1- Describa de manera clara el problema de predicción que está abordando, su campo de aplicación y explique si corresponde a un problema de clasificación o de regresión.

La tiroides es una pequeña glándula con forma de mariposa ubicada en la base del cuello y es parte de una red de glándulas llamadas sistema endocrino. Esta glándula produce hormonas que regulan el metabolismo del cuerpo, cuando una persona tiene problemas con esta glándula puede presentar los siguientes síntomas: (inquietud, nerviosismo, corazón acelerado, irritabilidad, aumento de la sudoración, sacudida, ansiedad, problemas para dormir, piel delgada, cabello y uñas quebradizas, debilidad muscular, pérdida de peso, fatiga, piel seca, aumento de la sensibilidad al frío, problemas de memoria, estreñimiento, depresión, aumento de peso, debilidad, frecuencia cardíaca lenta, coma).

El campo en que se presenta es vida, referente al área de la salud y el problema que se quiere abordar es determinar si un paciente referido a la clínica padece hipotiroidismo

Este problema corresponde a un problema de machine learning tipo clasificación, ya que sus valores son finitos y podemos categorizar las salidas.

La base de datos que se utilizará para este experimento es la siguiente: <https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease> y se eligió debido a que la base de datos está actualizada, es la mejor para tomar decisiones según el modelo de clasificación aplicar y las otras para completar los datos usaban….

2- Enumere las variables incluidas como entrada al sistema y la o las variables a predecir. Explique claramente el tipo de codificación de cada variable y, si la base de datos cuenta con valores faltantes, explique cómo se llenaron los vacíos en cada caso.

**Resumen de conjunto de datos:**

Number of classes: 3

Number of learning examples: 3772

Number of testing examples: 3428

Variables de entrada del sistema:

Número de características 21.

15 Binarias.

6 continuas.

Variables a predecir:

El problema es determinar si un paciente referido a la clínica padece de hipotiroidismo Por lo tanto, se construyen tres clases: normal (no hipotiroideo), hipotiroideo y funcionamiento subnormal. Porque el 92 por ciento de los pacientes No son hipotiroideos, un buen clasificador debe ser significativamente mejor que el 92%.

<https://www.cs.sfu.ca/~wangk/ucidata/dataset/thyroid-disease/ann-thyroid.names>

<https://github.com/EpistasisLab/penn-ml-benchmarks/tree/master/datasets/classification/ann-thyroid>

Nombre de los atributos: <http://www.is.umk.pl/~duch/projects/projects/datasets.html>

Attribute Name Possible Values

-------------- ---------------

age: continuous.

sex: 0, 1

on thyroxine: 0, 1

query on thyroxine: 0, 1

on antithyroid medication: 0, 1

sick: 0, 1

pregnant: 0, 1

thyroid surgery: 0, 1

I131 treatment: 0, 1

query hypothyroid: 0, 1

query hyperthyroid: 0, 1

lithium: 0, 1

goitre: 0, 1

tumor: 0, 1

hypopituitary: 0, 1

psych: 0, 1

TSH: continuous.

T3: continuous.

TT4: continuous.

T4U: continuous.

FTI: continuous.

Salidas nombre y valor:

normal (no hipotiroideo) 3

hipotiroideo 2

funcionamiento subnormal. 1

Se clasifican las variables de entrada en continuas y lógicas.

La base de datos no cuenta con valores faltantes.

3- Realice una búsqueda de al menos 4 artıculos que hayan abordado el mismo problema de aprendizaje que Uds están trabajando. Incluya, en la medida de lo posible, trabajos que hayan empleado la misma base de datos.

Describa brevemente:

• ¿Qué técnica(s) de aprendizaje usan en los artıculos?

• ¿Que metodologıa de validación usaron?

• ¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?

Se recomienda buscar trabajos en las bases de datos:

www.sciencedirect.com y www.ieeexplore.org. También se pueden buscar trabajos en la base de datos http://link.springer.com, pero se debe tener en cuenta que el acceso que tiene la Universidad es mucho más limitado para dicha base de datos. Incluir preferiblemente artıculos publicados en revista no en congresos o conferencias. No utilice más de una página del informe para esta descripción.

* <http://www.is.umk.pl/~duch/projects/projects/datasets.html>
* <https://www.ispacs.com/journals/ojids/2016/ojids-00002/article.pdf>
* <https://www.researchgate.net/publication/321145710_Prediction_of_Thyroid_Disease_Using_Data_Mining_Techniques> + <https://pdfs.semanticscholar.org/3ab1/6e04552beeabd98db2e5fa2eb55919393255.pdf>

• ¿Qué técnica(s) de aprendizaje usan en los artıculos?

Naive Bayes Learner

Decision Tree Learner

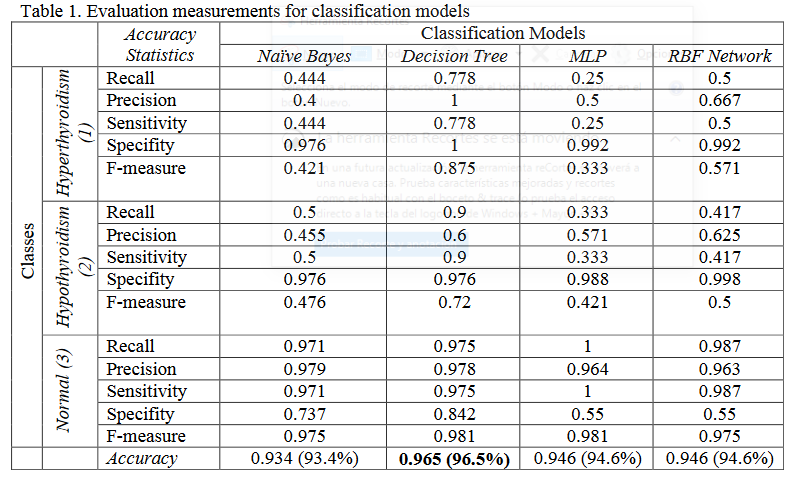
• ¿Que metodologıa de validación usaron?

Se utiliza una matriz de confusión para visualizar el desempeño que tiene el algoritmo. Los datos están normalizados para poderlos aplicar a los modelos de red neuronal MLP y RFB Network

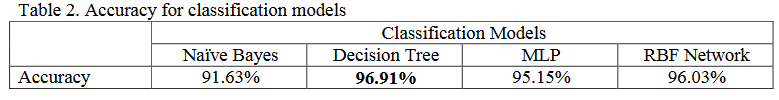
the nodes used to validate the model, Scorer

the node reports a confusion matrix and the accompanying quality measures in its view, Normalizer – the data set are normalized to be able to apply the neural network models, Multilayer Perceptron and RBFNetwork – the nodes corresponding to the neural network classification models, Weka Predictor – a node implemented in Weka to validate the models.

• ¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?



The accuracy of the classification models is over 90%, the best model being Decision Tree with 96.5% accuracy. If the number of epochs for MLP model increases to 1500 (initial was set to 500) the accuracy of the classification model increases to 95.34%. The authors initially considered the Decision Tree model Gini Index as quality measure and not pruning method. For the same data set, but choosing MDL as a post pruning method, the accuracy of the decision tree model increases to 96.9%. Other experiment consists in analyzing the modification occurred after changing the partition percent to 70%



After a careful analysis and data processing, a data set with 140 records resulted and consists in row material for the second experiment, discussed in the current paper. The attributes used to apply data mining algorithms are: TSH, FT4, ATPO, AGE, SEX and Class. The target is class with three possible values: 1- hyperthyroidism, 2- hypothyroidism and 3 – normal.

Conclusions

As the medical reports show serious thyroid dysfunctions among the population, more affected being women, thyroid classification is a very important subject for researchers in medical science. In literature are mentioned various research works in the field of thyroid classification based on different data mining techniques used to build robust classifier. In this paper the authors discussed about applying four classification models (Naïve Bayes, Decision Tree, MLP and RBF Network) on thyroid data set to identify more accurately the dysfunction of thyroid namely hyperthyroidism and hypothyroidism. The best classification model was the decision tree model in all the effectuated experiments. The future work will focus on the identification of factors that affect the thyroid diseases and on testing more data mining techniques for the classification of different diseases (diabetes, heart diseases etc.).

* <https://ieeexplore.ieee.org/document/8745910>
* <https://ieeexplore.ieee.org/document/8094070>
* <https://ieeexplore.ieee.org/document/7976548>

¿Que tecnica(s) de aprendizaje usan en los artıculos?

¿Que metodologıa de validacion usaron?

¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?

* <http://papersim.com/wp-content/uploads/Neural_Network_Thyroid_Disease_2015.pdf>
* <https://link.springer.com/article/10.1007/s10916-012-9825-3>
* <https://computerresearch.org/index.php/computer/article/view/1353> + <https://pdfs.semanticscholar.org/4110/afbe1a0511d4bc53f976ed02b53c9f6e385c.pdf>

¿Que tecnica(s) de aprendizaje usan en los artıculos?

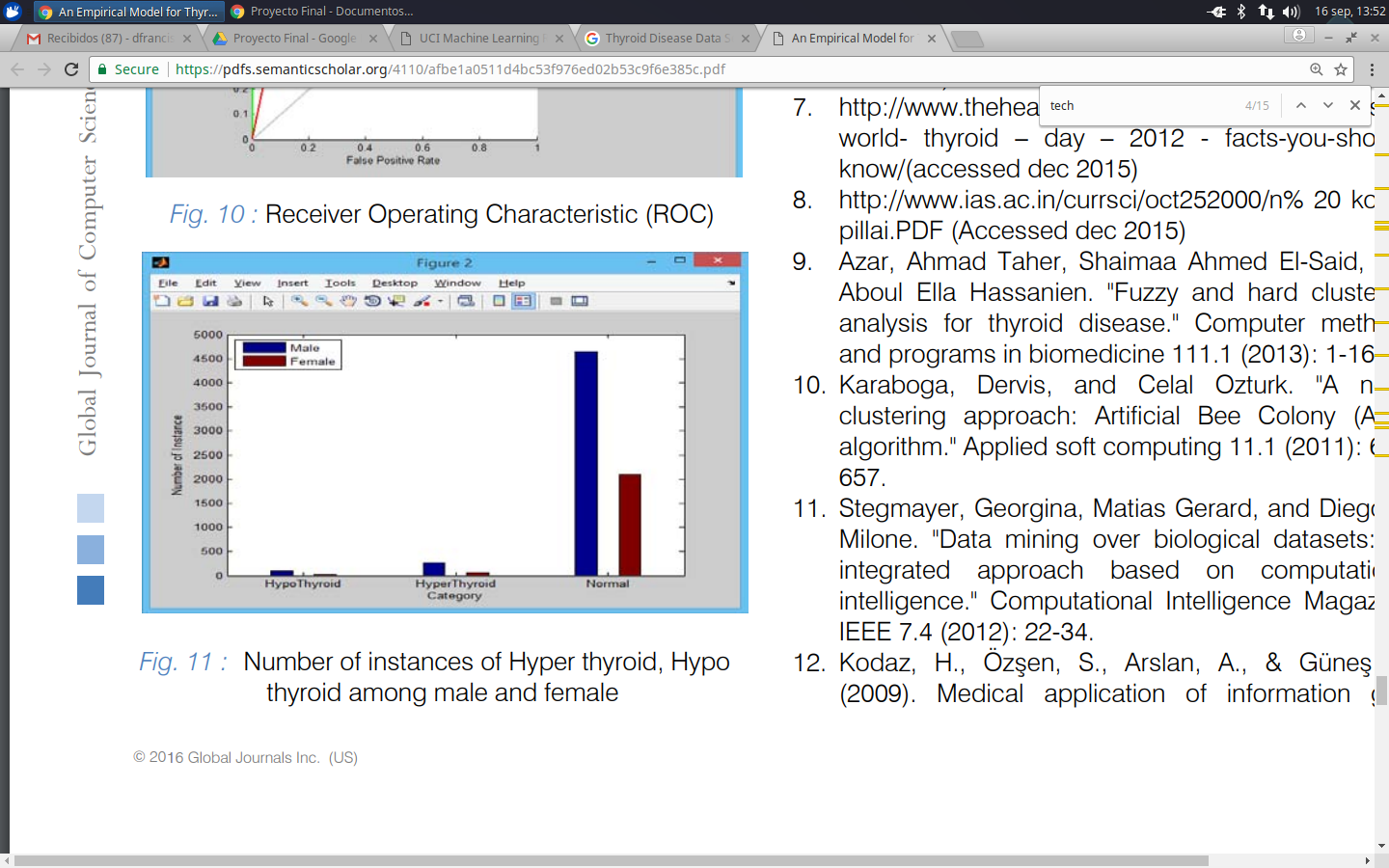
Naive Bayesian Classifier

¿Que metodologıa de validacion usaron?

kfold cross-validation

¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?

The accuracy of classification is achieved as 97.97%



The objective of this research work is aimed to show the classes of thyroid from the available raw medical dataset helps the physician to arrive at an accurate diagnosis. The results show that the proposed Evolutionary Multivariate Bayisean Prediction classifier model achieves remarkable dimensionality reduction from among the 7200 medical datasets obtained from the UCI repository with 21 attributes (Continuous -15; Discrete - 6). 21 epohs (runs) are carried out for the data and after stabilization, the data are classified as Hyper, Hypo and Normal classes. The results are evaluated based on ten evaluation metrics and the accuracy of classification is 97.97%.

* <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S1-S22>

¿Que tecnica(s) de aprendizaje usan en los artıculos?

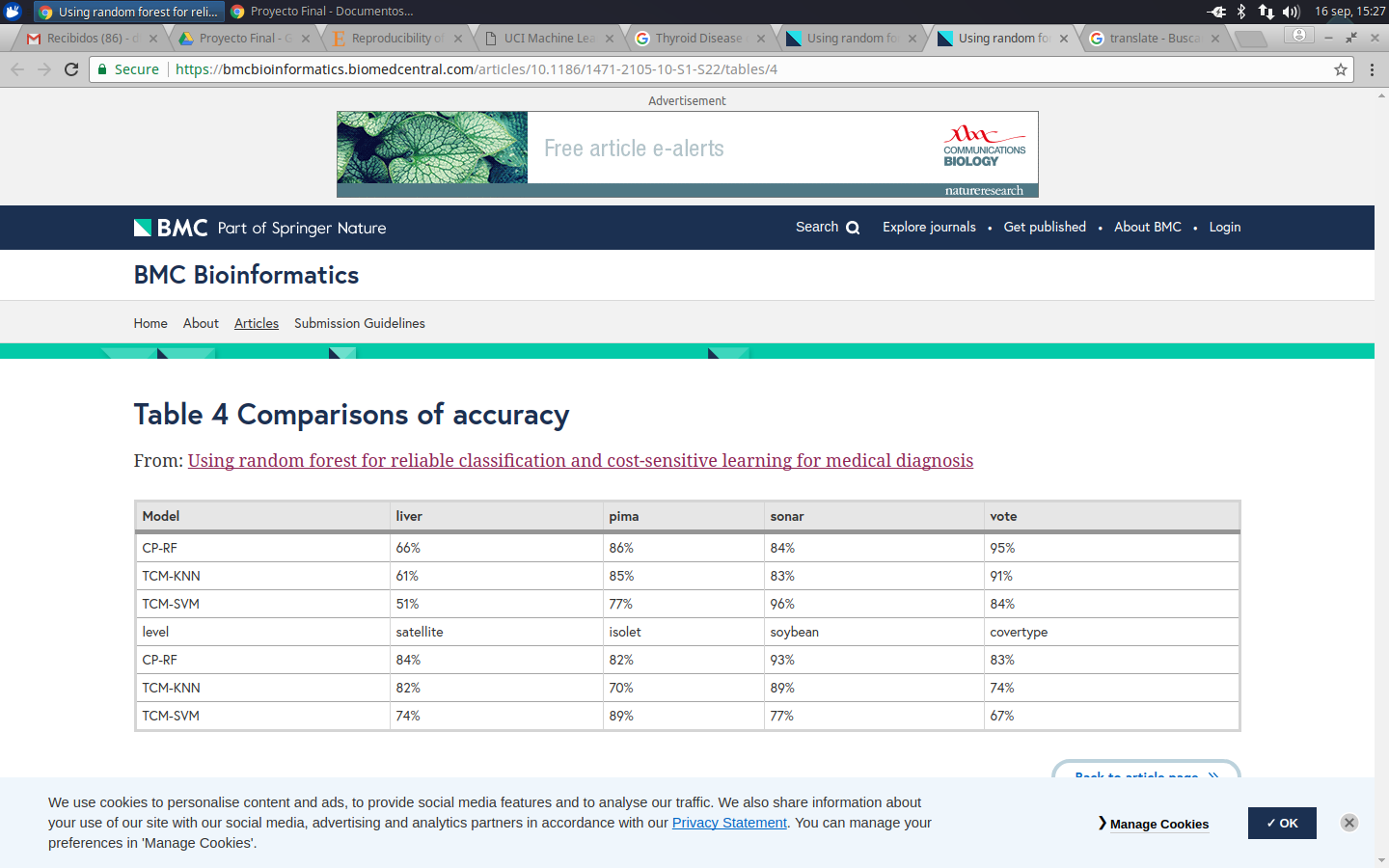
Random forest

¿Que metodologıa de validacion usaron?

10-fold cross validation

¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?

It is clear that CP-RF performs well at most of the datasets, especially on the datasets with categorical and mixed variable. CP-RF especially outperforms TCM-KNN for high-dimension dataset (isolet), and outperforms TCM-SVM for noisy data (covertype).



Most of state-of-the-art machine learning algorithms cannot provide a reliable measure of their classifications and predictions. This paper addresses the importance of reliability and confidence for classification, and presents a novel method based on a combination of random forest, and conformal predictor. The new algorithm hedges the predictions of RF and gives a well-calibrated region prediction by using the proximity matrix generated with RF as a nonconformity measure of examples.

* <https://www.researchgate.net/publication/303480840_MLTDD_Use_of_Machine_Learning_Techniques_for_Diagnosis_of_Thyroid_Gland_Disorder>

¿Que tecnica(s) de aprendizaje usan en los artıculos?

The classification algorithms such as decision trees and support vector machines (SVM)

¿Que metodologıa de validacion usaron?

Parece que no tiene

¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?

* <https://www.ijarcs.info/index.php/Ijarcs/article/download/4929/4289>

¿Que tecnica(s) de aprendizaje usan en los artıculos?

BPNN (Back Propagation Neural Network), KNN (K Nearest Neighbor) and SVM (Support Vector Machine) classifier

¿Que metodologıa de validacion usaron?

¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?

* <https://arxiv.org/pdf/1403.0522.pdf>
* <https://file.scirp.org/pdf/ENG_2013103114384127.pdf>

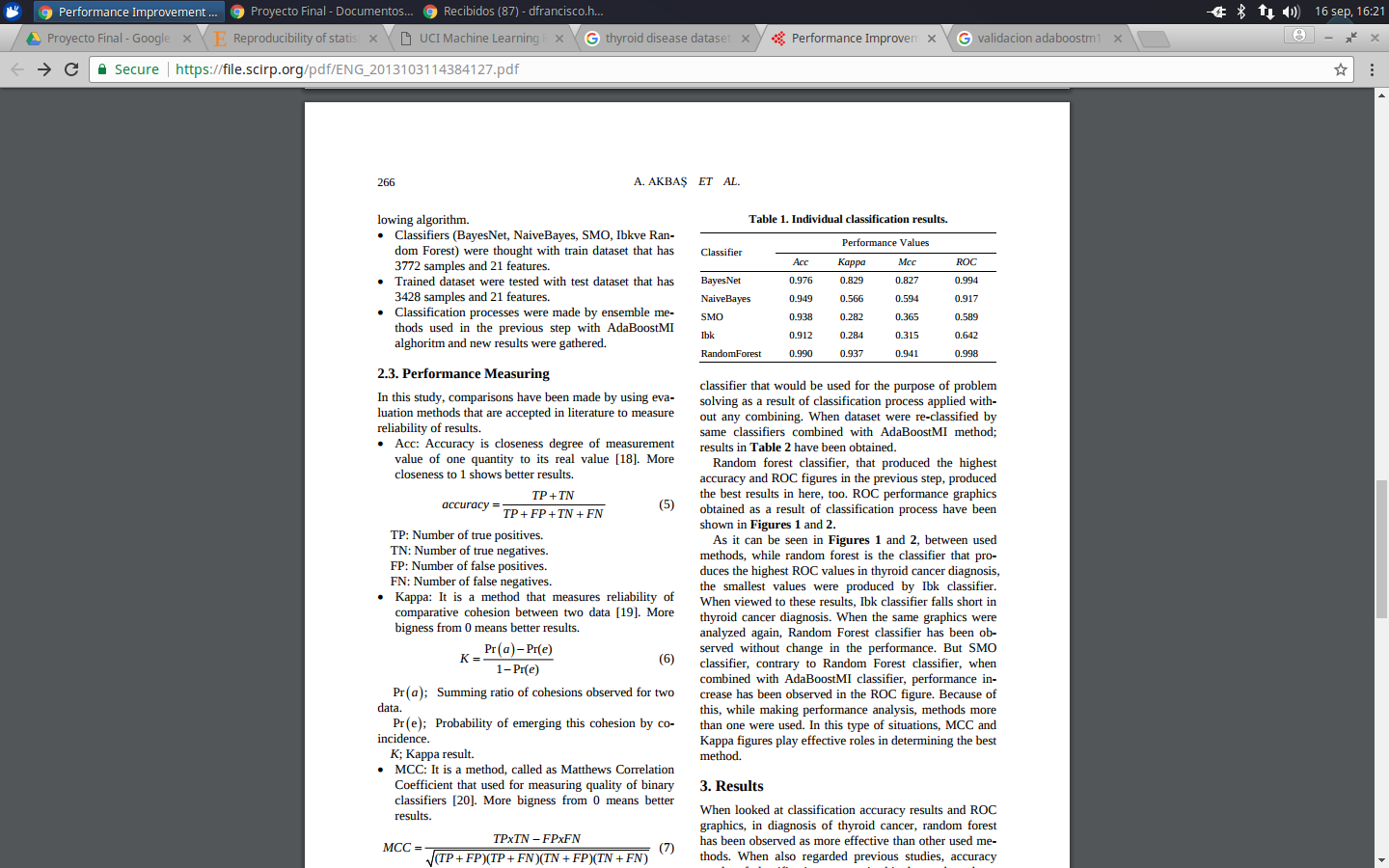
¿Que tecnica(s) de aprendizaje usan en los artıculos?

BayesNet, Naïve Bayes, Sequential Minimal Optimization (SMO), IBK, Random Forest, AdaBoostM1.

¿Que metodologıa de validacion usaron?

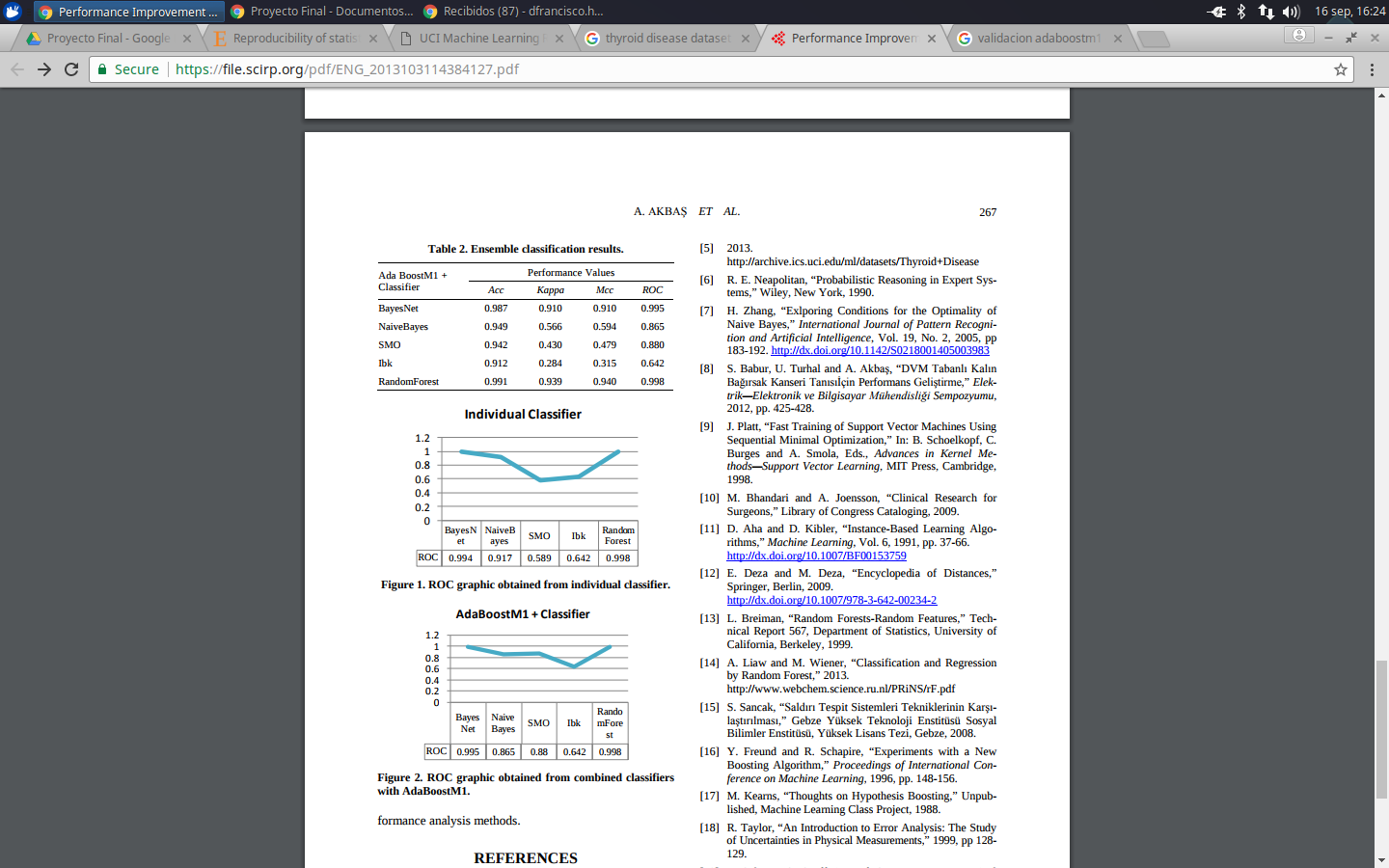
cross-validation

¿Cuales fueron los resultados obtenidos en cada uno de los trabajos citados?



When looked at classification accuracy results and ROC graphics, in diagnosis of thyroid cancer, random forest has been observed as more effective than other used methods.

When also regarded previous studies, accuracy results of classification process in this dataset have been observed with their coming fairly close to %100. Making predictions with such high accuracy values makes study in this area hard in the subject of thyroid cancer. When taken into account obtained results and used datasets being old (1992), a need can be seen for a new dataset to obtain more accurate and more valid results.



4- Incluya una sesion dentro de su informe con el nombre Experimentos, en la cual describa la metodologıa de validación usada y la base de datos que está usando para llevar a cabo el proyecto, incluyendo la fuente de la base de datos como referencia, el número de muestras, variables, etc. Si su problema es de clasificación, indique cuántas clases contiene la base de datos y cuantas muestras por clase. Si la base de datos está desbalanceada, deben considerar el uso de técnicas de submuestreo y sobremuestreo inteligente, además de usar validación estratificada.

Para este problema se aborda la base de datos “ann-train.data” del respositorio “thyroid disease data set” la cual tiene 21 atributos, 3772 muestras y tiene 3 diferentes clases las cuales la clase 1 tiene 2,47% de las muestras, la clase 2 tiene 5,06% de las muestras y la clase 3 tiene 92,47% de las muestras.

Como se puede observar se tiene un problema de desbalanceo y para ello vamos a utilizar SMOTEENN la cual es una librería de combinación de sobre muestreo y submuestreo inteligente para balancear los datos, creando muestras de la clase minoritaria para obtener una muestra promedio y a continuación utilizaremos la validación estratificada “Naive Bayes” utilizando diferentes porcentajes de sobremuestreo para lograr un mejor resultado.

Como se menciona en los artículos relacionados, el autor elige cuantos folds utiliza para la validación, para este vamos a utilizar 5-folds.

**Naive Bayes:**

|  |  |  |
| --- | --- | --- |
| Porcentaje de Balanceo | Error de Clasificación | Intervalo de Confianza |
| 0% | 0.2293085494665928 | 0.03077768184548598 |
| 5% | 0.18107616906144647 | 0.0214205093507078 |
| 10% | 0.1802797067101439 | 0.01583641840034585 |
| 15% | 0.18132703147953405 | 0.02910920455595323 |
| 20% | 0.17868468520535746 | 0.02253861285390327 |
| 25% | 0.18611351712879382 | 0.021841190335585424 |

Como se observa en la tabla solo hasta el 20% se obtiene un intervalo de confianza y un error de clasificación aceptables para el desbalanceo de datos sobre la clase minoritaria.

**KNN:**

|  |  |  |
| --- | --- | --- |
| **K-Vecinos** | **Error de Clasificación** | **Intervalo de Confianza** |
| **1** | 0.06495139313798451 | 0.0021859179513763066 |
| **2** | 0.06336233677016243 | 0.0015713932872648002 |
| **3** | 0.06415704154893095 | 0.00230878419526919 |
| **4** | 0.06362653663701129 | 0.0025051004153281 |
| **5** | 0.06495139313798451 | 0.0024862993599678415 |
| **6** | 0.06495244712310225 | 0.002218473473525061 |
| **7** | 0.06468789685855199 | 0.001968414692155269 |
| **8** | 0.06389284447409563 | 0.0030858383661151747 |
| **9** | 0.06548295482703545 | 0.003419225956252193 |
| **10** | 0.06548154765220278 | 0.0031790183116519374 |
| **11** | 0.06601240575383738 | 0.003866622372120999 |
| **12** | 0.06601275615153876 | 0.003064305912252658 |
| **13** | 0.06574715190187078 | 0.003076667288654344 |
| **14** | 0.06548154765220278 | 0.0029487700956473987 |
| **15** | 0.06627730641608903 | 0.0034402849297134695 |
| **16** | 0.06601240575383738 | 0.0035840755771531442 |
| **17** | 0.06601240575383738 | 0.0035840755771531442 |
| **18** | 0.06601240575383738 | 0.003866622372120999 |
| **19** | 0.06601240575383738 | 0.003866622372120999 |

Como se observa en la tabla anterior después de correr K vecinos mayores que 11 empiezan arrojar resultados iguales en sus vecinos, con solo 10 ya se obtienen los valores más exactos de los vecinos a evaluar.

**Redes Neuronales Artificiales:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Epocas** | **Capas Ocultas** | **Error Clasificación** | **Intervalo de confianza** |
| 10 | **10** | 0.022181086090481618 | 0.0008595557645385432 |
| **10** | **20** | 0.02385878935407092 | 0.0054453296321465275 |
| **10** | **30** | 0.02755875548229312 | 0.013384044016925624 |
| **10** | **40** | 0.0272060217962261 | 0.012699689840352817 |
| **10** | **50** | 0.02509675560113983 | 0.0002915442473657968 |
| **100** | **10** | 0.02509675560113983 | 0.0002915442473657968 |
| **100** | **20** | 0.020854589746879308 | 0.001277903426419425 |
| **100** | **30** | 0.031094392533932386 | 0.009917802060862492 |
| **100** | **40** | 0.02253464016319878 | 0.0010100516358206506 |
| **100** | **50** | 0.022622706785481735 | 0.0015866662391695295 |
| **1000** | **10** | 0.023152977168262607 | 0.0010267350236736876 |
| **1000** | **20** | 0.021562633216855177 | 0.0012341839056585065 |
| **1000** | **30** | 0.023330281197180048 | 0.0013589674147973035 |
| **1000** | **40** | 0.26523864573010153 | 0.11983287422617692 |
| **1000** | **50** | 0.31450977104697764 | 0.0013364512658075422 |

A continuación se observan las redes neuronales que se corrieron hasta mil épocas variando las capas ocultas hasta 50, no se corren más capas ni épocas mayores ya que los otros resultados no arrojan resultados eficientes.

**Random Forest**

|  |  |  |
| --- | --- | --- |
| **Arboles** | **Error Clasificación** | **Intervalo de confianza** |
| 10 | 0.051430179628380215 | 0.005513568623128987 |
| 20 | 0.05461708581072187 | 0.003935999698639382 |
| 30 | 0.04904642685783025 | 0.005105214227777381 |
| 40 | 0.053557480369701684 | 0.00414069982469638 |
| 50 | 0.05010884664851578 | 0.003331951975840308 |
| 60 | 0.05037128615081703 | 0.002779226511817212 |
| 70 | 0.04825313483792152 | 0.005164826465345447 |
| 80 | 0.05143229318264277 | 0.0010229341290607123 |
| 90 | 0.05169930739915701 | 0.0039683050154970135 |
| 100 | 0.05249400938591198 | 0.0018532237894483814 |

La cantidad de árboles se toman en cuenta para poder variar los resultados, pero como se observa en la tabla no varían mucho con diferentes tamaños de árbol.

**MÁQUINAS DE SOPORTE VECTORIAL CON KERNEL LINEAL**

**Hacer la tabla**

|  |  |  |  |
| --- | --- | --- | --- |
| **C** | **Error Clasificación** | **Intervalo de confianza** | **% Vectores de Soporte** |
| **0.001** | **0.929038** | **0.0110914** | **0.142923** |
| **0.01** | **0.935657** | **0.009526** | **0.132536** |
| **0.1** | **0.94742** | **0.0116786** | **0.121417** |
| **1** | **0.9958448** | **0.011023** | **0.107906** |
| **10** | **0.960654** | **0.0111415** | **0.100278** |
| **100** | **0.95992** | **0.00977058** | **0.100094** |

Como se observa en la figura se obtiene un buen resultado con un “C” mayor que muestra cómo se clasifican, enseñando mejor resultado en una buena eficiencia.

**MÁQUINAS DE SOPORTE VECTORIAL CON KERNEL RBF**

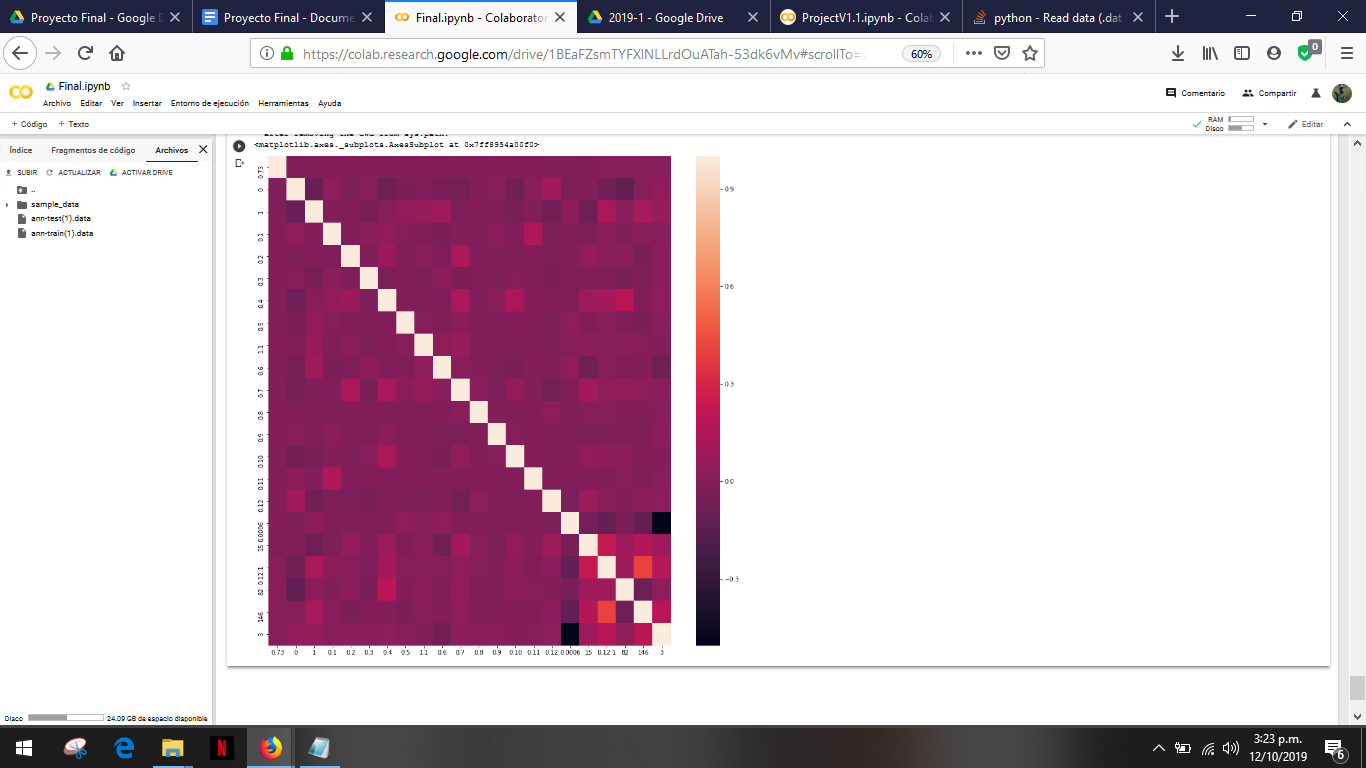
**Llenar tabla**

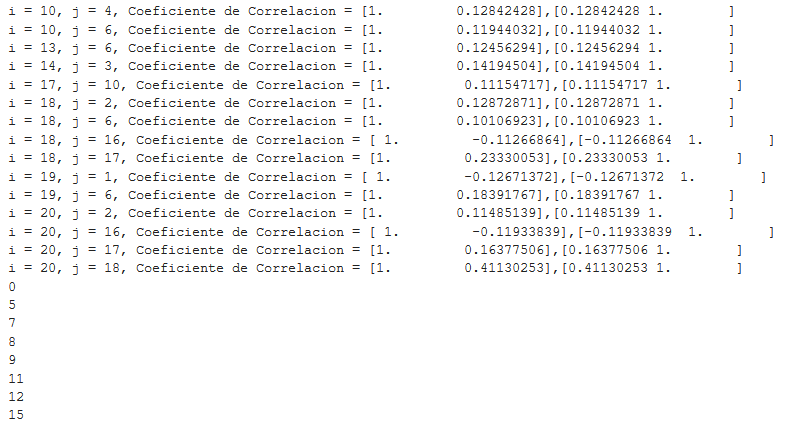
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **C** | **GAMMA** | **Error Clasificación** | **Intervalo de confianza** | **% Vectores de Soporte** |
| **0.001** | **0.01** | **0.926833** | **0.0136474** | **0.141728** |
| **0.001** | **0.1** | **0.926833** | **0.0136474** | **0.144303** |
| **0.001** | **1** | **0.926833** | **0.0136474** | **0.153493** |
| **0.01** | **0.01** | **0.926833** | **0.0136474** | **0.150093** |
| **0.01** | **0.1** | **0.926833** | **0.0136474** | **0.211579** |
| **0.01** | **1** | **0.926833** | **0.0136474** | **0.35872** |
| **0.1** | **0.01** | **0.92904** | **0.0116774** | **0.161121** |
| **0.1** | **0.1** | **0.927567** | **0.0133098** | **0.241359** |
| **0.1** | **1** | **0.926833** | **0.0136474** | **0.467446** |
| **1** | **0.01** | **0.938599** | **0.0101901** | **0.149815** |
| **1** | **0.1** | **0.947412** | **0.0124418** | **0.237865** |
| **1** | **1** | **0.934552** | **0.0115126** | **0.478658** |
| **10** | **0.01** | **0.953665** | **0.00973111** | **0.135018** |
| **10** | **0.1** | **0.949622** | **0.00924815** | **0.214703** |
| **10** | **1** | **0.934549** | **0.0146685** | **0.467536** |
| **100** | **0.01** | **0.958078** | **0.0077083** | **0.126103** |
| **100** | **0.1** | **0.943734** | **0.0112016** | **0.164429** |
| **100** | **1** | **0.926833** | **0.0137868** | **0.457611** |

La tabla anterior muestra como el gamma influencia en la precisión de los datos, menor el gamma mayor influencia tienen los vectores de soporte, como se observa en el gamma 0.01 con C 100.

**Análisis Individual**

**Correlación matriz**

****

****

Se observa en la matriz de correlación las características candidatas a ser eliminadas son las características a eliminar son la 0 age, 5 sick, 7 thyroid\_surgery, 8 I131\_treatment, 9 query\_hypothyroid, 11 lithium, 12 goitre, 15 psych

**Indice de Fisher**

**Este no he podido**

**Selección de Características**

Para la selección de características se usa Wrapper elige la SFS (Sequential Forward Selection) la cual da como resultado que la precisión mejora con 14 caracteristicas las cuales son:

(0, 2, 3, 4, 5, 6, 8, 9, 10, 16, 17, 18, 19, 20)

Ahora con la selección de estas características se procederá a evaluar nuevamente los 3 mejores modelos anteriores que son (KNN, RNA, RF):

Ahora analizamos los cambios con las características seleccionadas:

**KNN**

|  |  |  |
| --- | --- | --- |
| **K-Vecinos** | **Error de Clasificación** | **Intervalo de Confianza** |
| **1** | 0.06893651086812216 | 0.00973422215667973 |
| **2** | 0.07264828767902566 | 0.007554271920269801 |
| **3** | 0.0697308652491893 | 0.008257849884326455 |
| **4** | 0.07026382294501868 | 0.0072051523455491295 |
| **5** | 0.06548892834004044 | 0.006597693089658821 |
| **6** | 0.06840600316418896 | 0.005032849535723948 |
| **7** | 0.07079257586843138 | 0.007467319614179129 |
| **8** | 0.06840108363630171 | 0.005695105191175573 |
| **9** | 0.0697301588697594 | 0.005086131732871021 |
| **10** | 0.06839686769583078 | 0.004680305950073526 |

**RNA**

|  |  |  |  |
| --- | --- | --- | --- |
| **Epocas** | **Capas Ocultas** | **Error Clasificación** | **Intervalo de confianza** |
| **10** | **10** | 0.08505260549076457 | 0.11994645281886503 |
| **10** | **20** | 0.08505260549076457 | 0.11994645281886503 |
| **10** | **30** | 0.08505260549076457 | 0.11994645281886503 |
| **10** | **40** | 0.08505260549076457 | 0.11994645281886503 |
| **10** | **50** | 0.08716900760716668 | 0.11896765721543111 |
| **100** | **10** | 0.08505260549076457 | 0.11994645281886503 |
| **100** | **20** | 0.08505260549076457 | 0.11994645281886503 |
| **100** | **30** | 0.08187625032765443 | 0.12158090281273713 |
| **100** | **40** | 0.08505260549076457 | 0.11994645281886503 |
| **100** | **50** | 0.08496407074087522 | 0.11999105639446561 |
| **1000** | **10** | 0.08514114024065389 | 0.1199020941443981 |
| **1000** | **20** | 0.08505260549076457 | 0.11994645281886503 |
| **1000** | **30** | 0.08434432749164991 | 0.1203101155868251 |
| **1000** | **40** | 0.08505260549076457 | 0.11994645281886503 |
| **1000** | **50** | 0.08443286224153924 | 0.12026380540784115 |

**RF**

|  |  |  |
| --- | --- | --- |
| **Arboles** | **Error Clasificación** | **Intervalo de confianza** |
| 10 | 0.05222664477169635 | 0.0027267117836939493 |
| 20 | 0.05196771483043613 | 0.004900519271601527 |
| 30 | 0.05435183195875526 | 0.004409409238109414 |
| 40 | 0.049580096517116645 | 0.005043565522232653 |
| 50 | 0.048254533636713524 | 0.0031565094750441365 |
| 60 | 0.04931238150519964 | 0.002862330299469093 |
| 70 | 0.04904466928529618 | 0.0028864883135933514 |
| 80 | 0.05063583362335374 | 0.0012794830046381847 |
| 90 | 0.04984218562171651 | 0.0024814035972441956 |
| 100 | 0.05222488999117584 | 0.003488896817481885 |

Al revisar los tiempos de ejecución solo se noto gran mejora en el **RNA**, por tanto no se muestran los tiempos de ejecución ya que no dan datos diferentes a los ya asociados aquí mencionados.

Se puede concluir que el mejor modelo con las características seleccionadas es el **RF**.

**Extracción de Caracteristicas**

A continuación procedemos con la extracción de características mediante Análisis de componentes principales (PCA), principalmente usamos el entrenamiento:

|  |  |  |  |
| --- | --- | --- | --- |
| **# Características** | **Error Clasificación** | **Intervalo de confianza** | **Eficiencia** |
| **2** | **0.0755631** | **0.0143574** | **92.47018110913977** |
| **3** | **0.0755631** | **0.0143574** | **92.47018110913977** |
| **4** | **0.0755631** | **0.0143574** | **92.47018110913977** |
| **5** | **0.0654881** | **0.011329** | **92.47018110913977** |
| **6** | **0.0654881** | **0.011329** | **92.47018110913977** |
| **7** | **0.0654881** | **0.011329** | **92.47018110913977** |
| **8** | **0.0654884** | **0.0116363** | **92.47018110913977** |
| **9** | **0.0657537** | **0.0115763** | **92.47018110913977** |
| **10** | **0.0657537** | **0.0116377** | **92.47018110913977** |
| **11** | **0.063898** | **0.0121049** | **92.47018110913977** |
| **12** | **0.0636327** | **0.0115882** | **92.47018110913977** |
| **13** | **0.0633675** | **0.0116903** | **92.60280710383473** |
| **14** | **0.062837** | **0.011172** | **92.6292971700599** |
| **15** | **0.062837** | **0.011172** | **92.97394909269767** |
| **16** | **0.062837** | **0.011172** | **93.00043915892283** |
| **17** | **0.0631022** | **0.0114218** | **93.026929225148** |
| **18** | **0.0628373** | **0.0115756** | **93.07997962302598** |
| **19** | **0.0625724** | **0.0117514** | **93.61016740738138** |
| **20** | **0.0625724** | **0.0117514** | **93.68974300419836** |

Como conclusión de la extracción, la eficiencia no varía mucho en comparación comparación de las otras, posiblemente si se balancean los datos igualando todo con un sobremuestreo puede que se obtenga una mejor precisión, pero de igual forma como se observa en la tabla no se disminuye bastante la extracción de características, es decir no variaría mucho en el resultado esperado de los modelos a evaluar.

**Discusión**

<https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease>

La tiroides es una glándula pequeña con forma de mariposa ubicada en la base del cuello, justo debajo de la manzana de Adán. Es parte de una intrincada red de glándulas llamada sistema endocrino. El sistema endocrino es responsable de coordinar muchas de las actividades de su cuerpo. La glándula tiroides produce hormonas que regulan el metabolismo de su cuerpo.

Pueden surgir varios trastornos diferentes cuando la tiroides produce demasiada hormona (hipertiroidismo) o no suficiente (hipotiroidismo).

**Hipertiroidismo**

En el hipertiroidismo , la glándula tiroides es hiperactiva. Produce demasiada hormona. El hipertiroidismo afecta aproximadamente al 1 por ciento de las mujeres . Es menos común en hombres.

La enfermedad de Graves es la causa más común de hipertiroidismo y afecta a aproximadamente el 70 por ciento de las personas con tiroides hiperactiva. Los nódulos en la tiroides, una afección llamada bocio nodular tóxico o bocio multinodular , también pueden causar que la glándula produzca en exceso sus hormonas.

La producción excesiva de hormona tiroidea conduce a síntomas como:

inquietud

nerviosismo

corazón acelerado

irritabilidad

aumento de la sudoración

sacudida

ansiedad

problemas para dormir

piel delgada

cabello y uñas quebradizas

debilidad muscular

pérdida de peso

ojos saltones (en la enfermedad de Graves)

**Hipotiroidismo**

El hipotiroidismo es lo opuesto al hipertiroidismo. La glándula tiroides es poco activa y no puede producir suficientes hormonas.

El hipotiroidismo a menudo es causado por la enfermedad de Hashimoto , la cirugía para extirpar la glándula tiroides o el daño causado por el tratamiento con radiación . En los Estados Unidos, afecta a alrededor del 4,6 por ciento de las personas mayores de 12 años. La mayoría de los casos de hipotiroidismo son leves.

Muy poca producción de hormona tiroidea conduce a síntomas como:

fatiga

piel seca

aumento de la sensibilidad al frío

problemas de memoria

estreñimiento

depresión

aumento de peso

debilidad

frecuencia cardíaca lenta

coma

<https://www.healthline.com/health/common-thyroid-disorders>