a higher percentage of hypothyroid and hyperthyroid patients than other bins.

1.3 **Conclusion:** Age is a factor determining the thyroid condition.

2 Sex to Thyroid Condition:

2.1 Thyroid Condition Distribution in Sex Attribute:

Sex	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
Male	149	5.67%	71	2.70%	2409	91.63%	2629
Female	42	3.67%	22	1.92%	1079	94.40%	1143
Total	191	5.06%	93	2.47%	3488	92.47%	3772

2.2 **Observation**: Females have a lower percentage of hypothyroid and hyperthyroid patients than males.

2.3 **Conclusion:** Sex is a factor determining the thyroid condition.

3 On-Thyroxine to Thyroid Condition:

3.1 Thyroid Condition Distribution in On-Thyroxine Attribute:

On Thyroxine	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	191	5.78%	84	2.54%	3032	91.68%	3307
1	0	0.00%	9	1.94%	456	98.06%	465
Total	191	5.06%	93	2.47%	3488	92.47%	3772

3.2 **Observation**: Patients taking Thyroxine have a significantly low percentage of being diagnosed with hypothyroidism.

3.3 **Conclusion:** Whether taking Thyroxine or not is a critical factor in determining hypothyroidism.

4 Query On Thyroxine to Thyroid Condition:

4.1 Thyroid Condition Distribution in Query On-thyroxine Attribute:

Query On Thyroxine	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	188	5.05%	93	2.50%	3442	92.45%	3723
1	3	6.12%	0	0.00%	46	93.88%	49
Total	191	5.06%	93	2.47%	3488	92.47%	3772

4.2 **Observation**: The Query On-Thyroxine has no significant trends in determining Thyroid Conditions.

4.3 **Conclusion:** The Query On-Thyroxine is not a factor attribute. So, we will exclude it.

5 On Antithyroid-Medication to Thyroid Condition:

5.1 Thyroid Condition Distribution in Query On-thyroxine Attribute:

On Antithyroid-Medication	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	190	5.10%	93	2.49%	3446	92.41%	3729
1	1	2.33%	0	0.00%	42	97.67%	43
Total	191	5.06%	93	2.47%	3488	92.47%	3772

5.2 **Observation**: The group of On Antithyroid-Medication is way too smaller than its complementary group (43 / 3729 = 1.15%). It is easy to have a statistical bias if the sample group is already small and its subset's (possibly unnormal patients) percentage of it is also minor.

5.3 **Conclusion:** The group of Query On-Thyroxine is too small to be considered a factor in our model. We will exclude it.

6 Sick to Thyroid Condition:

6.1 Thyroid Condition Distribution in Sick Attribute:

Sick	**Hypothyroid**		**Hyperthyroid**		**Normal**		**Total**
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	180	4.96%	93	2.56%	3354	92.47%	3627
1	11	7.59%	0	0.00%	134	92.41%	145
Total	191	5.06%	93	2.47%	3488	92.47%	3772

6.2 **Observation**: The group of Sick patients is way too smaller than its complementary group (145 / 3627= 4%). And it has no significant data showing a high percentage of unnormal patients.

6.3 **Conclusion:** The group of Sick patients is too small and shows no significant data for determining the Thyroid Condition. We will exclude it.

7 Pregnant to Thyroid Condition:

7.1 Thyroid Condition Distribution in Pregnant Attribute:

Pregnant	**Hypothyroid**		**Hyperthyroid**		**Normal**		**Total**
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	191	5.14%	93	2.50%	3435	92.36%	3719
1	0	0.00%	0	0.00%	53	100.00%	53
Total	191	5.06%	93	2.47%	3488	92.47%	3772

7.2 **Observation**: The group of Pregnant patients is way too smaller than its complementary group (53 / 3719= 1.43%). It is easy to have a statistical bias if the sample group is already small and its subset's (possibly unnormal patients) percentage of it is also minor.

7.3 **Conclusion:** The group of Pregnant patients is too small to be considered a factor in our model. We will exclude it.

8 Thyroid Surgery to Thyroid Condition:

8.1 Thyroid Condition Distribution in Thyroid Surgery Attribute:

Thyroid Surgery	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	191	5.14%	91	2.45%	3436	92.39%	3718
1	0	0.00%	2	3.77%	52	98.11%	54
Total	191	5.06%	93	2.47%	3488	92.47%	3772

8.2 **Observation**: The group of Thyroid Surgery patients is way too smaller than its complementary group (54 / 3718= 1.45%). And it has no significant data showing a high percentage of unnormal patients.

8.3 **Conclusion:** The group of Thyroid Surgery patients is too small and shows no significant data for determining the Thyroid Condition. We will exclude it.

9 I131 Treatment to Thyroid Condition:

9.1 Thyroid Condition Distribution in I131 Treatment Attribute:

I131 Treatment	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	188	5.06%	91	2.45%	3436	92.49%	3715
1	3	5.26%	2	3.51%	52	91.23%	57
Total	191	5.06%	93	2.47%	3488	92.47%	3772

9.2 **Observation**: The group of I131 Treatment patients is way too smaller than its complementary group (57 / 3715= 1.45%). And the two groups, those who are taking the treatment and those who are not, show no significant difference in the percentage of unnormal patients.

9.3 **Conclusion:** The group of I131 Treatment patients is too small and shows no significant data for determining the Thyroid Condition. We will exclude it.

10 Query Hypothyroid to Thyroid Condition:

10.1 Thyroid Condition Distribution in Query Hypothyroid Attribute:

Query Hypothyroid	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	164	4.64%	82	2.32%	3292	93.05%	3538
1	27	11.54%	11	4.70%	196	83.76%	234
Total	191	5.06%	93	2.47%	3488	92.47%	3772

10.2 **Observation**: The group of the sample who queried hypothyroid has twice the percentage of being diagnosed as hypothyroid than the people who did not.

10.3 **Conclusion:** The attribute of Query Hypothyroid is a critical factor in determining whether a patient is hypothyroid or not.

11 Query Hyperthyroid to Thyroid Condition:

11.1 Thyroid Condition Distribution in Query Hyperthyroid Attribute:

Query Hyperthyroid	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	179	5.06%	90	2.54%	3271	92.40%	3540
1	12	5.17%	3	1.29%	217	93.53%	232
Total	191	5.06%	93	2.47%	3488	92.47%	3772

11.2 **Observation**: For the patients who queried hyperthyroid, have a lower percentage of being diagnosed as hyperthyroid. But its instances of being diagnosed as hyperthyroid are too few (only 3), which can introduce a statistical bias because its sample size is not large enough.

11.3 **Conclusion:** The attribute of Query Hyperthyroid cannot be included as a factor in our model because its subset of patients who queried hyperthyroid is not large enough compared to its complementary group.

12 Lithium to Thyroid Condition:

12.1 Thyroid Condition Distribution in Lithium Attribute:

Lithium	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	190	5.06%	93	2.48%	3470	92.46%	3753
1	1	5.26%	0	0.00%	18	94.74%	19
Total	191	5.06%	93	2.47%	3488	92.47%	3772

12.2 **Observation**: The group of patients who took lithium treatment is way too smaller than its complementary group (19 / 3753 = 0.5%). It is easy to have a statistical bias if the sample group is already small and its subset's (possibly unnormal patients) percentage of it is also minor.

12.3 **Conclusion:** The group of patients who took Lithium Treatment is too small to be considered a factor in our model. We will exclude it.

13 Goitre to Thyroid Condition:

13.1 Thyroid Condition Distribution in Goitre Attribute:

Goitre	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	191	5.11%	93	2.49%	3455	92.40%	3739
1	0	0.00%	0	0.00%	33	100.00%	33
Total	191	5.06%	93	2.47%	3488	92.47%	3772

13.2 **Observation**: The group of patients who had Goitre is way too smaller than its complementary group (33 / 3739 = 0.88%). It is easy to have a statistical bias if the sample group is already small and its subset's (possibly unnormal patients) percentage of it is also minor.

13.3 **Conclusion:** The group of patients who goitre is too small to be considered a factor in our model. We will exclude it.

14 Tumor to Thyroid Condition:

14.1 Thyroid Condition Distribution in Tumor Attribute:

Tumor	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	185	5.03%	91	2.47%	3401	92.49%	3677
1	6	6.32%	2	2.11%	87	91.58%	95
Total	191	5.06%	93	2.47%	3488	92.47%	3772

14.2 **Observation**: The patients who had tumor show no higher chance of diagnosing as unnormal (hypothyroid or hyperthyroid).

14.3 **Conclusion:** The Tumor attribute is not a critical factor to determine whether a person is potentially having hypothyroid or hyperthyroid. We will exclude it.

15 Hypopituitary to Thyroid Condition:

15.1 Thyroid Condition Distribution in Hypopituitary Attribute:

Hypopituitary	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	191	5.06%	93	2.47%	3487	92.47%	3771
1	0	0.00%	0	0.00%	1	100.00%	1
Total	191	5.06%	93	2.47%	3488	92.47%	3772

15.2 **Observation**: There is only one instance of patients who had hypopituitary, which is way too smaller than its complementary group (1 / 3771 = 0.027%). It is easy to have a statistical bias if the sample group is too small.

15.3 **Conclusion:** To avoid unnecessary and potentially biased attributes in our model, we will not include the Hypopituitary attribute.

16 Psych to Thyroid Condition:

16.1 Thyroid Condition Distribution in Psych Attribute:

Psych	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
0	183	5.10%	93	2.59%	3310	92.30%	3586
1	8	4.30%	0	0.00%	178	95.70%	186
Total	191	5.06%	93	2.47%	3488	92.47%	3772

16.2 **Observation**: There is no trend showing people who were mental illness had a higher chance of diagnosing with hypothyroid. However, they had a lower chance of being diagnosed as hyperthyroid than people who were mentally healthy. This result contradicts the presupposition, "Thyroid disorders can sometimes affect mood, cognition, and overall mental well-being." Hence, we assume there is a statistical bias in the sample group because the sample group is not large enough.

16.3 **Conclusion:** The Psych attribute is not a factor in determining Thyroid Condition. Hence, we will exclude it.

17 TSH to Thyroid Condition:

17.1 Thyroid Condition Distribution in Equal-width 10 Bins of TSH Attribute:

TSH	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
B1of10	188	5.06%	48	1.29%	3483	93.65%	3719
B2of10	2	7.69%	21	80.77%	3	11.54%	26
B3of10	1	9.09%	8	72.73%	2	18.18%	11
B4of10	0	0.00%	8	100.00%	0	0.00%	8
B5of10	0	0.00%	2	100.00%	0	0.00%	2
B6of10	0	-	0	-	0	-	0
B7of10	0	-	0	-	0	-	0
B8of10	0	0.00%	1	100.00%	0	0.00%	1
B9of10	0	0.00%	3	100.00%	0	0.00%	3
B10of10	0	0.00%	2	100.00%	0	0.00%	2
Total	191	5.06%	93	2.47%	3488	92.47%	3772

17.2 **Observation**: The patients in bin-1 of TSH have a lower chance (48 / 3719 = 1.29%) of getting hyperthyroid than the whole sample group's average (93/ 3772 = 2.47%). If we look at bin-2 to bin-4, the patients had a very high possibility of being diagnosed as hyperthyroid (from 72.73% to 100%).

17.3 **Conclusion:** The TSH attribute is a crucial factor in determining whether a patient is highly possible of having hyperthyroidism.

18 T3 to Thyroid Condition:

18.1 Thyroid Condition Distribution in Equal-width 10 Bins of T3 Attribute:

T3	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
B1of10	25	7.91%	64	20.25%	227	71.84%	316
B2of10	125	5.41%	24	1.04%	2162	93.55%	2311
B3of10	37	3.79%	5	0.51%	933	95.69%	975
B4of10	4	3.51%	0	0.00%	110	96.49%	114
B5of10	0	0.00%	0	0.00%	37	100.00%	37
B6of10	0	0.00%	0	0.00%	10	100.00%	10
B7of10	0	0.00%	0	0.00%	6	100.00%	6
B8of10	0	0.00%	0	0.00%	1	100.00%	1
B9of10	0	0.00%	0	0.00%	1	100.00%	1
B10of10	0	0.00%	0	0.00%	1	100.00%	1
Total	191	5.06%	93	2.47%	3488	92.47%	3772

18.2 **Observation**: The patients in bin-1 of TSH have a higher possibility (64 / 316 = 20.25%) of having hyperthyroid than the whole sample group's average (93/ 3772 = 2.47%). If we look at bin-2 and bin-3, the percentage of patients diagnosed as hyperthyroid significantly decreases sequentially (1.04% to 0.51%).

18.3 **Conclusion:** The T3 attribute is a crucial factor in determining whether a patient is possibly having hyperthyroidism or not.

19 TT4 to Thyroid Condition:

19.1 Thyroid Condition Distribution in Equal-width 10 Bins of TT4 Attribute:

TT4	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
B1of10	1	1.23%	62	76.54%	18	22.22%	81
B2of10	94	11.59%	30	3.70%	687	84.71%	811
B3of10	82	3.81%	1	0.05%	2071	96.15%	2154
B4of10	14	2.44%	0	0.00%	559	97.56%	573
B5of10	0	0.00%	0	0.00%	114	100.00%	114
B6of10	0	0.00%	0	0.00%	30	100.00%	30
B7of10	0	0.00%	0	0.00%	6	100.00%	6
B8of10	0	-	0	-	0	-	0
B9of10	0	0.00%	0	0.00%	1	100.00%	1
B10of10	0	0.00%	0	0.00%	2	100.00%	2
Total	191	5.06%	93	2.47%	3488	92.47%	3772

19.2 **Observation**:

19.2.1 **Hyperthyroid**: The patients in bin-1 of TT4 have a much higher possibility (62 / 81 = 76.54%) of having hyperthyroid than the whole sample group's average (93/ 3772 = 2.47%). If we look at bin-3 and bin-4, the sample group's instance numbers are large (2154 and 573), but meanwhile, the percentage of patients diagnosed as hyperthyroid significantly decreases sequentially (0.05% to 0%).

19.2.2 **Hypothyroid**: The patients in bin-2 of TT4 have a higher percentage (94 / 811 = 11.59%) of having hypothyroid than the whole sample group's average (191/ 3772 = 5.06%).

19.3 **Conclusion:** The TT4 attribute is a very critical factor in determining whether a patient is possibly having hyperthyroidism or not. It is also an important factor in determining whether a patient is potentially having hypothyroidism.

20 T4U to Thyroid Condition:

20.1 Thyroid Condition Distribution in Equal-width 10 Bins of T4U Attribute:

T4U	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
B1of10	0	0.00%	0	0.00%	7	100.00%	7
B2of10	1	2.70%	1	2.70%	35	94.59%	37
B3of10	24	5.10%	9	1.91%	438	92.99%	471
B4of10	111	4.83%	40	1.74%	2147	93.43%	2298
B5of10	42	5.86%	33	4.60%	642	89.54%	717
B6of10	9	6.57%	9	6.57%	119	86.86%	137
B7of10	4	5.71%	1	1.43%	65	92.86%	70
B8of10	0	0.00%	0	0.00%	28	100.00%	28
B9of10	0	0.00%	0	0.00%	6	100.00%	6
B10of10	0	0.00%	0	0.00%	1	100.00%	1
Total	191	5.06%	93	2.47%	3488	92.47%	3772

20.2 **Observation**: The T4U attribute shows no trends in determining whether patients have a higher possibility of being diagnosed hypothyroid or hyperthyroid.

20.3 **Conclusion:** The T4U attribute is not a factor in determining whether a patient is potentially hypothyroid or hyperthyroid. So, we will exclude it.

21 FTI to Thyroid Condition:

21.1 Thyroid Condition Distribution in Equal-width 10 Bins of FTI Attribute:

FTI	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
B1of10	0	0.00%	90	63.83%	51	36.17%	141
B2of10	180	6.50%	3	0.11%	2586	93.39%	2769
B3of10	11	1.44%	0	0.00%	755	98.56%	766
B4of10	0	0.00%	0	0.00%	79	100.00%	79
B5of10	0	0.00%	0	0.00%	10	100.00%	10
B6of10	0	0.00%	0	0.00%	3	100.00%	3
B7of10	0	0.00%	0	0.00%	2	100.00%	2
B8of10	0	0.00%	0	0.00%	1	100.00%	1
B9of10	0	-	0	-	0	-	0
B10of10	0	0.00%	0	0.00%	1	100.00%	1
Total	191	5.06%	93	2.47%	3488	92.47%	3772

21.2 **Observation**:

21.2.1 **Hyperthyroid**: The patients in bin-1 have a much higher percentage of being diagnosed as hyperthyroid (90 / 141 = 63.83%). In bin-2 and bin-3, the percentage of being diagnosed as hyperthyroid are 0.11% (= 2 / 2769) and 0% (= 0 / 766) separately, which are much lower than the average of the whole sample group (93 / 3773 = 2.46%).

21.2.2 **Hypothyroid:** Most of the patients diagnosed as hypothyroid are in bin-2, whereas its neighbours, bin-1 and bin-3, have less possibility of being diagnosed as hypothyroid, which are 0% (= 0 / 141) and 1.44% (= 11 / 766) separately compared to the sample group's overall average (191 / 3772 = 5.06%).

21.3 **Conclusion:** The FTI attribute is a very crucial factor in determining whether a patient is with hyperthyroidism. It is also a factor determining a patient is less likely to have hypothyroidism.

4. Verify Data Quality

- Missing Data: None
- Error Data: None

Data Preparation

1. Select Data

Attribute	Included / Excluded	Reasons
Age	**Included**	Bin 1 of age has a higher percentage of hypothyroid and hyperthyroid patients than other bins. Hence, the attribute “Age” is a factor determining the thyroid condition.
Sex	**Included**	Females have a lower percentage of hypothyroid and hyperthyroid patients than males. Hence, the attribute “Sex” is a factor determining the thyroid condition.
On Thyroxine	**Included**	Patients taking “on Thyroxine” have a significantly low percentage of being diagnosed with hypothyroidism. Whether taking Thyroxine or not is a critical factor in determining hypothyroidism.
Query On Thyroxine	Excluded	The attribute “Query On-Thyroxine" has no significant trends in determining Thyroid Conditions.
On Antithyroid Medication	Excluded	Given that the sample size of the "On Antithyroid-Medication" group is already small, and the percentage of potentially abnormal patients within this group is minor, it is concluded that this attribute does not provide sufficient data to be considered a significant factor in the model.
Sick	Excluded	The "Sick" group does not show significant data regarding a high percentage of abnormal patients compared to the complementary group and does not contribute significantly to determining the thyroid condition.
Pregnant	Excluded	Given that the sample size of the "Pregnant" group is already small, and the percentage of potentially abnormal patients within this group is minor, it is concluded that this attribute does not provide sufficient data to be considered a significant factor in the model. Therefore, it is decided to exclude the "Pregnant" attribute from the analysis.
Thyroid Surgery	Excluded	The attribute " Thyroid Surgery " does not provide distinct information for determining the thyroid condition.
I131 Treatment	Excluded	The attribute "I131 Treatment" does not provide distinct information for determining the thyroid condition.
Query Hypothyroid	**Included**	Considering the substantial difference in the percentage of hypothyroid patients between the group that queried hypothyroid and the group that did not, it is concluded that the "Query Hypothyroid" attribute is an essential factor in determining the thyroid condition.

Query Hyperthyroid	Excluded	The "Query Hyperthyroid" attribute does not provide distinct information for determining the thyroid condition.
Lithium	Excluded	The "Lithium" attribute does not provide distinct information for determining the thyroid condition.
Goitre	Excluded	The "Goitre" attribute does not provide distinct information for determining the thyroid condition.
Tumor	Excluded	The "Tumor" attribute does not provide distinct information for determining the thyroid condition.
Hypopituitary	Excluded	The "Hypopituitary" attribute does not provide distinct information for determining the thyroid condition.
Psych	Excluded	The "Psych" attribute does not provide distinct information for determining the thyroid condition.
TSH	**Included**	The "TSH" attribute plays a significant role in determining the likelihood of a patient being diagnosed with hyperthyroidism.
T3	**Included**	The "T3" attribute plays a significant role in determining the likelihood of a patient being diagnosed with hyperthyroidism.
TT4	**Included**	The "TT4" attribute plays a significant role in determining the likelihood of a patient being diagnosed with hyperthyroidism and hypothyroidism.
T4U	Excluded	The "T4U" attribute does not play a significant role in determining the likelihood of a patient being hypothyroid or hyperthyroid.
FTI	**Included**	The "FTI" attribute plays a significant role in determining the likelihood of a patient being diagnosed with hyperthyroidism.

2. Clean Data

- No Need

3. Construct Data

1 Young Age:

1.1 Separates Equal-width 10 Bins Age Attribute into two Groups, Young Age = 1 (Bin-1) and Young Age = 0 (Bin-2 to Bin 10):

Young Age	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
1	4	18.18%	3	13.64%	15	68.18%	22
0	187	4.99%	90	2.4%	3473	92.61%	3750
Total	191	5.06%	93	2.47%	3488	92.47%	3772

2 TSH Level:

2.1 Discretizes TSH into 3 Groups, Low (< 0.00585), Medium (≥ 0.00585 *and* < 0.037), and High (≥ 0.037):

TSH Level	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
Low (< 0.00585)	0	0.00%	0	0.00%	3399	100.00%	3399
Medium (≥ 0.00585 *and* < 0.037)	183	61.20%	33	11.04%	83	27.76%	299
High (≥ 0.037)	8	10.81%	60	81.08%	6	8.11%	74
Total	191	5.06%	93	2.47%	3488	92.47%	3772

3 T3 Level:

3.1 Discretizes T3 into 3 Groups, Low (< 0.0105), Medium (≥ 0.0105 *and* < 0.0265), and High (≥ 0.0265):

T3 Level	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
Low (< 0. 0105)	20	7.43%	62	23.05%	187	69.52%	269
Medium (≥ 0.0105 *and* < 0. 0265)	164	5.30%	31	1.00%	2898	93.70%	3093
High (≥ 0.0265)	7	1.71%	0	0.00%	403	98.29%	410
Total	191	5.06%	93	2.47%	3488	92.47%	3772

4 TT4 Level:

4.1 Discretizes TT4 into 3 Groups, Low (< 0.0525), Medium (≥ 0.0525 *and* < 0.0895), and High (≥ 0.0895):

TT4 Level	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
Low (< 0.0525)	2	1.74%	76	66.09%	37	32.17%	115
Medium (≥ 0.0525 *and* < 0.0895)	104	11.90%	17	1.95%	753	86.16%	874
High (≥ 0.0895)	85	3.05%	0	0.00%	2698	96.95%	2783
Total	191	5.06%	93	2.47%	3488	92.47%	3772

5 FTI Level:

5.1 Discretizes FTI into 3 Groups, Low (< 0.064965), Medium (≥ 0.064965 *and* < 0.08233), and High (≥ 0.08233):

FTI Level	Hypothyroid		Hyperthyroid		Normal		Total
	Instance	Percentage	Instance	Percentage	Instance	Percentage	Instance
Low (< 0.064965)	2	1.33%	93	62.00%	55	36.67%	150
Medium (≥ 0.064965 *and* < 0.08233)	68	20.86%	0	0.00%	258	79.14%	326
High (≥ 0.08233)	121	3.67%	0	0.00%	3175	96.33%	3296
Total	191	5.06%	93	2.47%	3488	92.47%	3772

4. Integrate Data

- **None**

5. Format Data

- **To better use the dataset in all the models of K-Means, Farthest First, Local Outlier Factor, and Isolation, we reformat all the attributes into the numeric type except the class.**

No.	Attribute	Original Format	Revised Format
1	Young Age	Nominal { 0 = No, 1 = Yes}	Numeric
2	Sex	Nominal { 0 = male, 1 = female}	Numeric
3	On Thyroxine	Nominal { 0 = No, 1 = Yes}	Numeric
4	Query Hypothyroid	Nominal { 0 = No, 1 = Yes}	Numeric
5	TSH Level	Nominal { Low, Medium, High}	Numeric (TSH Level=High)

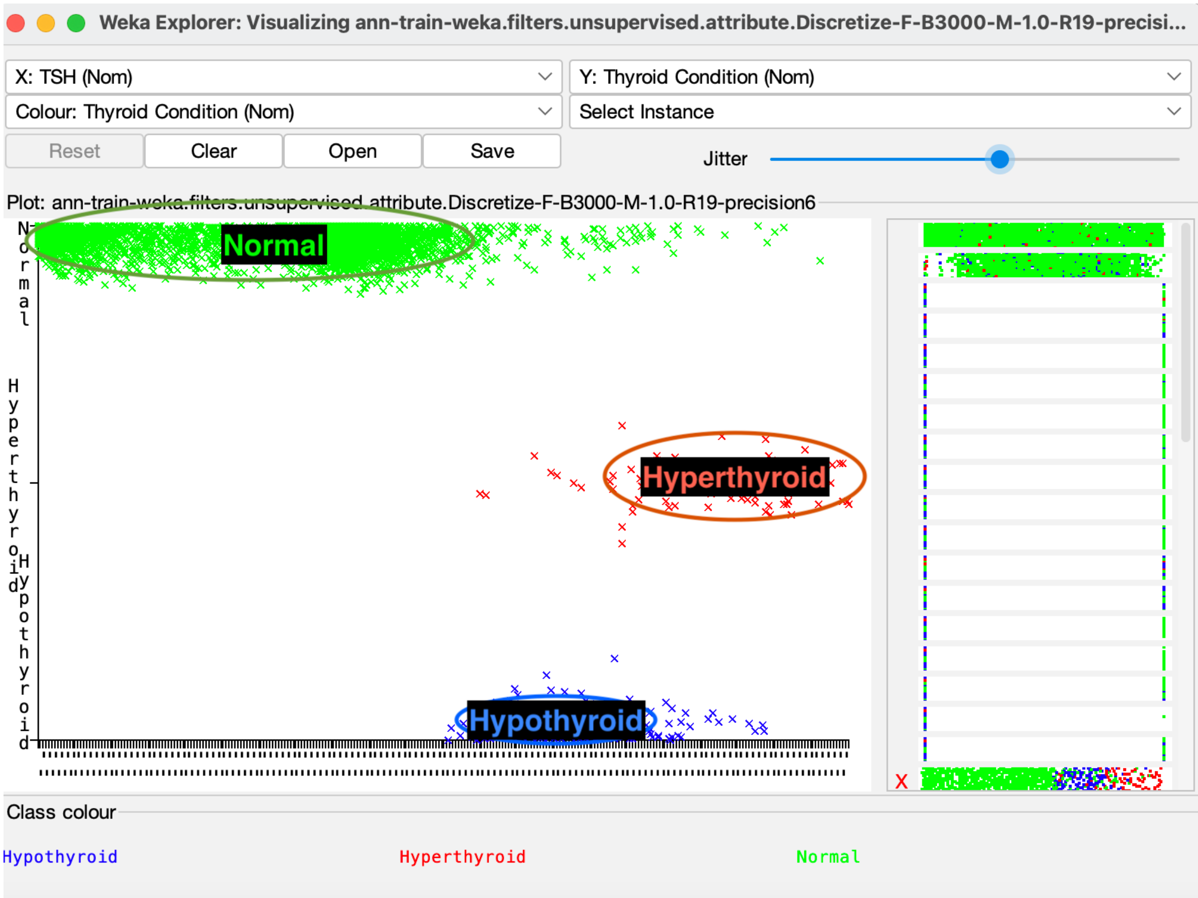
			Numeric (TSH Level=Medium)
			Numeric (TSH Level=Low)
6	T3 Level	Nominal { Low, Medium, High}	Numeric (T3 Level=High)
			Numeric (T3 Level=Medium)
			Numeric (T3 Level=Low)
7	TT4 Level	Nominal { Low, Medium, High}	Numeric (TT4 Level=High)
			Numeric (TT4 Level=Medium)
			Numeric (TT4 Level=Low)
8	FTI Level	Nominal { Low, Medium, High}	Numeric (FTI Level=High)
			Numeric (FTI Level=Medium)
			Numeric (FTI Level=Low)

- **Tabulate statistics and counts**

Attribute	Count		Min	Max	Mean	StdDev
	0	1				
Young Age	3750	22	0	1	0.006	0.076
Sex	2629	1143	0	1	0.303	0.460
On Thyroxine	3307	465	0	1	0.123	0.329
Query Hypothyroid	3538	234	0	1	0.062	0.241
TSH Level=High	3698	74	0	1	0.020	0.139
TSH Level=Medium	3473	299	0	1	0.079	0.270
TSH Level=Low	373	3399	0	1	0.901	0.299
T3 Level=High	3362	410	0	1	0.109	0.311
T3 Level=Medium	679	3093	0	1	0.820	0.382
T3 Level=Low	3503	269	0	1	0.071	0.257
TT4 Level=High	989	2783	0	1	0.738	0.440
TT4 Level=Medium	2898	874	0	1	0.232	0.422
TT4 Level=Low	3657	115	0	1	0.030	0.172
FTI Level=High	476	329	0	1	0.874	0.332
FTI Level=Medium	3446	326	0	1	0.086	0.281
FTI Level=Low	3622	150	0	1	0.040	0.195
Thyroid Condition=Hyperthyroid	93		-	-	-	-
Thyroid Condition=Hypothyroid	191		-	-	-	-
Thyroid Condition=Normal	3488		-	-	-	-

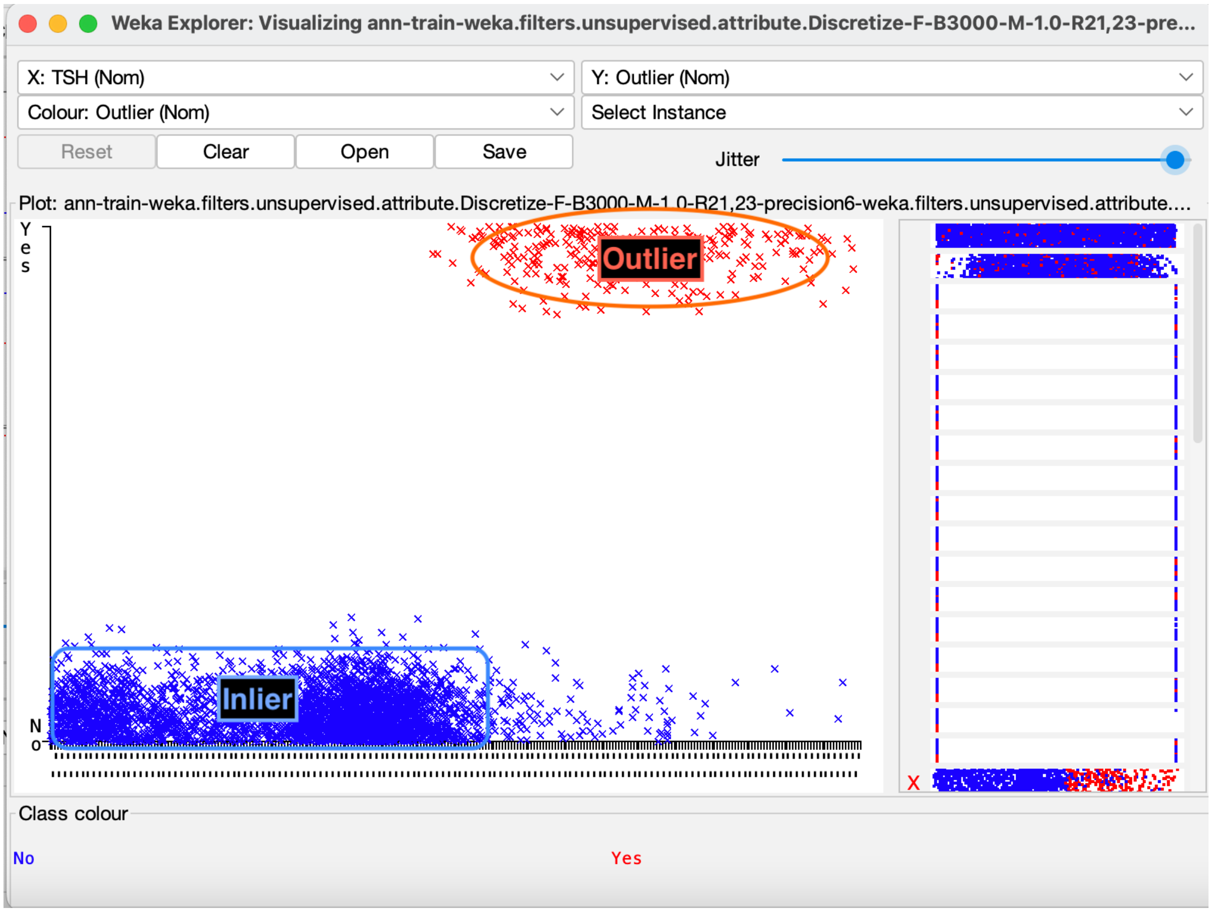
6. Three Interesting Charts:

1 X: TSH vs Y: Thyroid Condition (Colour: Thyroid Condition):



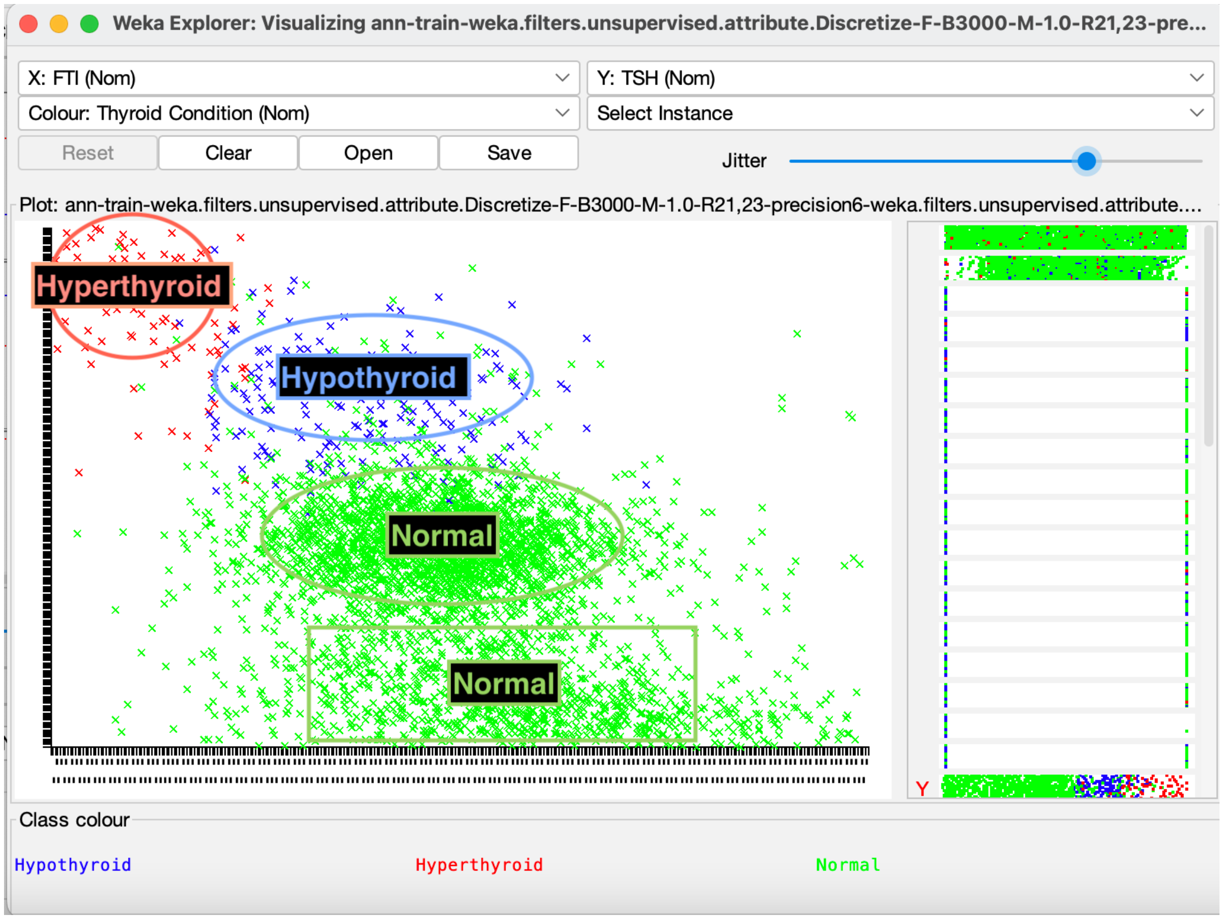
Observation: If we discretize TSH attribute into 280 equal-frequency bins, we can see from the plot that its density level effectively categorizes the patients into Normal, Hypothyroid, and Hyperthyroid groups.

2 X: TSH vs Y: Outlier (Colour: Outlier):



Observation: If we discretize TSH attribute into 280 equal-frequency bins, we can see from the plot that its density level very effectively categorizes the patients into Normal and Unnormal (Hypothyroid or Hyperthyroid) groups.

3 X: FTI vs Y: TSH (Colour: Thyroid Condition):



Observation: If we discretize the FTI attribute into 324 equal-frequency bins and the TSH attribute into 280 equal-frequency bins, we can see from the plot that their relative values effectively categorize the patients into Normal, Hypothyroid, and Hyperthyroid groups.

Modeling

1. Select Modeling Technique:

1 Clustering:

1.1 SimpleKMeans

1.2 FartherstFirst

2 Outlier Detection:

2.1 Local Outlier Factor

2.2 Isolation Forest

2. Generate Test Design:

1 Clustering:

1.1 SimpleKMeans:

1.1.1 Cluster mode: Classes to clusters evaluation - (Nom) Thyroid Condition.

1.1.2 Find the best k = ? by using the Elbow Point method.

1.1.3 Identify each generated cluster to which group in Thyroid Condition it belongs.

1.1.4 Merge all the identified clusters into their own groups.

1.2 FartherstFirst:

1.2.1 Cluster mode: Classes to clusters evaluation - (Nom) Thyroid Condition

1.2.2 Find the best k = ? by increasing the value of k, and stop when the newly generated clusters no longer further filter out instances from the previous clusters, where they were classified as the group they don't belong to.

1.2.3 Identify each generated cluster to which group in Thyroid Condition it belongs.

1.2.4 Merge all the identified clusters into their own groups.

2 Outlier Detection:

2.1 Local Outlier Factor:

2.1.1 Test options: 10 Folds Cross-validation

2.2 Isolation Forest:

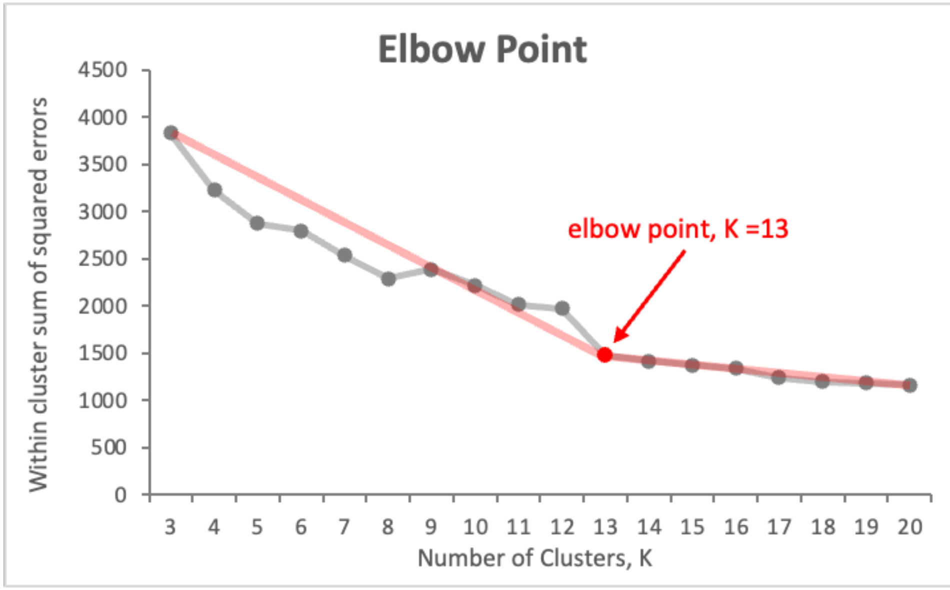
2.2.1 Test options: 10 Folds Cross-validation

3. Build Model

1 Clustering:

1.1 SimpleKMeans:

1.1.1 Elbow Graph:



1.1.2 Clusters (K = 13):

Cluster	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Hypothyroid	1	1	6	0	0	7	8	0	0	88	76	4	0
Hyperthyroid	0	0	0	0	0	0	0	87	0	6	0	0	0
Normal	125	1603	178	589	278	34	80	21	31	21	59	22	447
Classified (N = Normal, H = Hyperthyroid, S = Hypothyroid)	N	N	N	N	N	N	N	H	N	S	S	N	N

1.1.3 Merge Clusters:

Cluster	**0**	**1**	**2**
Hypothyroid	164	0	27
Hyperthyroid	6	87	0
Normal	80	21	3387
Precision (%)	65.60	80.56	99.21

1.2 <u>FartherstFirst</u>:

1.2.1 No Further Clustering After K = 31:

```
30   31   32   33   34   35   36   37   38   39   40   41
15    1    0    2    1    0    0    0    0    0    0    0
 0    1    0    0    0    2    0    0    0    0    0    0
 0    1    2    0    1    2  154  341   49    6   16    8
```

1.2.2 Clusters (K = 31):

Cluster	0	1	2	3	4	5	6	7	8	9	
Hypothyroid	0	0	1	15	0	11	0	41	1	0	
Hyperthyroid	0	2	13	0	17	0	0	4	5	0	
Normal	2022	1	0	2	1	14	20	2	4	73	
Classified (N = Normal, H = Hyperthyroid, S = Hypothyroid)	N	H	H	S	H	N	N	S	H	N	
Cluster	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	
Hypothyroid	0	0	0	1	0	0	0	1	50	32	
Hyperthyroid	0	0	0	7	1	0	0	0	0	0	
Normal	12	18	6	2	1	7	9	35	10	7	
Classified (N = Normal, H = Hyperthyroid, S = Hypothyroid)	N	N	N	H	H	N	N	N	S	S	
Cluster	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Hypothyroid	7	0	0	0	0	0	1	3	7	1	15
Hyperthyroid	0	0	5	0	0	36	0	0	0	0	0
Normal	0	16	0	245	260	4	0	4	0	4	0
Classified (N = Normal, H = Hyperthyroid, S = Hypothyroid)	S	N	H	N	N	H	S	N	S	N	S

1.2.3 Merge Clusters:

Cluster	0	1	2
Hypothyroid	153	3	16
Hyperthyroid	4	86	0
Normal	21	13	2745
Precision (%)	85.96	84.31	99.42

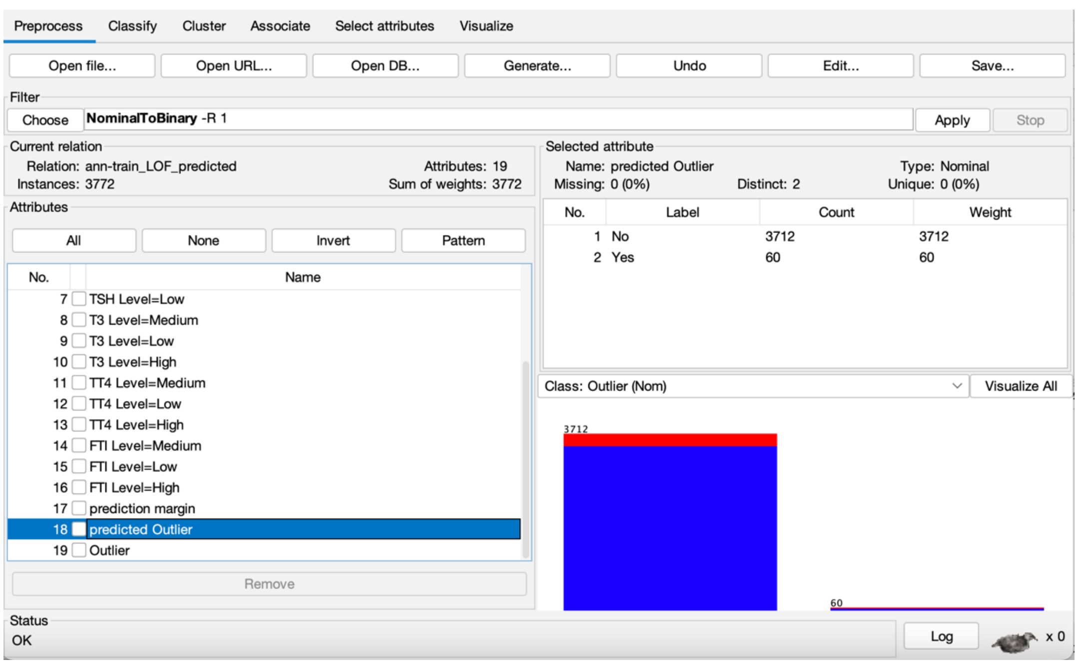
2 Outlier Detection:

2.1 Local Outlier Factor:

2.1.1 Result in Weka:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3462               91.7815 %
Incorrectly Classified Instances       310                8.2185 %
Kappa statistic                          0.0745
Mean absolute error                      0.0942
Root mean squared error                  0.2518
Relative absolute error                 67.5345 %
Root relative squared error             95.4378 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.940    0.928      0.988    0.957      0.100    0.848     0.980     No
                 0.060    0.012    0.283      0.060    0.099      0.100    0.848     0.276     Yes
Weighted Avg.    0.918    0.870    0.880      0.918    0.892      0.100    0.848     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 3445   43 |    a = No
  267   17 |    b = Yes
```

2.1.2 Predicted outlier of the result (datasetName_LOF.arff):

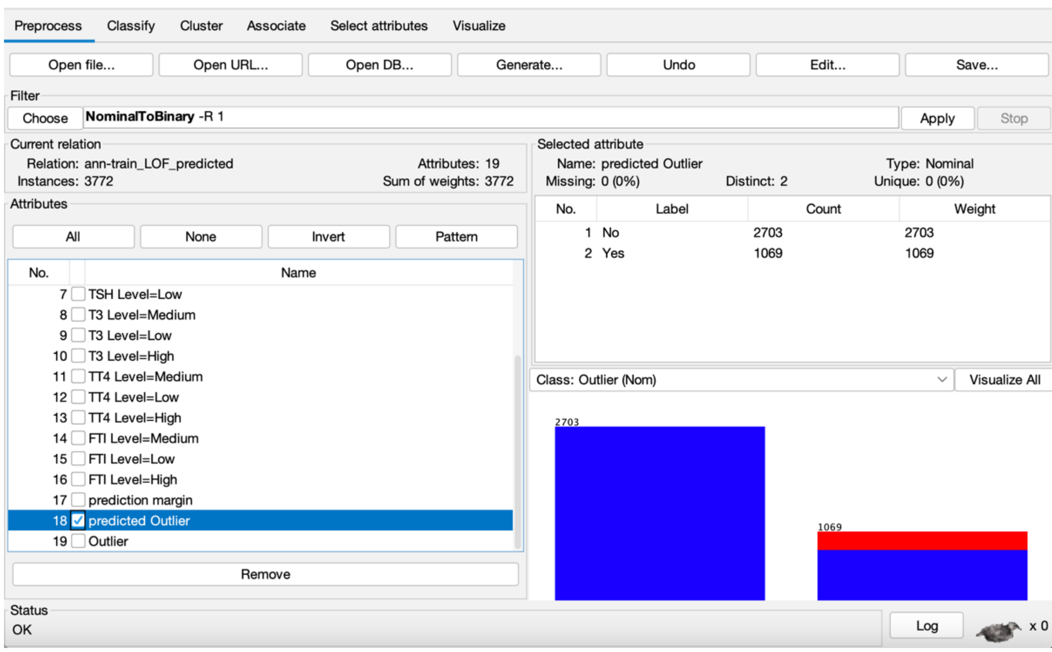


2.2 Isolation Forest:

2.2.1 Result in Weka:

```
=== Summary ===

Correctly Classified Instances        2987               79.1888 %
Incorrectly Classified Instances       785               20.8112 %
Kappa statistic                          0.3415
Mean absolute error                      0.4131
Root mean squared error                  0.4259
Relative absolute error                296.2989 %
Root relative squared error            161.4096 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.775    0.000    1.000      0.775    0.873      0.454    0.933     0.994     No
                 1.000    0.225    0.266      1.000    0.420      0.454    0.933     0.544     Yes
Weighted Avg.    0.792    0.017    0.945      0.792    0.839      0.454    0.933     0.960

=== Confusion Matrix ===

    a    b   <-- classified as
 2703  785 |    a = No
    0  284 |    b = Yes
```

2.2.2 Predicted outlier of the result (datasetName_ISF.arff):



4. Assess Model

1 Clustering:

1.1 SimpleKMeans:

1.1.1 Normal: The model has a very high precision (99.21%) in predicting normal patients. It is because the sample data has a high percentage (92.47%) of normal patients. It is easy to find attributes or discretize an attribute into groups to further distinguish normal patients from the unnormal.

1.1.2 Hyperthyroid: We found significant factors in TSH, TT4, and FTI attributes by discretizing them into equal-frequency bins. We then further grouped the bins into 3 groups relevant to the Thyroid Condition. This made our model have high precision in predicting Hyperthyroid patients (80.56%)

1.1.3 Hypothyroid: Our model has less precision in Hypothyroid patients (65.60%) even though we merged all the most relevant generated clusters to increase it. It is because we could not find crucial factors to distinguish Hypothyroid patients from others except the attribute TSH, which has a medium-level density subgroup to separate the patients from the rest with an accuracy of 58.64% (= $\frac{183}{191} \times 61.2\%$).

The subgroups in other attributes with less precision (less than 21%), but they help increase our model's precision to 65.60% by 6.96% (= 65.60% - 58.64%).

1.2 FartherstFirst:

1.2.1 The precision comparison table between SimpleKMeans and FartherstFirst models:

Model	**Precision (%)**		
	Hypothyroid	Hyperthyroid	Normal
SimpleKMeans	65.60	80.56	99.21
FartherstFirst	85.96	84.31	99.42

1.2.2 Observation: The FartherstFirst model has a higher precision in all three Thyroid conditions.

1.2.3 Conclusion: FarthestFirst algorithm initially starts by selecting the initial cluster's centroids that are farthest from each other. This can help in creating well-separated initial clusters, which can lead to better precision. On the other hand, SimpleKMeans initializes the cluster centroids randomly, which might result in starting with less well-separated (suboptimal) points and producing a less precise model.

2 Outlier Detection:

2.1 <u>Local Outlier Factor</u>:

2.1.1 The model has high accuracy (91.78%) in correctly classifying the instances:

Classified Instances	Instance	Percentage
Correctly	3462	91.7815%
Incorrectly	310	8.2185%

2.1.2 The model has a poor specificity ($5.99\% = \frac{17}{284}$) in correctly classifying all the outliers but a high sensitivity in correctly identifying the inliers ($98.77\% = \frac{3445}{3488}$):

- Confusion matrix:

	Classified As Inlier	Classified As Outlier
Actual Inlier	3445	43
Actual Outlier	267	17

1.1 <u>Isolation Forest</u>:

1.1.1 The model has high accuracy (79.19%) in correctly classifying the instances:

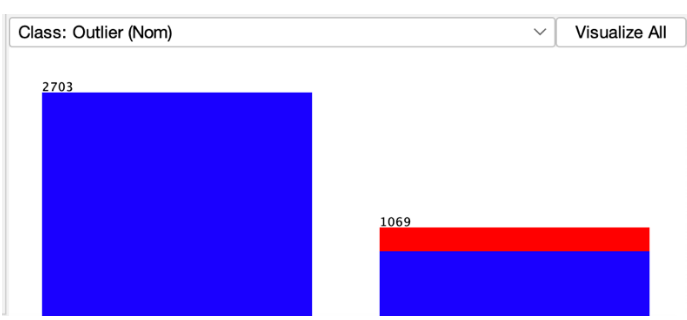
Classified Instances	Instance	<u>Percentage</u>
Correctly	2987	79.1888 %
Incorrectly	785	20.8112 %

1.1.2 The model successfully identifies all the outliers but incorrectly classifies some inliers:

1.1.2.1 Confusion matrix:

```
=== Confusion Matrix ===

   a    b   <-- classified as
 2703  785 |    a = No
    0  284 |    b = Yes
```

1.1.2.2 Predicted outliers visualization:



Evaluation

1. Evaluate Results:

1 **Clustering:** The FartherstFirst model outwins the SimpleKMeans with higher precision in every thyroid condition because of its algorithm in selecting the initial cluster's centroids that are farthest from each other. This secures the initial points to be more well-separated from each other and results in a higher precise model.

2 **Outlier Detection:** The Isolation Forest model can correctly classify all the outliers but with the disadvantage of falsely over-classifying some of the inliers. On the contrary, the Local Outlier Factor model has a very poor specificity (5.99%) of correctly identifying the outliers, but it has a high sensitivity (91.78%) in correctly classifying the inliers.

3 **The comparison table of the Local Outlier Factor and Isolation Forest model's Sensitivity and Specificity:**

	Sensitivity (%)	Specificity (%)
Local Outlier Factor	91.78	5.99
Isolation Forest	77.49	100

2. Review Process:

1 **Clustering:** The trickiest part of finding the most influential factors in the attributes is that we have to discretize numeric data into equal bins, observe the Thyroid condition distribution trending among those bins, and finally, group them into beneficial groups. It took time to achieve this, not to mention that we tried different bins to better discretize them into more appropriate groups.

2 **Outlier Detection:** We use the same selected attributes in both the Clustering and Outlier Detection Models because the influential factors in determining the Hyperthyroid and Hypothyroid patients (both are outliers) can also play an important role in determining outliers. This saves us time from generating another training dataset.

3. Determine Next Steps:

1 **Clustering:** We are satisfied with our FartherstFirst model's high precision.

2 **Outlier Detection:** Although the Local Outlier Factor Model cannot effectively identify the outliers, but our Isolation Forest Model successfully identifies all the outliers with the cost of over-classifying some of the inliers.

3 Decision: No further iteration of the process. Move to next step.

Discussion of Results

1 The Combined Outlier Detection Result:

1.1 The screenshot of combinedResults_datasetname.xlsx:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	'Young Ag	Sex	'On Thyro	'Query Hy	'TSH Level	'TSH Level	'TSH Level	'T3 Level=	'T3 Level=	'T3 Level=	'TT4 Level	'TT4 Level	'TT4 Level	'FTI Level=	'FTI Level=	'FTI Level=	Outlier	LOF: predi	LOF: predi	'prediction	'predicted	Ensemble
88	0	1	0	1	0	0	1	1	0	0	0	1	0	0	1	0	No	-0.050851	1	-0.369895	1	2
94	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	No	-0.204422	1	-0.442793	1	2
116	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1	No	-0.014476	1	-0.144461	1	2
354	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	Yes	0.045003	1	0.458594	1	2
369	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	1	Yes	0.010156	1	0.399748	1	2
443	0	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	No	-0.445404	1	-0.400967	1	2
486	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	No	-0.008304	1	-0.198885	1	2
609	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	No	-0.008304	1	-0.198885	1	2
630	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	No	-0.008304	1	-0.198885	1	2
741	1	1	0	1	0	1	0	1	0	0	1	0	0	1	0	0	Yes	0.50052	1	0.406235	1	2
752	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	1	Yes	0.416341	1	0.362727	1	2

1.2 **Observation**: The ensemble value of two does not guarantee the entry is actually an outlier. In fact, there are only 16 out of 59 are actual outliers.

1.3 **Conclusion**: This phenomenon can be caused by the inaccurate prediction of outliers by both models, the Local Outlier Factor and Isolation Forest.

2 **Clustering**: The "FarthestFirst" algorithm chooses the initial cluster centers by maximizing the distance between them. This initialization strategy helps in avoiding convergence to suboptimal solutions and can lead to better results compared to random initialization, which is commonly used in k-means. In contrast, the "Single k-means" algorithm often uses random initialization, which may result in a higher likelihood of converging to suboptimal solutions.

3 **Outlier Detection:** The Isolation Forest algorithm tends to be more sensitive to variations and deviations in the data, this sensitivity is primarily due to the underlying principles and mechanisms of the Isolation Forest algorithm. By design, Isolation Forest is more inclined to isolate data points that exhibit variations or deviations from the majority, resulting in a higher sensitivity to outliers.

4 Comparison Between Clustering and Outlier Detection:

4.1 **The disadvantage of the Outlier Detection Method**: The Local Outlier Factor Model identifies data points that deviate significantly from their local neighbourhood. It is possible that the data points are well distributed, causing the result of only a few outliers. Meanwhile, the Isolation Forest Model inclines to isolate data points that exhibit variations or deviations from the majority, resulting in a higher sensitivity to outliers.

4.2 **The advantage of the Clustering Method:** The Clustering Method can categorize each data point into a group. Based on domain knowledge, we can then further cluster them into ideal groups according to our needs. In this case, our goal is to cluster all the data points into Normal, Hyperthyroid, and Hypothyroid.

5 **Overall:** The Local Outlier Factor Model cannot effectively identify outliers if the data points are relatively well-distributed. The Isolation Forest Model can be so sensitive that it identifies all the data outside the highly dense centroid as outliers. In this scenario, we should use the Clustering Method over the Outlier Detection Method. Firstly, classify each data point with its nearest neighbours into a group. Secondly, identify each group by the major subgroup within it. Finally, test the model with test datasets and iterate to modify and improve the model.

Conclusion

In conclusion, the analysis of the Thyroid Disease dataset using clustering and outlier detection models has provided valuable insights into the factors determining thyroid conditions. The findings reveal that age, sex, the usage of Thyroxine, the attribute of Query Hypothyroid, as well as the attributes TSH, T3, TT4, and FTI, are influential factors in determining the thyroid condition.

The analysis demonstrates that age is a determinant, with Bin 1 showing a higher percentage of hypothyroid and hyperthyroid patients compared to other age bins. Similarly, sex is found to be a determining factor, with females having a lower percentage of hypothyroid and hyperthyroid patients compared to males.

The usage of Thyroxine emerges as a critical factor in determining hypothyroidism, as patients taking Thyroxine exhibit a significantly lower percentage of being diagnosed with hypothyroidism. Additionally, the attribute of Query Hypothyroid proves to be essential, with individuals who queried hypothyroid having twice the percentage of being diagnosed with hypothyroidism compared to those who did not.

Furthermore, the attributes TSH, T3, TT4, and FTI play significant roles in determining the thyroid condition. These factors are indicative of thyroid hormone levels and provide valuable insights into the functioning of the thyroid gland. By considering these attributes, healthcare professionals can gain a deeper understanding of the patient's thyroid status and make more accurate diagnoses and treatment decisions.

These findings, incorporating age, sex, Thyroxine usage, Query Hypothyroid attribute, as well as TSH, T3, TT4, and FTI, have significant implications for improving diagnostic accuracy, facilitating personalized treatment approaches, and enhancing patient outcomes in thyroid disease. By leveraging these factors, healthcare professionals can better understand and predict thyroid conditions, leading to more effective interventions and improved patient care.

References

[1] "UCI Machine Learning Repository," *archive.ics.uci.edu.* https://archive.ics.uci.edu/dataset/102/thyroid+disease

[2] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," *BioMed Research International*, vol. 2022, pp. 1–10, Jun. 2022, doi: https://doi.org/10.1155/2022/9809932.

[3] "Thyroid Disease Unsupervised Anomaly Detection," *www.kaggle.com.* https://www.kaggle.com/datasets/zhonglifr/thyroid-disease-unsupervised-anomaly-detection

[4] "Thyroid Disease dataset – ODDS." http://odds.cs.stonybrook.edu/thyroid-disease-dataset/ (accessed Jun. 15, 2023).