CLASIFICADOR MEDIANTE IMÁGENES DE CÁNCER DE PIEL

Facultad de Ingeniería Informática

Departamento de Ingeniería Telemática

Prashant Jeswani Tejwani

Bajo la dirección del doctor
José María Quinteiro González

Universidad de Las Palmas de Gran Canaria

Memoria para optar al

Grado de Ingeniería Informática

Las Palmas de Gran Canaria, 2021

Abstract

Las decisiones médicas son difíciles ya que menudo deben tomarse con información insuficiente e incierta. Además, el resultado del proceso de decisión tiene implicaciones de gran alcance en el bienestar humano o incluso vidas.

El desempeño humano en la toma de decisiones disminuye con la complejidad de los problemas y la presión del tiempo. Por lo tanto, el apoyo de médicos a la toma de decisiones es crucial, especialmente en su fase inicial cuando un especialista debe elaborar un diagnóstico preliminar y especificar las posibles direcciones para el tratamiento del paciente.

Las herramientas informáticas tienen el potencial de marcar la diferencia en la medicina. Especialmente las redes neuronales profundas, que pueden aprovechar tanto el gran número de datos disponibles como la experiencia clínica.

La aplicación de redes profundas al diagnóstico se ha propuesto hace casi dos décadas, teniendo potencial para beneficiar significativamente la atención médica. Sin embargo, la difusión práctica de este enfoque sigue siendo mínima.

Por lo tanto, en este trabajo se introduce un clasificador de diagnóstico de cáncer de piel. Se realiza comparaciones con varios modelos, mostrando formas de resolver el desbalanceo de datos en datos clínicos. Se expone la conveniencia de calibrar los modelos de clasificación para que las probabilidades mostradas en el diagnóstico se ajusten a los datos. Además, se implementa una red bayesiana que permita incorporar otros factores que influyen en el cáncer de piel.

Finalmente, se argumenta que un diagnóstico con la ayuda de las redes neuronales puede ser beneficiosa para sus usuarios y más acertada, pudiendo en ocasiones mejorar la precisión diagnóstica de los especialistas en la detección de melanoma.

Índice de contenido

Índice de figuras

Índice de tablas

1

Introducción

El cáncer de piel es una enfermedad de carácter más o menos grave, según el tipo de tumor al que dé lugar. El melanoma es un tipo de cáncer de piel que se origina cuando los melanocitos Figura 1.1 (las células responsables de la pigmentación normal de la piel) comienzan a crecer fuera de control [1]. Este es mucho menos frecuente que otros tipos de cánceres de piel, aunque es el más peligroso y agresivo debido a que evoluciona muy rápidamente, es de fácil diseminación por el organismo y muchas veces mortal si no es detectado en una fase temprana.

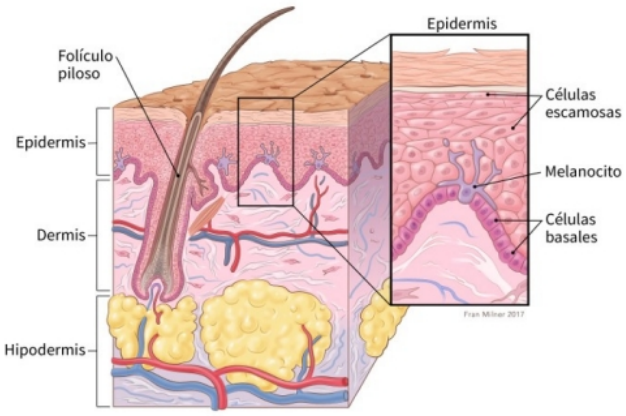

Figura 1.1: Capas de la piel [1].

Muchos tipos de tumores benignos (no cancerosos) se pueden originar de los diferentes tipos de células de la piel. Un lunar es un tumor benigno de la piel que también se origina a partir de los melanocitos. No obstante, casi todos los lunares son inofensivos, aunque algunos tipos pueden aumentar su riesgo de melanoma.

Un tipo de lunar que a veces se parece al melanoma se llama “nevo Spitz”. Este lunar es más común en niños y adolescentes, aunque a veces se presenta en adultos. Por lo general, estos tumores son benignos y no se propagan. Sin embargo, en ocasiones los médicos tienen problemas para distinguir entre un “nevo Spitz” y un melanoma, aun cuando los observan con un microscopio.

En general, las variedades de melanoma se caracterizan por su Asimetría, Borde irregular, Color heterogéneo, Diámetro grande (> 6mm) y Evolución rápida, también conocido como la regla del ABCDE. Se reconocen actualmente como factores de riesgo de melanoma:

Factores personales	Características fenotípicas: piel y ojos claros, pelo rubio o pelirrojo, presencia de efélides y tendencia a quemarse fácilmente y a no pigmentarse o hacerlo con dificultad tras la exposición al sol. Antecedentes familiares de melanoma y de síndrome del nevus displásico familiar. Tener un elevado número de nevi benignos (mayor que 50). Presencia de nevi congénito grande o gigante. Otros: Xeroderma pigmentoso (padecimiento hereditario que afecta a la capacidad de las células de la piel de reparar el daño causado a su ADN), Inmunodepresión (síndromes linfoproliferativos, VIH, inmundeficiencias primarias, transplantados que están con tratamiento inmunosupresor...).
Factores ambientales	Radiación Ultravioleta, que es el factor etiopatogénico más importante en el desarrollo del melanoma cutáneo.

Tabla 1.1: Factores de riesgo de melanoma [8].

Según la etapa del cáncer y otros factores, las opciones de tratamiento podrían incluir: cirugía, inmunoterapia, medicamentos de terapia, quimioterapia, radioterapia para el cáncer de piel tipo melanoma, entre otros.

1.1 Motivación

El cáncer de piel está aumentando en muchos países, debido a que cada vez la capa de ozono que protege a la tierra de la radiación solar se debilita. Uno de los principales responsables del cáncer de piel de melanoma es la radiación ultravioleta proveniente del sol, o de cualquier otra fuente artificial de radiación ultravioleta, como las lámparas bronceadoras [2].

El Índice de Radiación Ultravioleta (Índice UV o UVI) se mide en una escala a partir de 0 sin límite superior, en la que un índice UV menor de 2, representa un riesgo bajo (código de color verde) y un índice UV mayor de 11 representa un riesgo extremadamente alto (código de color violeta).

En Canarias, no solo el sol acompaña todo el año con una exposición constante, también una exposición estacional; es decir, personas que durante todo el año no toman demasiado sol pero que en verano sí lo hacen de forma constante. Al sol de Canarias también se suman los niveles de radiación que existe en la región dada su proximidad al Ecuador. Esto hace que el índice de radiación UV sea extremadamente elevado comom se observa en la Figura 1.2, siendo el más alto de las Comunidades Autónomas de España.

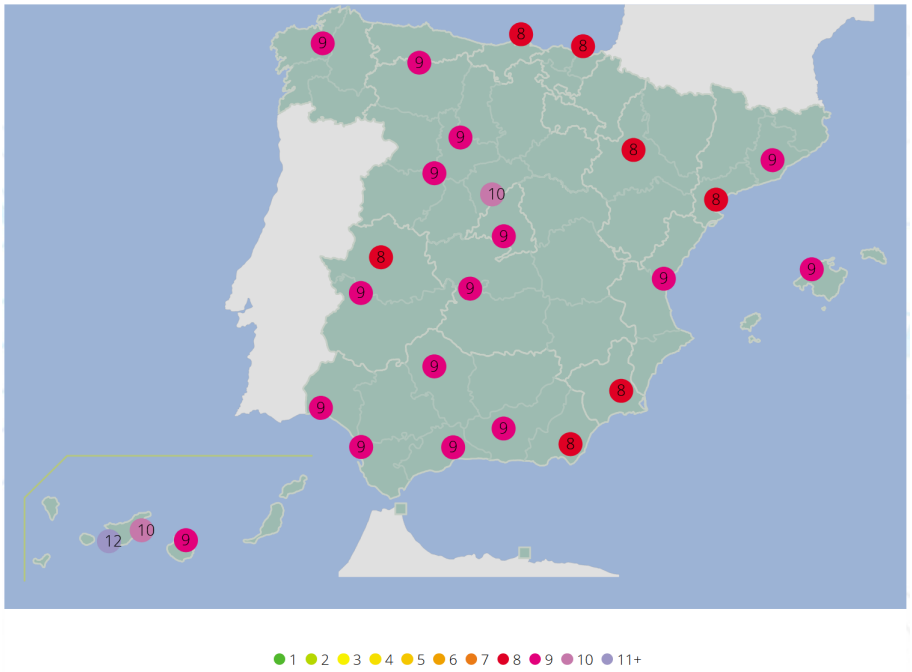

Figura 1.2: Índice UV de las Comunidades Autónomas de España (datos del sábado, 17 julio 2021) [2].

Por lo que, la motivación principal del trabajo ha sido la identificación e incorporación de los factores que influyen en el cáncer de piel para obtener la probabilidad de tener melanoma a partir de dichos factores y del resultado de un clasificador basado en redes neuronales convolucionales.

1.2 Trabajos previos

Este trabajo está fundamentado principalmente sobre las siguientes publicaciones: *On Calibration of Modern Neural Networks* [6] en el cual se descubre que las redes neuronales modernas, a diferencia de los de hace una década, están mal calibradas y explica la importancia de la calibración de los modelos mediante una serie de experimentos. Por otro lado, la publicación *Bayesian Network Classifiers* [7] evalúa los enfoques para inducir clasificadores a partir de datos, basados en la teoría del aprendizaje de redes bayesianas las cuales son representaciones factorizadas de distribuciones de probabilidad que generalizan el clasificador *Naive Bayes*.

Una de las noticias más recientes en esta línea de trabajo es la publicación del 18 de mayo de 2021: *Google AI tool can help patients identify skin conditions* [9] en la cual Google ha presentado una herramienta que utiliza inteligencia artificial para ayudar a detectar las condiciones de la piel, el cabello y las uñas, basada en imágenes cargadas por los pacientes.

1.3 Objetivos

El objetivo principal es apoyar el diagnóstico médico, ayudar a los médicos para tomar mejores decisiones y detectar la enfermedad en una fase tempranera con el propósito de poder realizar el tratamiento adecuado. Para ello se implementa un clasificador capaz de discernir imágenes malignas (particularmente, de tipo melanoma) y benignas a partir de un conjunto de datos. Asimismo, implementar una red bayesiana que permita identificar e incorporar otros factores que influyen en el cáncer de piel.

1.4 Metodología

El desarrollo se ha hecho principalmente en la plataforma de Kaggle utilizando Jupyter Notebook, en el lenguaje de programación Python. Se ha utilizado las librerías más populares como numpy, pandas, sklearn, keras, ML Insights, pyArgrum, entre otros.

2

Diagnosis en Hematología/Oncología

La Hematología/Oncología es la rama de la medicina que se centra en el diagnóstico y tratamiento de numerosos trastornos sanguíneos, incluidos el cáncer, la anemia, la leucemia y la enfermedad de Hodgkin, y el diagnóstico y tratamiento de cánceres y tumores benignos y malignos.

La mejor manera de detectar el melanoma es examinando continuamente la piel de los pacientes, especialmente los lunares. Una llaga, una protuberancia o un tumor en la piel también puede ser un signo de melanoma u otro cáncer de piel. El melanoma se puede encontrar en varios lugares incluyendo la espalda, las nalgas, las piernas, el cuero cabelludo, el cuello, detrás de la oreja, las plantas de los pies, las palmas de las manos, dentro de la boca, los genitales y debajo de las uñas [10].

Según la Academia Americana de Dermatología (AAD), aproximadamente del 20% al 40% de los melanomas se desarrollan a partir de un lunar. Una llaga o tumor que sangra o cambios en el color de la piel también puede ser un signo de cáncer de piel [10].

La clave para tratar con éxito el melanoma es reconocer los síntomas a tiempo. Se recomienda realizar exámenes corporales anuales por un dermatólogo y examinarse la piel una vez al mes. Además, si un paciente ha tenido cáncer de piel, debe hacerse chequeos regulares para que un médico pueda examinar su piel. Hay un número de pruebas que se pueden ordenar para diagnosticar el cáncer de piel:

- Biopsia
- Tomografía computarizada (TC)
- Imagen de resonancia magnética (IRM)
- Tomografía por emisión de positrones (TEP)

Por lo que un clasificador que sea capaz de discernir entre imágenes de piel de cáncer maligno y benigno ayudaría al diagnóstico mensual que deben hacer los médicos a sus pacientes además de reconocer los síntomas a tiempo.

3

Conjunto de datos

Se ha hecho uso de un conjunto de datos generados por la International Skin Imaging Collaboration (ISIC) en donde las imágenes provienen de las siguientes fuentes: *Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland* y *University of Athens Escuela de Medicina* [11].

Todos los diagnósticos malignos se han confirmado mediante histopatología, y los diagnósticos benignos se han confirmado mediante el acuerdo de expertos, el seguimiento longitudinal o la histopatología.

El conjunto de datos fue seleccionado para el SIIM-ISIC Melanoma Classification Challenge organizado en Kaggle durante el verano de 2020 [12]. También, se ha hecho uso del conjunto de entrenamiento de la competición de 2019 [13]. El propósito de la competición consiste en predecir un objetivo binario para cada imagen. El modelo debe predecir la probabilidad entre 0.0 y 1.0 de que la lesión en la imagen sea maligna. La métrica utilizada para la competición el la curva AUC - ROC.

En este trabajo se utiliza el conjunto de datos de 2019 para el entrenamiento de modelos y el conjunto de datos de 2020 como un conjunto que ha sido proporcionado por una clínica con el fin de que esta pueda utilizar el modelo como un servicio hacia sus pacientes.

3.1 Análisis de datos

Las imágenes se proporcionan en formato DICOM, JPEG y TFRecord. En este trabajo se ha utilizado las imágenes en formato JPEG. También se proporcionan en los metadatos en archivos CSV.

	image_name	sex	age_approx	anatom_site_general_challenge	target
0	ISIC_0000000.jpg	female	55.0	anterior torso	0.0
1	ISIC_0000001.jpg	female	30.0	anterior torso	0.0
2	ISIC_0000002.jpg	female	60.0	upper extremity	1.0
3	ISIC_0000003.jpg	male	30.0	upper extremity	0.0
4	ISIC_0000004.jpg	male	80.0	posterior torso	1.0

Figura 3.1: Primeras cinco filas de los datos tabulares de entrenamiento de 2019.

	image_name	patient_id	sex	age_approx	anatom_site_general_challenge	diagnosis	benign_malignant	target
0	ISIC_2637011.jpg	IP_7279968	male	45.0	head/neck	unknown	benign	0
1	ISIC_0015719.jpg	IP_3075186	female	45.0	upper extremity	unknown	benign	0
2	ISIC_0052212.jpg	IP_2842074	female	50.0	lower extremity	nevus	benign	0
3	ISIC_0068279.jpg	IP_6890425	female	45.0	head/neck	unknown	benign	0
4	ISIC_0074268.jpg	IP_8723313	female	55.0	upper extremity	unknown	benign	0

Figura 3.2: Primeras cinco filas de los datos tabulares de entrenamiento de 2020.

Las columnas de los datos tabulares (Figura 3.1 y Figura 3.2) representan:

- image_name: identificador único, apunta al nombre de archivo de la imagen relacionada.
- patient_id: identificador único del paciente.
- sex: el sexo del paciente (cuando se desconozca, estará en blanco).
- age_approx: edad aproximada del paciente en el momento de la obtención de la imagen.
- anatomsite_general_challenge: ubicación de la lesión en el cuerpo.
- diagnosis: información del diagnóstico.
- benign_malignant: indicador del tipo de lesión.
- target: valor binario (siendo 0 benigno y 1 maligno).

Al representar el número de muestras de cada tipo, observamos que hay más registros benignos que malignos. Esto es habitual ya que en el mundo real la frecuencia de tumores benignos son mayores a los malignos. Por lo que produce un desequilibrio del conjunto de datos (Figura 3.3) el cual será comentado en la sección 3.2.1.

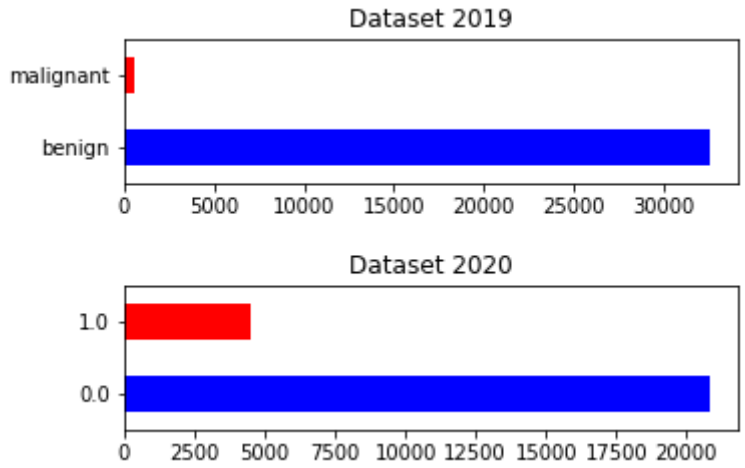

Figura 3.3: Distribución de clases de los conjuntos de datos.

Al visualizar de qué parte del cuerpo provienen la mayoría de las imágenes los conjuntos de imágenes, se observa que son del torso, extremidad inferior y superior (Figura 3.4)

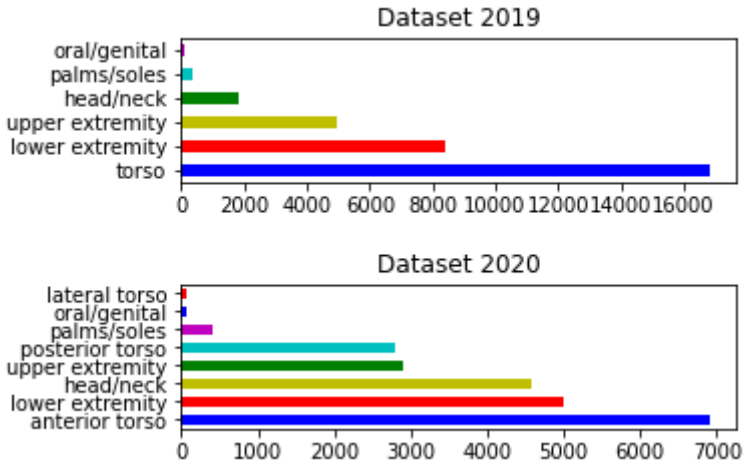

Figura 3.4: Distribución de la localización de las lesiones de los conjuntos de datos.

Se observa que la mayoría de las imágenes provienen de personas de mediana edad entre 40 y 60 años (Figura 3.5).

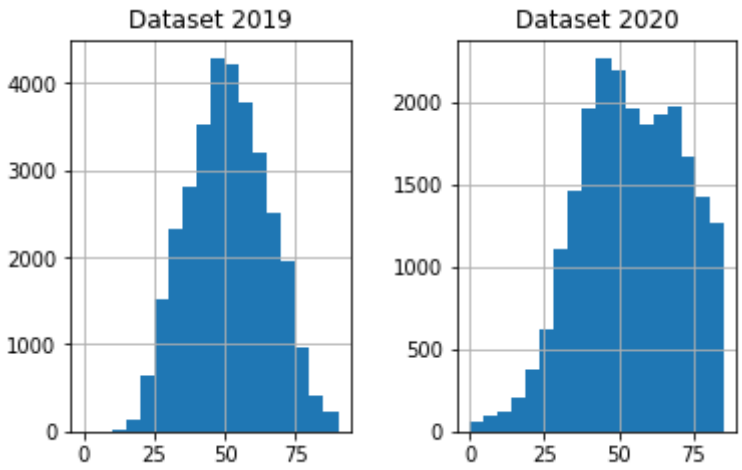

Figura 3.5: Distribución de edades de los conjuntos de datos.

Finalmente, se observa que la distribución del sexo está bastante equilibrada (Figura 3.6).

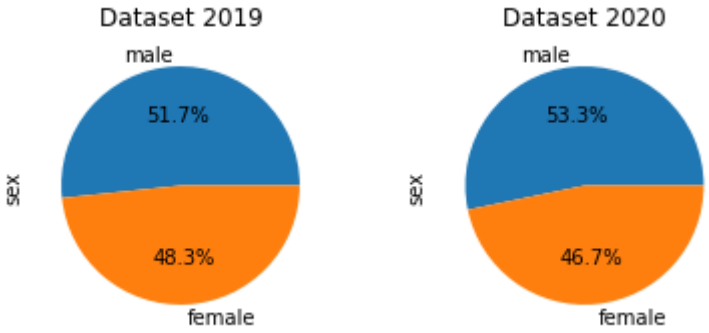

Figura 3.6: Distribución del sexo de los conjuntos de datos.

3.2 Preprocesamiento de datos

Respecto al preprocesamiento de las imágenes JPEG (con el fin de utilizarlas como entrenamiento en modelos basados en redes neuronales convolucionales) se han redimensionado las imágenes a una escala menor (224x224 píxeles) mediante la librería PIL.

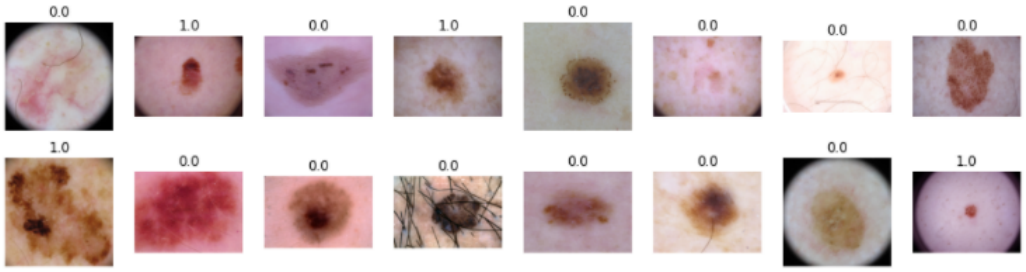

Figura 3.7: Muestras de imágenes junto a su etiqueta del conjunto de datos de 2019.

Para ello, se ha mantenido la relación de aspecto ya que si no es así, esto afectaría a que la lesión de la imagen cambie de aspecto, cambiando la simetría, la cual como se comentaba anteriormente es un factor muy importante en la detección de la enfermedad.

Sin embargo, al redimensionar las imágenes realmente se establece el alto a 224 píxeles y ancho se adapta con el fin de que se mantenga la relación de aspecto, por lo que las imágenes no son exactamente 224x224 píxeles como se observa en la Figura 3.8.

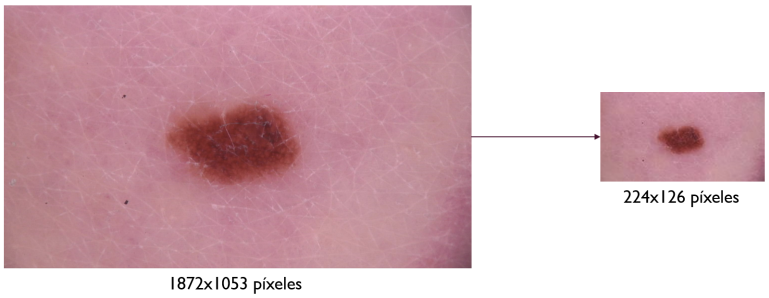

Figura 3.8: Ejemplo de redimensionamiento de la imagen ISIC_0052212.

En los datos tabulares del conjunto de datos de 2020, se ha observado que existen datos faltantes (595 casos). Se ha optado por eliminar estos registros junto a las columnas *diagnosis* y *benign_malignant*, quedando un total 32.531 número de filas.

Por otro lado, para el conjunto de datos de 2019 la competición proporciona dos archivos CSV los cuales se han extraído, unido, eliminado y renombrado columnas (Figura 3.9) creando un único archivo CSV, con el fin de que los datos tabulares de 2019 sean idénticos a los datos tabulares de entrenamiento de 2020.

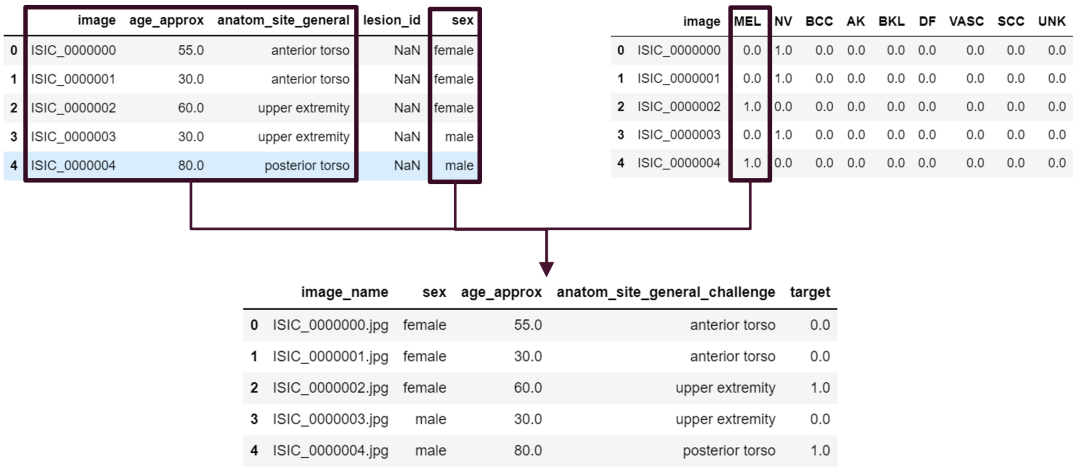

	image	age_approx	anatom_site_general	lesion_id	sex
0	ISIC_0000000	55.0	anterior torso	NaN	female
1	ISIC_0000001	30.0	anterior torso	NaN	female
2	ISIC_0000002	60.0	upper extremity	NaN	female
3	ISIC_0000003	30.0	upper extremity	NaN	male
4	ISIC_0000004	80.0	posterior torso	NaN	male

	image	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
0	ISIC_0000000	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	ISIC_0000001	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	ISIC_0000002	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	ISIC_0000003	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	ISIC_0000004	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

	image_name	sex	age_approx	anatom_site_general_challenge	target
0	ISIC_0000000.jpg	female	55.0	anterior torso	0.0
1	ISIC_0000001.jpg	female	30.0	anterior torso	0.0
2	ISIC_0000002.jpg	female	60.0	upper extremity	1.0
3	ISIC_0000003.jpg	male	30.0	upper extremity	0.0
4	ISIC_0000004.jpg	male	80.0	posterior torso	1.0

Figura 3.9: Conversión de los datos tabulares de 2019.

En el conjunto de 2019 también se ha observado que existen datos faltantes (2.851 casos) los cuales se han eliminado, quedando un total 22.480 número de filas.

Finalmente, se ha dividido el conjunto de datos de 2019 en cuatro subconjuntos: entrenamiento, validación, calibración y prueba. Se ha realizado una estratificación con el fin de mantener la misma frecuencia de casos benignos y malignos en todos los subconjuntos.

3.2.1 Desbalanceo de datos

La mayoría de los modelos utilizados para aprender de un conjunto de datos se diseñaron en torno al supuesto de que la distribución de las clases sea igual. Cuando estas están desequilibradas, muchos algoritmos de aprendizaje automático fallan y las métricas utilizadas para evaluar esos modelos, como la precisión de la clasificación (o *accuracy*), se vuelven engañosos [14].

Como se observa en la Figura 3.3, las clases conocidas no están balanceadas. Existe una frecuencia de imágenes sin cáncer del 98,2% y 80% e imágenes malignas del 1,8% y 20% en el conjunto de 2020 y 2019, respectivamente. Esto provoca que la clase minoritaria sea más difícil de predecir porque hay pocos ejemplos de esta clase, por definición, y aprender las características de los ejemplos para diferenciar esta clase frente a la clase mayoritaria. Además se observa que el debalanceo es distinto en los datos de 2019 y 2020.

Como consecuencia, la mayoría de los modelos tienen un rendimiento predictivo deficiente, específicamente para la clase minoritaria. Esto presenta un inconveniente ya que es más importante clasificar la clase minoritaria (casos malignos) correctamente y, por lo tanto, el problema es más sensible a los errores de clasificación para la clase minoritaria que para la clase mayoritaria.

Para intentar solucionar este problema, se han propuesto modificaciones a los algoritmos existentes para hacer que sean útiles para la clasificación desequilibrada, la selección de métricas de rendimiento y nuevas técnicas de preparación de datos y algoritmos de modelado.

3.2.1.1 Sobremuestreo y submuestreo

Un enfoque para abordar los conjuntos de datos desequilibrados es sobremuestrear la clase minoritaria en el conjunto de datos de entrenamiento antes de entrenar el modelo. Esto implica realizar *data augmentation* de los ejemplos de la clase minoritaria (Figura 3.10) volviendo a muestrear las muestras menos frecuentes para ajustar su cantidad en comparación con la clase predominante como se observa en la Figura 3.11.

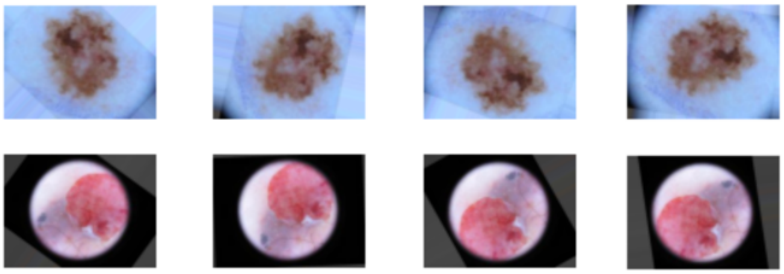

Figura 3.10: Ejemplo de sobremuestreo de muestras malignas mediante *data augmentation* del conjunto de entrenamiento.

La mayor parte de la atención de los métodos de muestreo para la clasificación se basa en el sobremuestreo de la clase minoritaria. Sin embargo, también se han desarrollado un conjunto de técnicas para submuestrear la clase mayoritaria.

La técnica de submuestreo más simple implica la selección aleatoria de ejemplos de la clase mayoritaria, eliminándolos del conjunto de datos de entrenamiento con el fin de equilibrar mejor la distribución de clases.

Una extensión de este enfoque es ser más perspicaz con los ejemplos de la clase mayoritaria que se eliminan implicando modelos heurísticos que intentan identificar ejemplos redundantes o críticos. Existen muchas técnicas de submuestreo que utilizan este tipo de heurísticas aunque no se profundizará en ellas en este trabajo.

Cabe destacar que los métodos de submuestreo se pueden utilizar junto con una técnica de sobremuestreo para la clase minoritaria, y esta combinación a menudo da como resultado mejor rendimiento que el uso únicamente del sobremuestreo o submuestreo en el conjunto de datos de entrenamiento [14], aunque se ha decidido no experimentar esta opción.

Figura 3.11: La figura de la izquierda es una representación visual del procedimiento del submuestreo y la figura de la derecha del sobremuestreo de datos [3].

3.2.1.2 Función de pérdida ponderada

Aunque cualquiera de las dos estrategias comentadas anteriormente equilibra el conjunto de datos, no aborda directamente los problemas causados por el desequilibrio de clase, sino que corre el riesgo de introducir nuevos problemas, dado que el sobremuestreo introduce muestras duplicadas, podría ralentizar el entrenamiento y también provocar *overfitting* en un modelo. Por otro lado, el submuestreo elimina cierto número de muestras y podría llevar al modelo a perder el aprendizaje de ciertas características importantes que podría haber aprendido de las muestras [15].

Otra alternativa para superar estos problemas es modificar la función de pérdida. Aplicar diferentes ponderaciones a la función de pérdida para diferentes muestras en función de si pertenecen a la clase mayoritaria o minoritaria. Básicamente, se asigna un mayor peso a las muestras asociadas con la clase minoritaria. Esta técnica se conoce como función de pérdida ponderada o *Weighted Loss Function*.

Al realizar el entrenamiento del modelo en Keras, se define un diccionario con las etiquetas y sus pesos asociados el cual se pasa al parámetro *class_weight*. Para calcular estas ponderaciones o pesos de cada clase se ha hecho uso de la librería *sklearn* (Figura 3.12) el cual utiliza la heurística inspirada en *Logistic Regression in Rare Events Data, King, Zen, 2001*.

```
from sklearn.utils import class_weight as cw

y_train = train_df['target'].to_numpy()
class_weight_ = cw.compute_class_weight('balanced',
                                        np.unique(y_train),
                                        y_train)
```

Figura 3.12: Cálculo de los pesos para cada clase.

4

Métricas

4.1 Evaluación de modelos

A continuación se explican los métodos de calibración usados y las métricas que se han empleado para evaluar y comparar los distintos modelos de redes neuronales convolucionales.

4.1.1 Precisión

Normalmente, al calcular la precisión en un conjunto de prueba, se considera la proporción de los ejemplos totales que el modelo clasifica correctamente. Es decir, nº de ejemplos clasificados correctamente / nº total de ejemplos.

Se interpreta la precisión como probabilidad de ser correcto. Pero como se ha comentado en la sección 3.2.1 esta métrica puede ser engañosa. Podemos descomponer esta probabilidad de ser correcto como la suma de dos probabilidades; la probabilidad de que el modelo sea correcto y un paciente tenga la enfermedad más la probabilidad de que el modelo sea correcto y el paciente esté sano:

$$Precisión = P(correcto \cap enfermedad) + P(correcto \cap sano)$$

La ley de la probabilidad condicional nos permite expandir aún más. Esta dice que la probabilidad de A y B es la probabilidad de A dado B multiplicado por la probabilidad de B.

$$Precisión = P(correcto)$$
$$Precisión = P(correcto \cap enfermedad) + P(correcto \cap sano)$$

Utilizando $P(A \cap B) = P(A|B)\ P(B)$ se tiene que:

$$Precisión = P(correcto|enfermedad)\ P(enfemedad) + P(correcto|sano)P(sano);$$
$$Precisión = P(+|enfermedad)\ P(enfemedad) + P(-|sano)P(sano)$$

Por lo que se reemplaza $P(+|enfermedad)$ por la probabilidad de predecir positivo dado que un paciente tiene la enfermedad (sensibilidad). De manera similar, $P(-|sano)$ por la probabilidad de predecir negativo dado un paciente es sano (especificidad).

4.1.2 Sensibilidad y especificidad

Como se ha visto anteriormente, a partir de la precisión se derivan otras métricas de evaluación muy importantes como son la sensibilidad y la especificidad.

La probabilidad de que un paciente tenga una enfermedad en una población se conoce como prevalencia. La probabilidad de estar sano es simplemente uno menos la probabilidad de tener una enfermedad o uno menos la prevalencia. Por lo que se puede reescribir la precisión como:

$$Precisión = Sensibilidad \text{ x } P(enfemedad) + Especificidad \text{ x } P(sano);$$
$$Precisión = Sensibilidad \text{ x } prevalencia + Especificidad \text{ x } (1 - prevalencia)$$

La sensibilidad expresa, dado que sabemos que un paciente tiene una enfermedad, ¿cuál es la probabilidad de que el modelo prediga positivo? En la clínica, el médico que utiliza un modelo de IA puede estar interesado en una pregunta diferente. Por ejemplo, si el modelo predice positivo en un paciente, ¿cuál es la probabilidad de que realmente tenga la enfermedad?. A esto se le conoce como valor predictivo positivo o PPV (*Positive Predictive Value*) del modelo.

Del mismo modo ocurre con la especificidad, el médico puede estar interesado en conocer la probabilidad de que un paciente esté sano, dado que la predicción del modelo es negativa, en otras palabras, el valor predictivo negativo o NPV (*Negative Predictive Value*) del modelo.

4.1.3 Matriz de confusión

Para ver su relación, se puede utilizar la matriz de confusión. Esta se emplea para observar el rendimiento de un clasificador en forma de tabla (Figura 4.1). Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, es decir, en términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo el modelo a la hora de pasar por el proceso de aprendizaje con los datos.

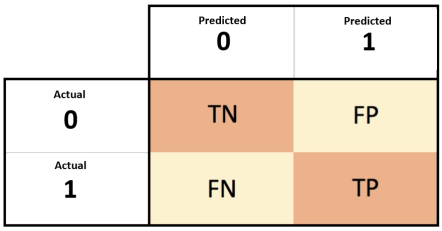

Figura 4.1: TN (verdadero positivo): recuento de resultados que fueron originalmente negativos y se predijeron como negativos. FP (falso positivo): recuento de resultados que originalmente fueron negativos pero que se pronosticaron positivos. FN (falso negativo): recuento de resultados que originalmente fueron positivos pero que se pronosticaron negativos. TP (verdadero positivo): recuento de resultados que originalmente fueron positivos y se predijeron como positivos [4].

4.1.4 Curva AUC - ROC

La curva AUC - ROC es una medida de rendimiento que se utiliza a menudo para los problemas de clasificación en varios valores de umbral. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Esta indica cuánto es capaz el modelo de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para distinguir entre pacientes con cáncer y sanos [5].

Para definir la curva AUC - ROC (Figura 4.3) se utilizan otras métricas como TRP (*True Positive Rate*, sensibilidad), especificidad y FPR (*False Positive Rate*):

$$TPR = \frac{TP}{TP+FN}$$

$$Especificidad = \frac{TN}{TN+FP}$$

$$FPR = 1 - especificidad = \frac{FP}{TN+FP}$$

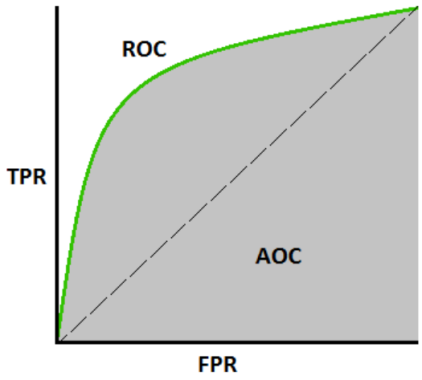

Figura 4.2: Curva AUC - ROC [5]

El objetivo es obtener un modelo que tenga un AUC cercano al 1, lo cual significa que mantiene una buena medida de separabilidad. En el caso de que el AUC es 0,5, el modelo no tiene capacidad de separación de clases en absoluto.

4.2 Calibración de modelos

Para entender la calibración, se debe conocer los distintos tipos de predicciones que un modelo puede realizar [16]:

1. *"Hard" Prediction*: donde se desea predecir el valor exacto de y dado x.
2. *"Ranking" Prediction*: donde se proporciona una puntuación para x, en el cual una puntuación más alta indica que es más probable que y sea verdadera. Las puntuaciones pueden no ser necesariamente significativas como probabilidades, de hecho, los valores pueden ni siquiera ser números entre 0 y 1.
3. *"Probabilistic" Prediction*: donde se desean predecir las probabilidades de todos los valores de y dado x.

Dado los tipos de predicciones, es relativamente fácil transformar un tipo de predicción a otra. En el caso de tener un modelo que calcula una predicción del segundo tipo, la calibración permite transformarlo al tercer tipo de predicción. Por lo tanto, la calibración consiste en ajustar las predicciones de un modelo para que sean probabilísticamente significativas.

En los sistemas de toma de decisiones del mundo real, las redes de clasificación no solo deben ser precisas, sino que también deben indicar cuando es probable que sean incorrectas. Como indican los experimentos de la publicación *On Calibration of Modern Neural Networks* [6], se descubre que las redes neuronales modernas, a diferencia de las de hace unas década, están mal calibradas.

Esto es un aspecto importante ya que queremos que nuestro modelo sea muy preciso, es decir, cuando se prediga un 99% signifique que cometerá un error 1 vez cada 100 predicciones. Otra razón por la cual se debe tener en cuenta la calibración es con el fin de evaluar la calidad del modelo.

Para calibrar un modelo se debe ajustar una función que mapee la puntuación no calibrada a la probabilidad real. A continuación, se explican brevemente los diferentes métodos de calibración que varían en el tipo de función de ajuste.

4.2.1 *Platt Scaling*

Es un método de calibración paramétrico que utiliza predicciones no probabilísticas como entrada para la calibración. La entrada se utiliza para ajustar a un modelo de regresión logística que devuelve probabilidades calibradas [6].

4.2.2 *Isotonic Regression*

Es un método no paramétrico que se usa ampliamente para la calibración. Aprende una función constante f por partes que genera probabilidades calibradas $\hat{q}_i = f(\hat{p}_i)$. La función f intenta minimizar el error cuadrático [17],

$$\begin{aligned} \min_{\substack{M \\ \theta_1,\ldots,\theta_M \\ a_1,\ldots,a_{M+1}}} \quad & \sum_{m=1}^{M} \sum_{i=1}^{n} \mathbf{1}(a_m \leq \hat{p}_i < a_{m+1}) \, (\theta_m - y_i)^2 \\ \text{subject to} \quad & 0 = a_1 \leq a_2 \leq \ldots \leq a_{M+1} = 1, \\ & \theta_1 \leq \theta_2 \leq \ldots \leq \theta_M. \end{aligned}$$

Figura 4.3: Regresión Isotónica [6]

donde M es el número de intervalos, $a_1, ..., a_{M+1}$ son los límites de los intervalos y $C_1, ..., C_M$ son los valores de salida de la función f [6].

4.2.3 *Beta Calibration*

Es un método de calibración paramétrico que se basa en la distribución beta. Es similar a la regresión logística (*Platt Scaling*), excepto que la calibración beta se deriva en base a la distribución beta, la regresión logística se deriva en base a la distribución gaussiana. La calibración beta tiene 3 parámetros a, b y c, y puede ajustar más distribuciones en comparación con la regresión logística,

$\mu_{beta}(\hat{p}; a, b, c) = \frac{1}{1+1/(e^c \frac{\hat{p}^a}{(1-\hat{p})^b})}$ donde $a, b \geq 0$ y $c \in \mathbb{R}$ [17]

Las probabilidades predichas se convierten a un espacio logarítmico para utilizar la regresión logística para el ajuste. Los parámetros ajustados se utilizan con $\mu_{beta}(\hat{p}_{test}; a, b, c)$ para calcular la probabilidades calibradas $\hat{p}_{test}$ [17].

4.2.4 *Spline Calibration*

Este método utiliza *smoothing splines* para calibrar probabilidades. Este enfoque no es paramétrico, como la regresión isotónica. En cambio, típicamente supera a la regresión isotónica ya que tiene la libertad de ajustar una *spline* cúbica en lugar de solo una función constante por partes [18].

4.3 Métricas de calibración

4.3.1 *Reliability Diagram*

Los diagramas de fiabilidad son una representación de la calibración del modelo. Estos diagramas representan la precisión esperada de la muestra en función de la confianza. Si el modelo está perfectamente calibrado, entonces el diagrama debe trazar la función de identidad. Cualquier desviación de una diagonal perfecta representa un error de calibración. Para estimar la precisión esperada a partir de muestras finitas, se agrupan las predicciones en contenedores de intervalo M (cada uno de tamaño 1 / M) y se calcula la precisión de cada contenedor [6].

4.3.2 *Brier Score*

La puntuación de Brier mide la diferencia cuadrática media entre la probabilidad predicha y el resultado real. La puntuación de Brier siempre toma un valor entre cero y uno, ya que esta es la mayor diferencia posible entre una probabilidad predicha (que debe estar entre cero y uno) y el resultado real (que puede tomar valores de solo 0 y 1). Cuanto menor sea la pérdida de puntuación de Brier, mejor es el modelo [19].

4.3.3 *Log Loss*

La pérdida logarítmica, también llamada pérdida de entropía cruzada, se define en estimaciones de probabilidad. Se usa comúnmente en regresión logística (multinomial) y redes neuronales. Para la clasificación binaria con una etiqueta verdadera $y \in \{0, 1\}$ y una estimación de probabilidad $p = Pr(y = 1)$, la pérdida logarítmica por muestra es la probabilidad logarítmica negativa del clasificador dada la etiqueta verdadera:

$$L_{log}(y, p) = -logPr(y|p) = -(ylog(p) + (1 - y)log(1 - p)) \text{ [20]}$$

5

Análisis de resultados

5.1 Modelos implementados

Se han implementado todos los modelos empleando la red DenseNet-121 [21] ya que ha demostrado su eficacia en el diagnóstico de enfermedades con imágenes médicas como explica la publicación *DenseNet Convolutional Neural Networks Application for Predicting COVID-19 Using CT Image* [22].
Se subraya que la solución del ganador de la competición de 2020 [23] consistió en ensamblar 18 modelos diferentes entrenados con diferentes tipos de redes neuronales convolucionales, el cual requiere mucho tiempo de entrenamiento y de computación.
A continuación se mencionan los modelos implementados en este trabajo:

- **Modelo 1**. Se entrena una red neuronal únicamente con el conjunto de imágenes de 2019 sin tener en cuenta el desequilibrio de clases ni los datos tabulares.
- **Modelo 2**. Se implementan 3 modelos con el conjunto de imágenes de 2019 los cuales intentan solucionar el desequilibrio de clases (sobremuestreo, submuestreo y función de pérdida ponderada).
- **Modelo 3**. Se realiza la calibración mediante los distintos métodos sobre el modelo 2 y se introducen los datos tabulares de 2019 a través de una red bayesiana.
- **Modelo 4**. Se entrena una red neuronal únicamente con el conjunto de imágenes de 2020 intentando solucionar el desequilibrio de clases mediante la función de pérdida ponderada.

5.2 Red bayesiana

Para los modelos 3 y 4 se debe diseñar e implementar una red bayesiana con el fin de identificar e incorporar los datos tabulares (factores) que influyen en el cáncer de piel. A continuación, se muestra el diseño de la red bayesiana y un análisis de los modelos. Cabe recordar que el conjunto de datos 2020 se consideran datos proporcionados por una hipotética clínica que desea proporcionar un servicio de diagnóstico. Se asume que las imágenes de la clínica han sido obtenidas en las mismas condiciones y herramientas que el conjunto de datos de 2019.

5.2.1 *Naive Bayes*

Un clasificador *Naive Bayes* aprende a partir de los datos de entrenamiento la probabilidad condicional de cada atributo A_i dada la etiqueta de clase C. Después, la clasificación se realiza aplicando la regla de Bayes para calcular la probabilidad de C dada la instancia particular de $A_1, ..., A_n$, y luego predecir la clase con la probabilidad posterior más alta [7].

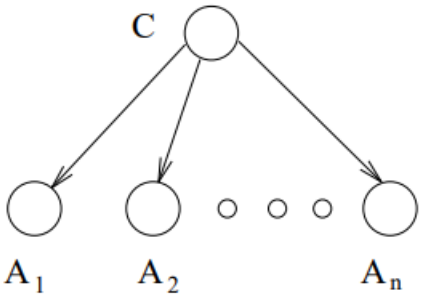

Figura 5.1: Estructura de una red *Naive Bayes* [7] donde C es la clase a predecir y $A_1, ..., A_n$, los factores o atributos, por ejemplo edad, sexo, localización que influyen en el caso del cáncer.

Esto se puede observar con mayor detalle mediante un ejemplo [24]: supongamos que un paciente se encuentra mal, acude a una clínica y el especialista le sugiere hacer una serie de exámenes. Obtenido los resultados, resulta que el paciente da positivo para una enfermedad muy rara que solo afecta a 0,1% de la población. Supongamos que el test identifica correctamente a un 99% de las personas que tienen la enfermedad e identifica incorrectamente al 1% de las personas. Entonces, ¿cuál es la probabilidad de que realmente el paciente no posea esta enfermedad habiendo dado positivo el test, o lo que es lo mismo, su valor predictivo positivo (visto en la sección 4.1.2). En un principio, se podría suponer que sería de un 99%, porque es la certeza del test, pero eso no es correcto.

Mediante el teorema de Bayes se puede obtener la probabilidad de tener la enfermedad H, habiendo dado positivo el test E. Para ello se necesita considerar la probabilidad a priori de que la hipótesis fuera verdad, es decir, cuán probable era tener la enfermedad antes de los resultados de los exámenes $P(H)$. Se multiplica por la probabilidad del evento (que el test de positivo) teniendo la enfermedad, $P(E|H)$, en otras palabras, cuán probable es un resultado positivo si el paciente tuviera la

enfermedad. Finalmente, se divide entre la probabilidad total de que ocurra ese evento (probabilidad de un resultado positivo). En otros términos:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

donde $P(E)$ es la combinación de la probabilidad de tener la enfermedad y resultar positivo correctamente, sumada a la probabilidad de no tener la enfermedad y obtener un falso positivo. Por lo que:

$$P(H|E) = \frac{P(E|H)P(H)}{P(H)P(E|H)+P(\neg H)P(E|\neg H)}$$

La probabilidad a priori de que la hipótesis sea verdad $P(H)$, es en general, la parte más difícil de resolver de la ecuación. Y, a veces, no es nada más que una suposición. Pero en este caso, sería razonable empezar por la frecuencia de la enfermedad en la población, es decir, 0,1%:

$$P(H|E) = \frac{0{,}99*0{,}001}{(0{,}001*0{,}99)+(0{,}999*0{,}01)} = 0,09$$

Por lo que el paciente tiene un 9% de probabilidad de realmente tener la enfermedad después de resultar positivo.
Pero, ¿qué sucedería si el paciente fuera a otro doctor para pedir una segunda opinión y realizar otro test independiente del anterior?. Resulta que ese examen vuelve a ser positivo, ¿cuál es la probabilidad de que el paciente tenga la enfermedad?. Se utiliza la fórmula de Bayes nuevamente, excepto de que esta vez, para la probabilidad a priori de tener la enfermedad P(H), se debe poner la probabilidad posterior que se tuvo antes (9%) ya que el paciente obtuvo un examen positivo. Si se realiza el cálculo:

$$P(H|E) = \frac{0{,}99*0{,}09}{(0{,}09*0{,}99)+(0{,}91*0{,}01)} = 0,907$$

Se obtiene que la probabilidad total basada en 2 exámenes positivos independientes es del casi 91%. Lo que tiene mucho sentido ya que es muy improbable obtener 2 resultados positivos, independientes, que estén mal.

El teorema de Bayes supone que cada variable de entrada depende de todas las demás variables. Ésta es una causa de complejidad en el cálculo. Se puede eliminar este supuesto y considerar cada variable de entrada como independiente entre sí. Lo cual cambia el modelo de un modelo de probabilidad condicionalmente dependiente a un modelo de probabilidad condicionalmente independiente. Esta simplificación del Teorema de Bayes es común y se usa ampliamente para problemas de modelado predictivo de clasificación y generalmente se conoce como *Naive Bayes* [25].

Este proceso se puede llevar a cabo en el caso de cáncer de piel. Para ello se utilizan los datos tabulares como evidencias (edad, sexo y localización) y la predicción calibrada del clasificador como la probabilidad a priori. Por ejemplo, supongamos un caso en el cual el clasificador realiza una predicción sobre una imagen y se obtiene un 90% de que sea una lesión maligna. El paciente tiene 20 años, es mujer y la imagen proviene del torso. Por lo que se puede aplicar:

$$posteriori(cáncer) = \frac{P(edad(20)|cáncer)P(mujer|cáncer)P(torso|cáncer)P(clasificador)}{constante\ de\ normalización}$$

donde la constante de normalización se puede calcular:

$P(edad(20)|cáncer)P(mujer|cáncer)P(torso|cáncer)P(clasificador) +$
$P(edad(20)|\neg\ cáncer)P(mujer|\neg\ cáncer)P(torso|\neg\ cáncer)P(\neg\ clasificador)$

Todos los atributos como el sexo, edad y localización (A_i) son condicionalmente independientes dado el valor de la clase C (Figura 5.1).

La suposición anterior podría ignorar las correlaciones entre la edad, el sexo y la ubicación de una lesión en el cáncer de piel. Esta observación plantea la pregunta de si se puede mejorar el rendimiento de un clasificador *Naive Bayes* evitando suposiciones injustificadas de independencia. Para ello, se debe intentar realizar algún tipo de análisis de las asociaciones entre los atributos.

5.2.2 Red de asociación

Una forma de explorar los datos es con el uso del método HNet (*Graphical Hypergeometric Networks*) que permite un examen profundo de las asociaciones entre características. Este utiliza pruebas estadísticas para determinar y encontrar relaciones significativas entre las variables [26]. Cabe mencionar que cuando se habla de asociación o relaciones, significa que ciertos valores de una variable tienden a coexistir con ciertos valores de otra variable.

Siguiendo el tutorial publicado *Explore and understand your data with a network of significant associations* [26], se utilizan los datos tabulares de 2019 para obtener el siguiente grafo de asociación:

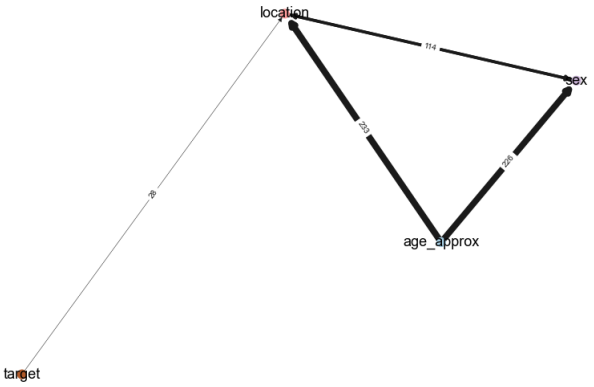

Figura 5.2: Grafo de asociación obtenido de HNet.

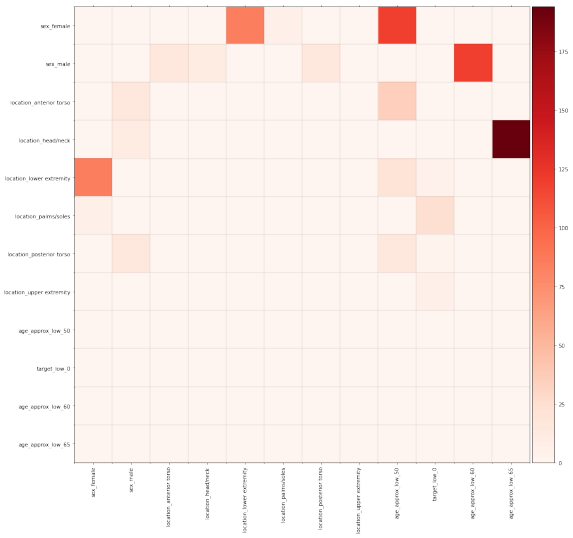

Figura 5.3: Mapa de calor obtenido de HNet.

Las variables también se pueden analizar desde la red calculando la importancia de la característica por categoría. Las variables de alta puntuación demuestran que muchos de los nodos separados son importantes en la red [26]. Se observa que el sexo y la localización son las 2 variables principales.

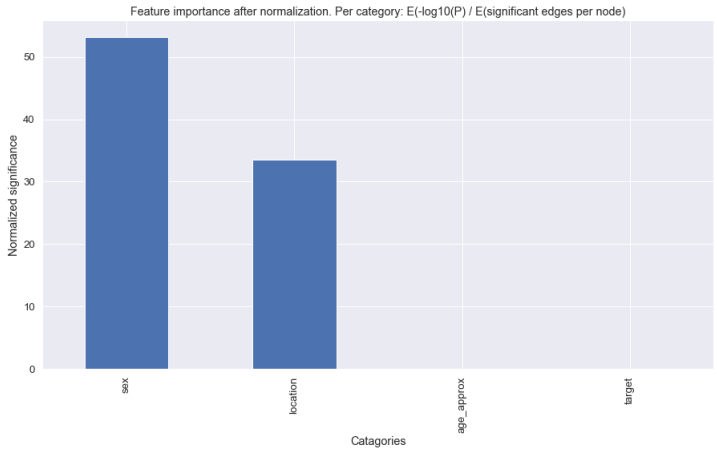

Figura 5.4: Importancia de cada característica obtenido de HNet.

Teniendo en cuenta la asociación que ha encontrado HNet a partir de los datos tabulares, se implementarán dos modelos bayesianos: *Naive Bayes* y otro con las asociaciones de HNet (Figura 5.5). Las redes se implementan mediante la librería pyAgrum [27].

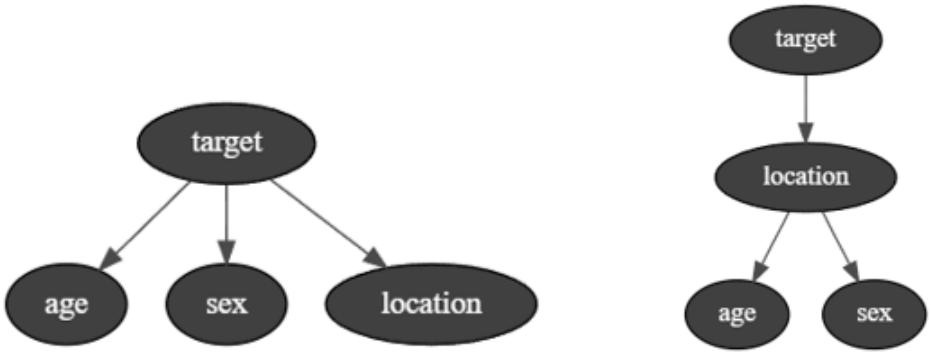

Figura 5.5: *Naive Bayes* frente a la red bayesiana diseñada a partir de HNet.

5.3 Comparación

A continuación, se muestran los resultados de los modelos. Cabe mencionar que todos los modelos se han entrenado durante 100 épocas.

Modelo 1. Este modelo utiliza las imágenes de 2019. Se obtiene una curva ROC - AUC en el conjunto de test de 0,75. Como es de esperar, al observar la matriz de confusión, el modelo tiende a clasificar las casos benignos mejor que los malignos ya que existes escasas imágenes malignas y el modelo no tiene la suficiente información para poder tener una buena separabilidad de clases.

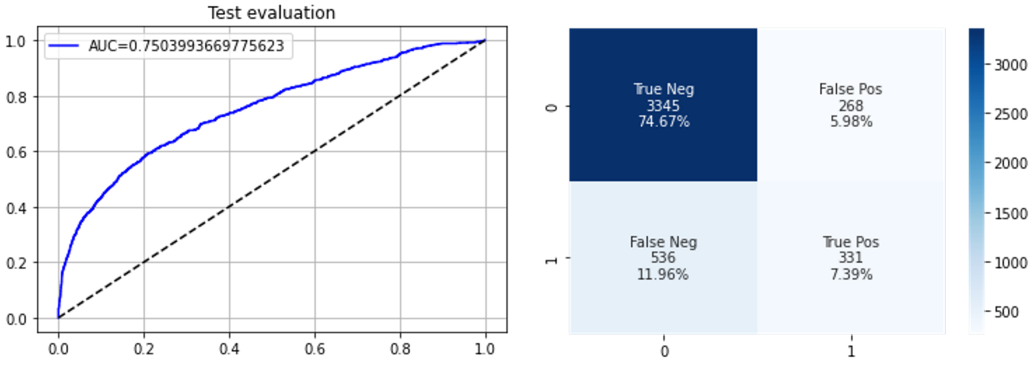

Figura 5.6: Métricas de evaluación del Modelo 1.

Modelo 2.1 (submuestreo). Este modelo utiliza las imágenes de 2019. Se obtiene una curva ROC - AUC en el conjunto de test de 0,78. Al observar la matriz de confusión, en este caso el modelo clasifica los casos malignos mejor que el Modelo 1, aunque debido a la eliminación de los casos benignos (los cuales probablemente son importantes, hasta quizás críticos) se observa que tiene una peor identificación que el Modelo 1.

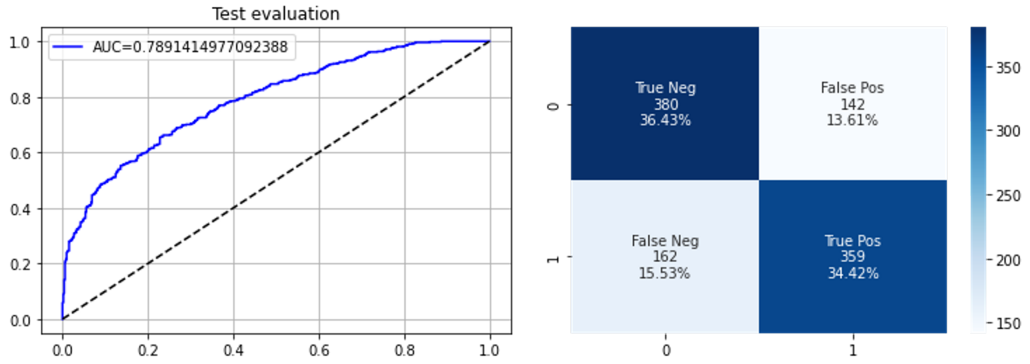

Figura 5.7: Métricas de evaluación del Modelo 2.1.

Modelo 2.2 (función de pérdida ponderada). Este modelo utiliza las imágenes de 2019. Se obtiene una curva ROC - AUC en el conjunto de test de 0,79. Al observar la matriz de confusión, se obtiene un modelo que clasifica mejor que el modelo de submuestreo.

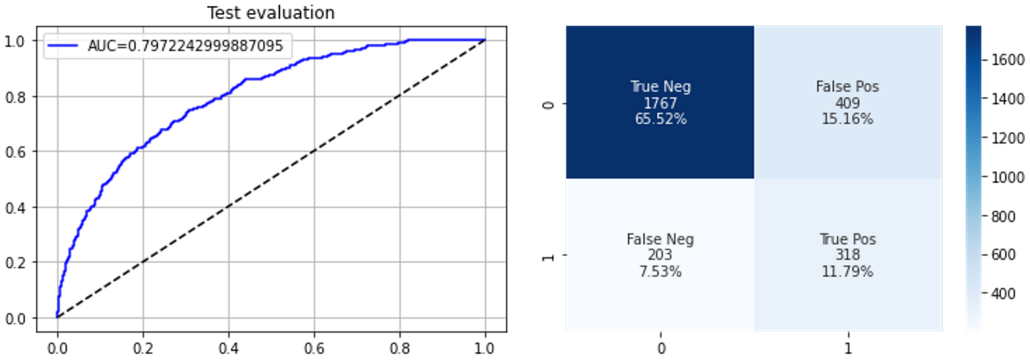

Figura 5.8: Métricas de evaluación del Modelo 2.2.

Modelo 2.3 (sobremuestreo). Este modelo utiliza las imágenes de 2019. Se obtiene una curva ROC - AUC en el conjunto de test de 0,63. Al observar la matriz de confusión, es el peor modelo a nivel de clasificación, además no es capaz de distinguir correctamente los casos malignos ya que se obtiene muy pocos verdadero positivos.

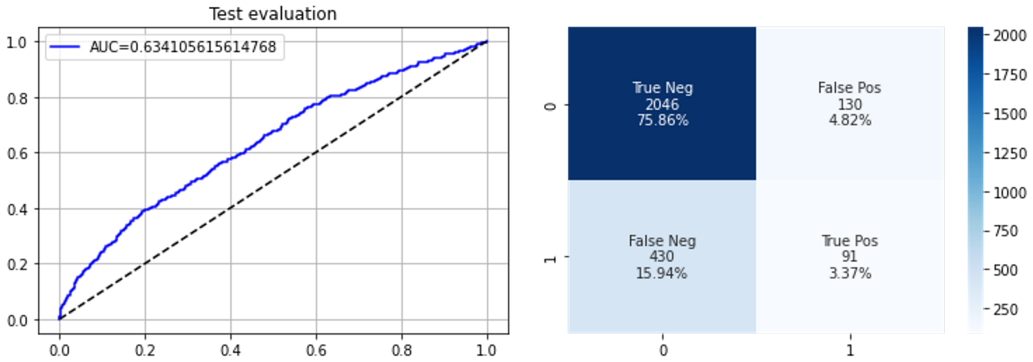

Figura 5.9: Métricas de evaluación del Modelo 2.3.

Modelo 3. Estos modelos calibrados utilizan las imágenes y los datos tabulares de 2019. Se han calibrado los tres modelos anteriores mediante los métodos de calibración vistos anteriormente. Respecto al conjunto de datos de calibración, este se ha aumentado uniéndolo al conjunto de validación, obteniendo así más datos para la calibración.
Se observa que una vez calibrados los modelos, los diagramas de fiabilidad se ajustan más a la diagonal perfecta. A continuación, se muestran los diagramas de fiabilidad antes de calibrar y después los resultados de cada método.

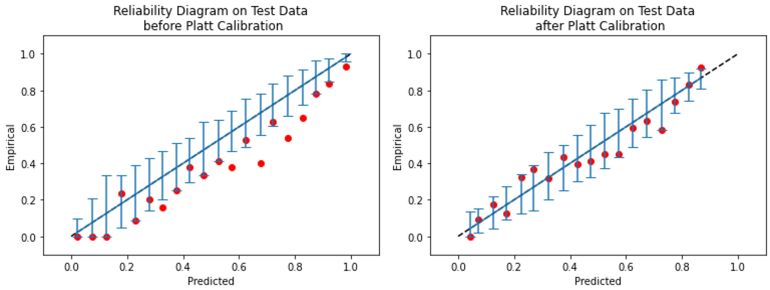

Figura 5.10: Resultados de *Platt Calibration* para el modelo de submuestreo.

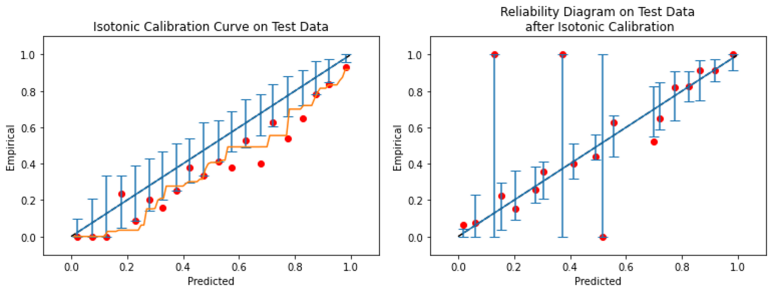

Figura 5.11: Resultados de *Isotonic Calibration* para el modelo de submuestreo.

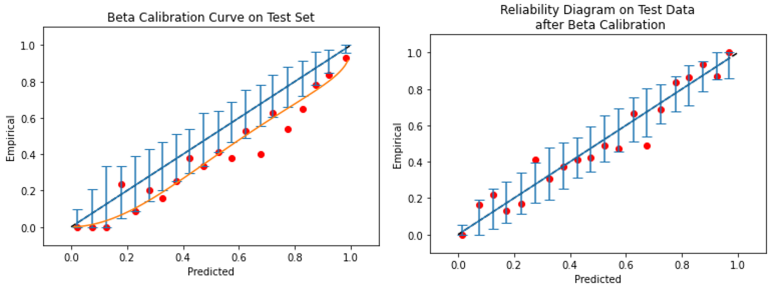

Figura 5.12: Resultados de *Beta Calibration* para el modelo de submuestreo.

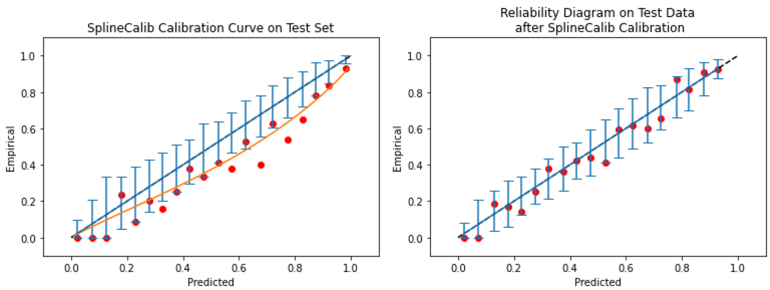

Figura 5.13: Resultados de *Spline Calibration* para el modelo de submuestreo.

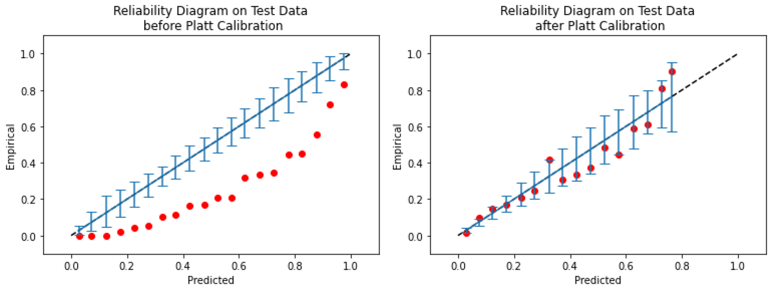

Figura 5.14: Resultados de *Platt Calibration* para el modelo con la función de pérdida ponderada.

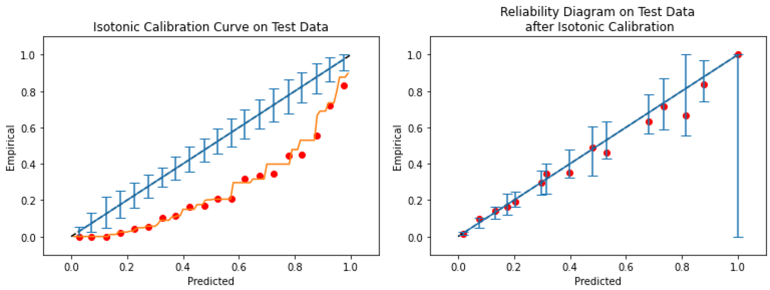

Figura 5.15: Resultados de *Isotonic Calibration* para el modelo con la función de pérdida ponderada.

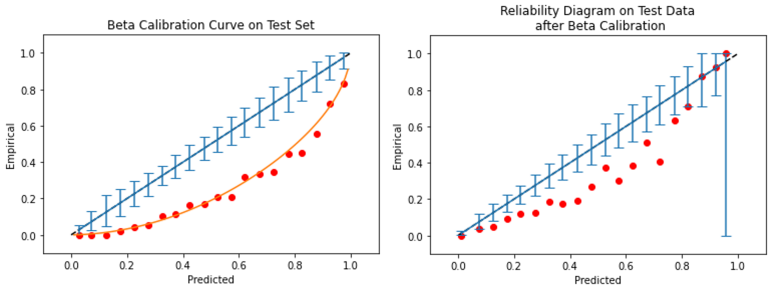

Figura 5.16: Resultados de *Beta Calibration* para el modelo con la función de pérdida ponderada.

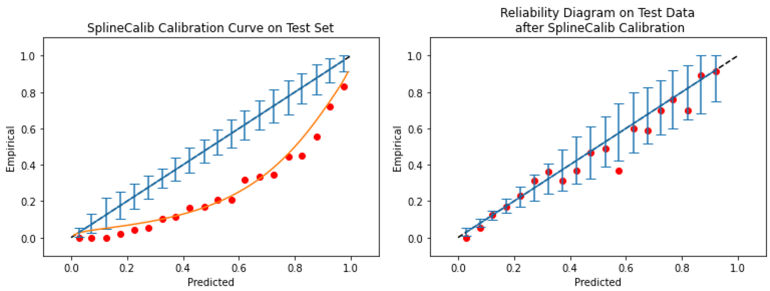

Figura 5.17: Resultados de *Spline Calibration* para el modelo con la función de pérdida ponderada.

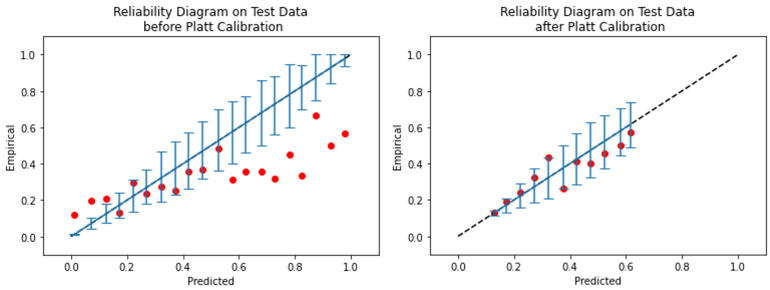

Figura 5.18: Resultados de *Platt Calibration* para el modelo de sobremuestreo.

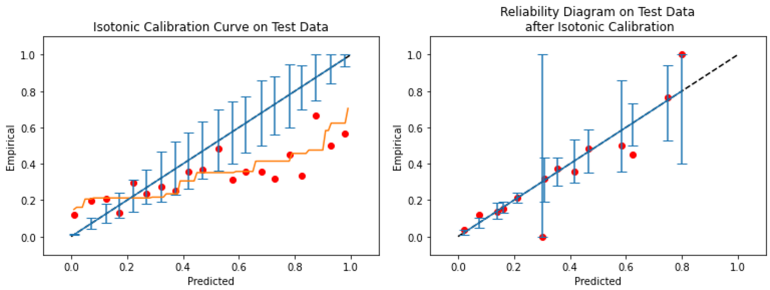

Figura 5.19: Resultados de *Isotonic Calibration para el modelo de sobremuestreo.*

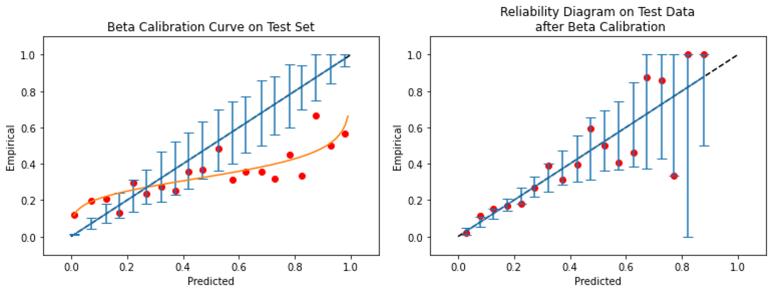

Figura 5.20: Resultados de *Beta Calibration* para el modelo de sobremuestreo.

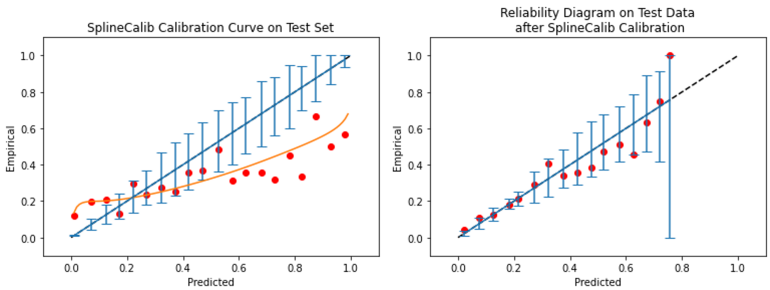

Figura 5.21: Resultados de *Spline Calibration* para el modelo de sobremuestreo.

Undersampling	Weighted loss	Oversampling
No calibration [[268 254] [74 447]] Platt [[373 149] [146 375]] Iso [[427 95] [195 326]] Beta [[371 151] [145 376]] Spline [[383 139] [158 363]]	No calibration [[1533 643] [144 377]] Platt [[2059 117] [346 175]] Iso [[2089 87] [374 147]] Beta [[1854 322] [235 286]] Spline [[2078 98] [365 156]]	No calibration [[1987 189] [370 151]] Platt [[2097 79] [437 84]] Iso [[2130 46] [467 54]] Beta [[2131 45] [468 53]] Spline [[2106 70] [442 79]]

Figura 5.22: Matrices de confusión de los modelos calibrados con los distintos métodos.

Para este trabajo y debido a los pocos datos de la clase minoritaria en el grupo de prueba utilizado para la calibración, se observa un comportamiento con intervalos muy amplios, por lo que se deja como trabajo futuro la obtención de más datos que permita una calibración más ajustada, con intervalos de confianza menores.

Posteriormente, se implementa un clasificador *Naive Bayes* mediante la librería pyAgrum. Antes, se muestra un caso concreto: para la imagen 'ISIC_0031543.jpg' maligna, el clasificador de imágenes predijo 0,9745324 (probabilidad a priori) y los atributos de la imagen indica que la imagen proviene de la extremidad superior de una mujer de sesenta años.
Para aplicar Bayes, se calculan las probabilidades necesarias para ello:

Predicción del clasificador de imágenes →

target	
False	True
0.0255	0.9745

	sex	
target	female	male
False	0.4728	0.5272
True	0.4501	0.5499

	age						
target	teen	twenties	thirties	forties	fifties	sixties	old
False	0.0256	0.0512	0.1264	0.1980	0.1824	0.1602	0.2561
True	0.0016	0.0145	0.0624	0.1440	0.1921	0.2009	0.3845

	location							
target	anterior torso	head/neck	lateral torso	lower extremity	oral/genital	palms/soles	posterior torso	upper extremity
False	0.3041	0.2040	0.0017	0.2300	0.0022	0.0109	0.1276	0.1195
True	0.3042	0.2016	0.0032	0.1811	0.0044	0.0451	0.0971	0.1634

Figura 5.23: Probabilidades obtenidas de los datos tabulares de 2019 para aplicar *Naive Bayes*.

Por lo que, aplicando *Naive Bayes* se obtiene:

$$Numerador = P(edad(60)|cáncer)P(mujer|cáncer)$$
$$P(extremidad\ superior|cáncer)P(clasificador)$$

$$Constante\ de\ normalización = Numerador + P(edad(60)|\neg cáncer)$$
$$P(mujer|\neg cáncer)P(extremidad\ superior|\neg cáncer)(1 - P(clasificador))$$

$$posteriori(cáncer) = \frac{Numerador}{constante\ de\ normalización} =$$
$$\frac{0{,}2009 \cdot 0{,}4501 \cdot 0{,}1634 \cdot 0{,}9745324}{(0{,}2009 \cdot 0{,}4501 \cdot 0{,}1634 \cdot 0{,}9745324) + (0.1602 \cdot 0.4728 \cdot 0.1195 \cdot (1 - 0{,}9745324))} = 0,9842 = 98,42\%$$

A continuación, se realiza la inferencia y se calcula la probabilidad a posteriori del mismo ejemplo con la librería pyAgrum:

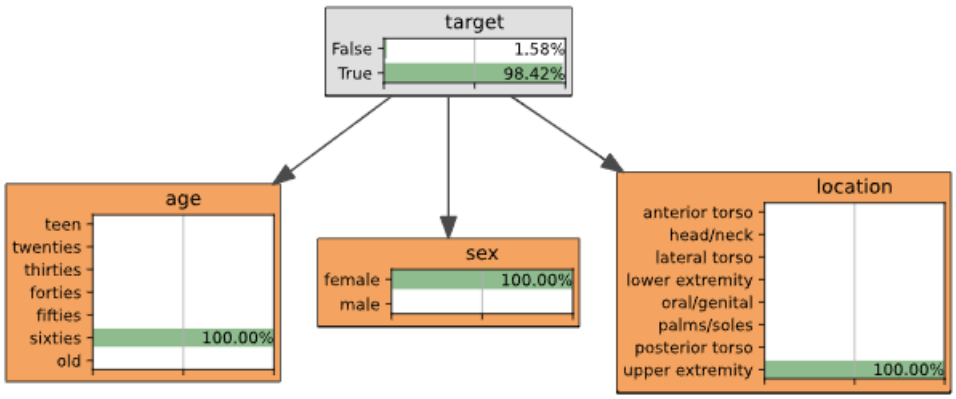

Figura 5.24: *Naive Bayes* con la librería pyAgrum.

Una vez comprobado que los resultados son idénticos, se procede a obtener las nuevas predicciones de los tres modelos. Cabe mencionar que las matrices de confusión del clasificador de imágenes son ligeramente distintas a las mostradas anteriormente ya que se ha vuelto a realizar predicciones sobre el conjunto de prueba.

Se observa que para todos los modelos, *Naive Bayes* es capaz de mejorar la clasificación de casos benignos y malignos mediante los datos tabulares. Con esto podemos afirmar que los atributos como la edad, sexo y localización (entre otros) deben tenerse en cuenta al realizar predicciones ya que son factores que influyen al diagnóstico de cáncer de piel, además de la propia imagen.

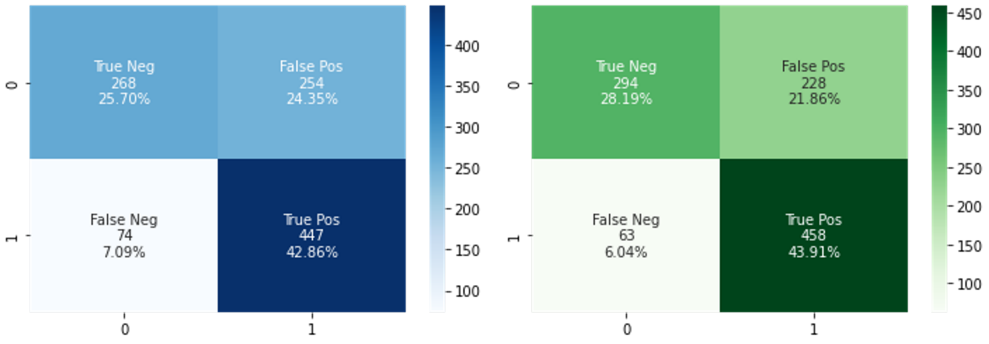

Figura 5.25: Matriz de confusión del Modelo 3.1 (antes de *Naive Bayes* y despúes).

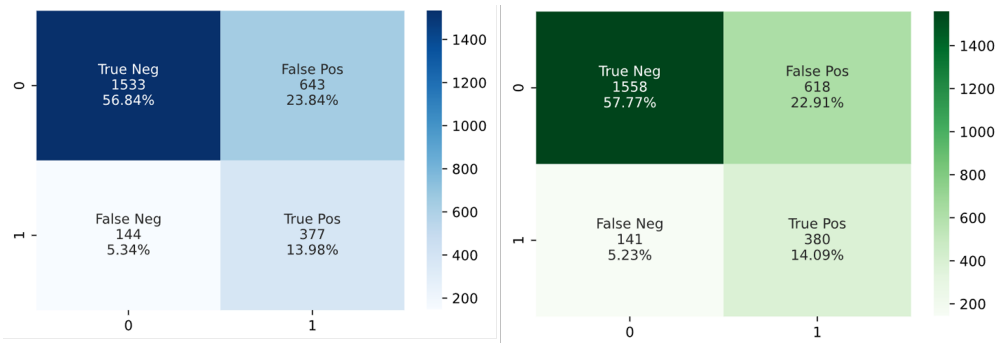

Figura 5.26: Matriz de confusión del Modelo 3.2 (antes de *Naive Bayes* y despúes).

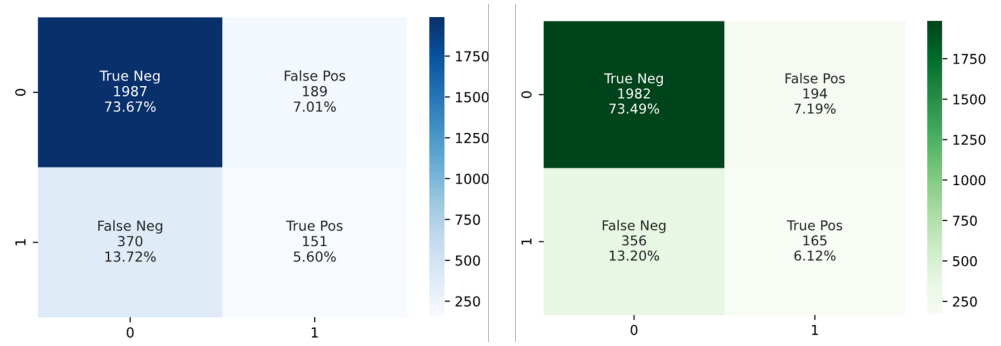

Figura 5.27: Matriz de confusión del Modelo 3.3 (antes de *Naive Bayes* y despúes).

A continuación, se implementa la segunda red bayesiana mediante la librería pyAgrum con las relaciones obtenidas de HNet con el fin de comparar las predicciones frente a *Naive Bayes* (Figura 5.5). Se observa que las matrices de confusión siguen siendo mejor que las del clasificador de imágenes aunque son peores que *Naive Bayes*. Los resultados obtenidos para cada modelo son los siguientes:

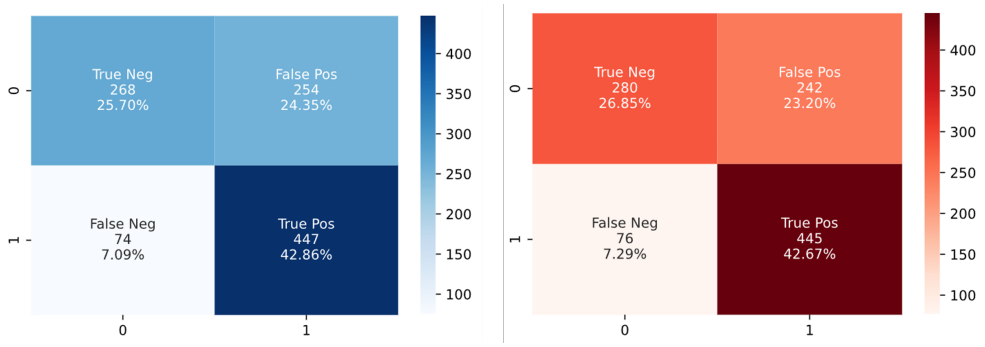

Figura 5.28: Matriz de confusión del Modelo 3.1 (antes de la red bayesiana HNet y despúes).

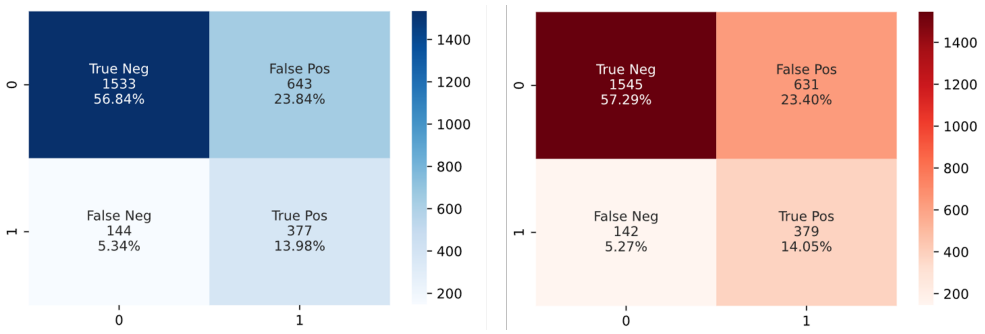

Figura 5.29: Matriz de confusión del Modelo 3.2 (antes de la red bayesiana HNet y despúes).

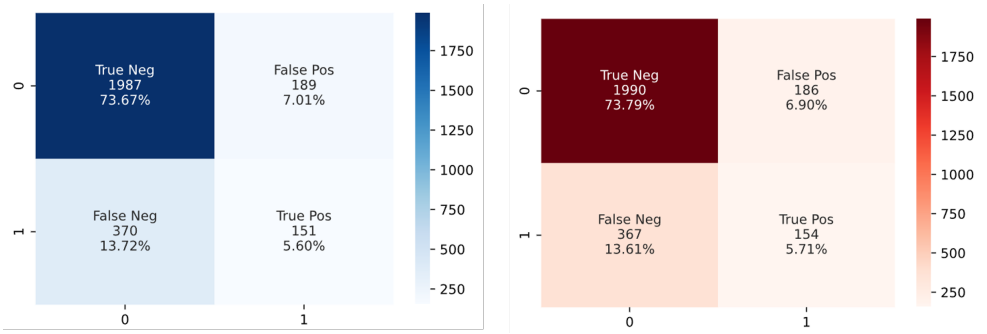

Figura 5.30: Matriz de confusión del Modelo 3.3 (antes de la red bayesiana HNet y despúes).

Modelo 4 (función de pérdida ponderada). Este modelo utiliza únicamente las imágenes de 2020. Se obtiene una curva ROC - AUC en el conjunto de test de 0,73. Al observar la matriz de confusión, tiene un buen nivel de clasificación malignos aunque al darle menos peso a los benignos no es capaz de clasificar estas muy bien y tiende a sobrestimar el diagnóstico. Este modelo es similar al modelo 2 con los datos de 2019.

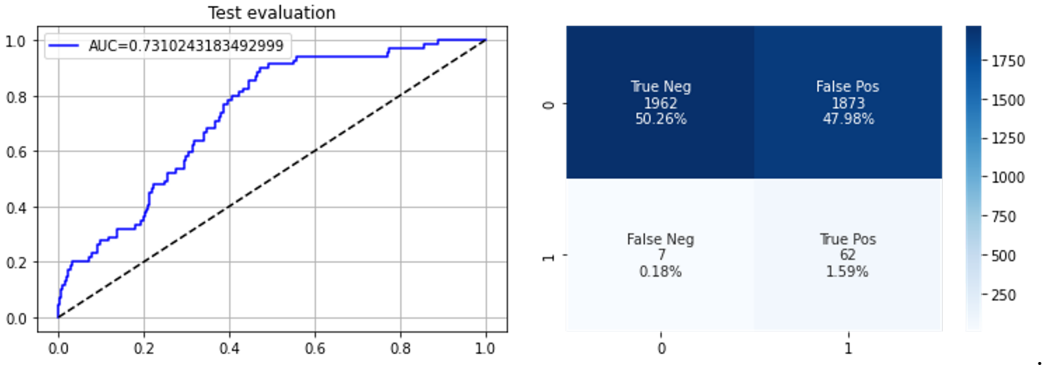

Figura 5.31: Métricas de evaluación del Modelo 5.

Finalmente, en la Tabla 5.1 se muestra un resumen de la curva ROC - AUC y la matriz de confusión de cada modelo implementado.

Modelo (Datos del año)	Test AUC - ROC	Matriz de confusión
Modelo base (2019)	0,75	$\begin{bmatrix} 3.345 & 268 \\ 536 & 331 \end{bmatrix}$
Modelo submuestreo (2019)	0,78	$\begin{bmatrix} 268 & 254 \\ 74 & 447 \end{bmatrix}$
Modelo con pérdida ponderada (2019)	0,79	$\begin{bmatrix} 1.533 & 643 \\ 144 & 377 \end{bmatrix}$
Modelo sobremuestreo (2019)	0,63	$\begin{bmatrix} 1.987 & 189 \\ 370 & 151 \end{bmatrix}$
Modelo submuestreo [*Naive Bayes*] (2019)	0,83	$\begin{bmatrix} 294 & 228 \\ 63 & 458 \end{bmatrix}$
Modelo con pérdida ponderada [*Naive Bayes*] (2019)	0,80	$\begin{bmatrix} 1.558 & 618 \\ 141 & 380 \end{bmatrix}$
Modelo sobremuestreo [*Naive Bayes*] (2019)	0,71	$\begin{bmatrix} 1.982 & 194 \\ 356 & 165 \end{bmatrix}$
Modelo submuestreo [Red bayesiana HNet] (2019)	0,81	$\begin{bmatrix} 280 & 242 \\ 76 & 445 \end{bmatrix}$
Modelo con pérdida ponderada [Red bayesiana HNet] (2019)	0,79	$\begin{bmatrix} 1.545 & 631 \\ 142 & 379 \end{bmatrix}$
Modelo sobremuestreo [Red bayesiana HNet] (2019)	0,70	$\begin{bmatrix} 1.990 & 186 \\ 367 & 154 \end{bmatrix}$
Modelo con pérdida ponderada (2020)	0,73	$\begin{bmatrix} 1.962 & 1.873 \\ 7 & 62 \end{bmatrix}$

Tabla 5.1: Comparación de modelos junto a su matriz de confusión.

6

Conclusión

6.1 Conclusiones

Como se puede comprobar en este trabajo, implementar una red bayesiana a partir de los datos tabulares ayuda y mejora la clasificación frente a únicamente tener en cuenta la imagen de la lesión. Además se expone la importancia de buscar relaciones entre las variables de los datos tabulares que puede ayudar a diseñar una red bayesiana más que se ajuste a los datos. Aunque en este caso se da que *Naive Bayes* resulta obtener mejores resultados frente a la red bayesiana obtenida por HNet.
Por lo que un diagnóstico con la ayuda de las redes neuronales convolucionales junto a una red bayesiana puede ser beneficiosa para sus usuarios y más acertada, mejorando la precisión diagnóstica de los especialistas en la detección de melanoma.

6.2 Trabajo futuro

Respecto a llevar este proceso a la práctica a una clínica, esta puede utilizar el Modelo 4 y a partir de sus datos tabulares crear una red bayesiana para incorporar los factores y obtener nuevas predicciones. Otra opción interesante podría ser utilizar un modelo que ha sido entrenado con el conjunto de imágenes de 2019, al cual se puede hacer *fine-tuning* con el conjunto de imágenes propias del hospital e introducir los datos tabulares o evidencias de los pacientes mediante la red bayesiana la cual puede ir creciendo a medida que se obtenga más información y nuevas características del paciente como la región en la que vive (la prevalencia de una población varía de un lugar a otro), si ha tenido antecedentes familiares de melanoma u otros factores que influyen en el cáncer de piel como mencionado en la Tabla 1.1.
A medida que se obtiene más evidencias, se actualizará la probabilidad, ya que Bayes no está pensado para usarse solo una vez, sino para utilizarse múltiples veces obteniendo más evidencias y revisando la probabilidad de que algo sea cierto.
Otra línea de trabajo sería utilizar datos e imágenes del conjunto de datos de 2017 y 2018 para disponer de suficientes datos para una correcta calibración de los modelos.
Finalmente, a corto plazo sería atrayente observar y revisar los resultados obtenidos a partir del modelo ganador de la competición de Kaggle.

Referencias

[1] Equipo de redactores y equipo de editores médicos de la Sociedad Americana Contra El Cáncer. **¿Qué son los cánceres de piel de células basales y de células escamosas?** 2019. iii, 1

[2] Alberto Armas Navarro. **Servicio Canario de la Salud**. 2011. iii, 3

[3] Malak A. Abdullah. **ResearchGate**. 2020. iii, 14

[4] Applied Deep Learning with Keras. **Packt**. iii, 16

[5] Sarang Narkhede. **Understanding AUC - ROC Curve**. 2018. iii, 17

[6] Yu Sun Kilian Q. Weinberger Chuan Guo, Geoff Pleiss. **On Calibration of Modern Neural Networks**. 2017. iii, 4, 18, 19, 20

[7] Moises Goldszmidt Nir Friedman, Dan Geiger. **Bayesian Network Classifiers**. 1997. iii, 4, 22

[8] G.M. Garnacho Saucedo. **Trastornos de la pigmentación: lentigos, nevus y melanoma**. v, 2

[9] Zoe Kleinman. **Google AI tool can help patients identify skin conditions**. 4

[10] Instituto de Cáncer del Centro Médico de la Universidad de Tennessee. **Melanoma y cáncer de tejidos blandos**. 5

[11] ISIC 2020 Challenge Dataset. **Official dataset of the SIIM-ISIC Melanoma Classification Challenge**. 2020. 7

[12] SIIM & ISIC. **SIIM-ISIC Melanoma Classification Kaggle Challenge**. 2020. 7

[13] SIIM & ISIC. **ISIC Challenge Datasets**. 2020. 7

[14] Jason Brownlee. **Imbalanced Classification with Python**. 2020. 12, 13

[15] Ishan Shrivastava. **Handling Class Imbalance by Introducing Sample Weighting in the Loss Function**. 2020. 14

[16] Brian Lucena. **Probability Calibration Workshop**. 2020. 18

[17] Markus Kängsepp. **Calibration of Convolutional Neural Networks**. 2018. 19

[18] Brian Lucena. **Spline-Based Probability Calibration**. 2018. 19

[19] Scikit-Learn Developers. **Sklearn Library, Brier Score**. 2020. 20

[20] Scikit-Learn Developers. **Sklearn Library, Log Loss**. 2020. 20

[21] Keras. **Keras DenseNet Documentation**. 2021. 21

[22] Ashadullah Shawon Yanmei Huang Najmul Hasan, Yukun Bao. **DenseNet Convolutional Neural Networks Application for Predicting COVID-19 Using CT Image**. 2021. 21

[23] haqishen. **SIIM-ISIC Melanoma Classification 1st Place Solution**. 2020. 21

[24] Veritasium. **Explore and understand your data with a network of significant associations.** 2019. 22

[25] Jason Brownlee. **A Gentle Introduction to Bayes Theorem for Machine Learning.** 2019. 23

[26] Erdogan Taskesen. **Explore and understand your data with a network of significant associations.** 2021. 24, 25

[27] pyAgrum Team. **pyAgrum Tutorial.** 2018-21. 25