



UNIVERSIDAD NACIONAL DE COLOMBIA

Análisis de Calibración en Modelos de Aprendizaje de Máquina Cuántico

Glenn Harry Amaya Cruz

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2023

Análisis de Calibración en Modelos de Aprendizaje de Máquina Cuántico

Glenn Harry Amaya Cruz

Tesis o trabajo de grado presentado como requerimiento parcial para optar por el título de:
Magister en Ingeniería de Sistemas y Computación

Director de tesis:

Fabio Augusto González Osorio, Ph. D.

Co-director de Tesis:

Santiago Toledo Cortés, MSc. Ph. D. (c)

Áreas de investigación:

Sistemas inteligentes, computación cuántica y aprendizaje de máquina

Grupo de investigación:

MindLab - Machine Learning, Perception and Discovery Lab

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2023

Para mi madre Ana & hermano David

Cada vez que cometo un error me parece descubrir una verdad que no conocía.

Maurice Maeterlinck

Agradecimientos

Quiero agradecer a la Universidad Nacional de Colombia por brindarme durante varios años la oportunidad de crecer profesional, académica y personalmente influyendo en mi significativamente. A mi director Fabio González por su guía a lo largo de este proceso y a Santiago Toledo Cortés, por sus continuos comentarios, apoyo y guía que fueron de gran utilidad para culminar este proceso de forma exitosa. Finalmente, a mi madre Ana y hermano David por su comprensión en todo momento.

Resumen

Análisis de Calibración en Modelos de Aprendizaje de Máquina Cuántico

El análisis de calibración de modelos de aprendizaje de máquina cobra gran importancia en distintos contextos como evaluación del riesgo, diagnósticos y sistemas críticos para la seguridad, donde hay decisiones influenciadas por las predicciones de los modelos. El área del aprendizaje de máquina cuántico ha recibido una mayor atención en los últimos años, en particular, se han desarrollado modelos que obtienen resultados competitivos en tareas de clasificación y regresión a comparación con métodos ampliamente utilizados. No obstante, las propiedades de este tipo de clasificadores en términos de calibración no han sido exploradas en la literatura. Por esta razón, en el presente trabajo se realiza un estudio de las propiedades de calibración que tienen algunos modelos de aprendizaje de máquina cuántico frente a modelos ampliamente usados en la literatura como máquinas de soporte vectorial, árboles de decisión, regresión logística, entre otros para tareas de clasificación binaria y de múltiples clases. Adicionalmente, se realiza un experimento para explorar el efecto de algunos clasificadores cuánticos en combinación con una red neuronal. Los resultados experimentales muestran que algunos de los clasificadores cuánticos analizados tienen un rendimiento competitivo e incluso mejor en métricas de calibración y las tareas de clasificación.

Palabras claves: aprendizaje de máquina, aprendizaje de máquina cuántico, calibración, análisis de confianza.

Abstract

Calibration Analysis in Quantum Machine Learning Models

Calibration of machine learning models is of great importance in different contexts such as risk assessment, diagnostics, and safety-critical systems, in which decisions are influenced by model predictions. The area of quantum machine learning has received an increased attention in recent years, in particular, models have been developed that obtain competitive results in classification and regression tasks compared to widely used methods. However, the properties of this type of classifiers in terms of calibration have not been explored in the literature. As a result, in this work a study of the properties of calibration is conducted for recent quantum machine learning models in comparison to state-of-the-art models such as support vector machines, decisions trees, logistic regression, and others for binary and multiclass classification tasks. Moreover, an experiment to explore the effect of some quantum classifiers in combination with a neural network is made. The experimental results show that some of the analyzed quantum classifiers have competitive and even better performance in calibration metrics and the classification tasks.

Keywords: machine learning, quantum machine learning, calibration, confidence analysis.

Este Trabajo Final de maestría fue calificado en marzo de 2023 por el siguiente evaluador:

Germán Jairo Hernández Pérez PhD
Profesor Facultad de Ingeniería
Universidad Nacional de Colombia

Content

Agradecimientos	vii
Resumen	ix
1. Introducción	1
1.1. Planteamiento del problema	2
1.2. Objetivos	3
1.2.1. Objetivos generales	3
1.2.2. Objetivos específicos	3
1.3. Contribuciones principales	3
1.4. Estructura de la tesis	4
2. Antecedentes	5
2.1. Marco Teórico	5
2.1.1. Definición de calibración	5
2.1.2. Métricas de calibración	8
2.1.3. Métodos de calibración	11
2.1.4. Aprendizaje de máquina cuántico	13
2.2. Estudios comparativos	13
2.3. Importancia de la calibración de modelos	15
2.4. Revisión de literatura	16
3. Quantum Measurement Classification	19
3.1. Quantum Measurement Classification	19
3.1.1. Mapeo de características cuánticas	19
3.1.2. Matrices de densidad	20
3.1.3. Medición cuántica	21
3.1.4. Predicción	21
3.2. Kernel Quantum Measurement	22
4. Conjuntos de datos	23
5. Análisis de calibración	25
5.1. Modelos de aprendizaje de máquina	25

5.2. Configuración experimental	25
5.2.1. Preprocesamiento	25
5.2.2. Búsqueda de hiperparámetros	26
5.2.3. Métricas de calibración y desempeño	29
5.3. Resultados	30
5.3.1. Clasificación binaria	30
5.3.2. Clasificación de múltiples clases	37
5.3.3. LeNet5-QMDM	41
5.4. Discusión	44
6. Conclusiones y trabajo futuro	46
A. Apéndice: Curvas de aprendizaje	48
Referencias	49

1. Introducción

El interés en métodos de aprendizaje de máquina cuántico ha aumentado en años recientes. Muestra de ello es la creciente investigación que se ha llevado a cabo en los últimos años debido a su potencial de sobrepasar métodos de aprendizaje de máquina tradicionales en términos de velocidad de procesamiento, métricas de rendimiento y otras propiedades [7, 51].

Principalmente, existen dos enfoques relacionados al aprendizaje de máquina cuántico. El primero se enfoca en la habilidad de aplicar algoritmos tradicionales de aprendizaje de máquina en computadores cuánticos con el fin de obtener una ejecución más eficiente [51]. El segundo enfoque se relaciona con la construcción de nuevos algoritmos de aprendizaje de máquina inspirados en conceptos de física cuántica, esto para aprovechar propiedades probabilísticas y desarrollar nuevos mecanismos de representación y procesamiento de la información [50].

En particular, los métodos propuestos en “*Classification with quantum measurements*” y “*Learning with density matrices and random features*” [28, 21] se enfocan en aprovechar conceptos de física cuántica, tales como la proyección cuántica y matrices de densidad, para construir algoritmos de aprendizaje de máquina que puedan ser utilizados en tareas supervisadas y no supervisadas. Por ejemplo, los autores de [52, 14] demostraron que los modelos cuánticos DMKDC y DQMOR son capaces de evidenciar mejores resultados en algunas métricas de rendimiento cuando se comparan con modelos de referencia en contextos como clasificación de imágenes médicas o en detección de anomalías [22].

Sin embargo, en aplicaciones de sistemas críticos para la seguridad, escenarios donde los costos de clasificación errónea son importantes y, en general, donde se toman decisiones que pueden tener implicaciones en la vida de los seres vivos, es relevante medir otros aspectos de los modelos como la incertidumbre y la confianza en torno a las predicciones [57]. Surge así el concepto de calibración de modelos estadísticos y de aprendizaje automático, que consiste en determinar la capacidad de los métodos para generar predicciones que puedan interpretarse como las probabilidades de pertenencia a las categorías a predecir [24, 16, 36, 35]. Así, las salidas de los clasificadores no solo ofrecen una herramienta para la clasificación, sino que también nos permiten conocer la confianza en sus predicciones.

En otras palabras, la calibración hace referencia al hecho que las probabilidades estimadas reflejan el comportamiento de algún evento en el estado real de la naturaleza [35]. Por ejem-

plo, si un modelo predice que lloverá con una probabilidad de 0.7 en 10 días diferentes, el clasificador se dice que está calibrado si llueve 7 de los días 10. En otros casos, el modelo predice el fenómeno con una mayor o menor confianza a comparación de la realidad [31].

Muchos de los modelos ampliamente empleados en aprendizaje de máquina tales como *naive Bayes*, árboles de decisión, máquinas de soporte vectorial, entre otros, no brindan probabilidades de clasificación calibradas [43, 32]. Esto puede conducir a decisiones incorrectas en áreas relacionadas a medicina o la salud [32, 10] generando un gran costo por dichas decisiones [57]. Estos problemas se pueden presentar debido a sobreajuste durante el entrenamiento, uso de estimadores no probabilísticos o características particulares de la función objetivo [16].

Adicionalmente, las probabilidades de un modelo de clasificación son relevantes en contextos donde se desea tener una idea sobre la incertidumbre de las predicciones. Por ejemplo, en el caso de tener un modelo no calibrado en medicina personalizada en vez de estimaciones de riesgo, las decisiones médicas basadas en los resultados podrían ser incorrectas [32, 10].

En la literatura no se encuentran estudios sobre la calibración de modelos de aprendizaje de máquina cuántico, lo cual puede deberse a que este tipo de modelos se han desarrollado recientemente y, en consecuencia, aún hay avances por alcanzar en esta área.

El propósito de este trabajo es el estudio de las propiedades de calibración de los modelos de aprendizaje de máquina cuántico mencionados en [28, 21], y su rendimiento contra otros métodos no cuánticos para diversas tareas de clasificación.

1.1. Planteamiento del problema

El interés en los métodos de aprendizaje de máquina cuántico ha aumentado en años recientes. Prueba de ello es la creciente investigación que se ha realizado en el área debido al potencial que tienen estos métodos para sobrepasar los métodos clásicos de aprendizaje de máquina en términos de velocidad de procesamiento e incluso exactitud de los modelos [7, 51].

En particular, los métodos propuestos en “*Classification with quantum measurements*” y “*Learning with density matrices and random features*” han mostrado mejores resultados en algunas métricas de rendimiento cuando se comparan con modelos del estado del arte para contextos como clasificación de imágenes médicas [52].

No obstante, en aplicaciones de sistemas críticos para la seguridad, escenarios donde los costos de clasificación errónea son importantes, y en general, donde se toman decisiones que pueden tener implicaciones en la vida de los seres vivos, es relevante medir otros tipos de métricas tales como la incertidumbre alrededor de las predicciones [57]. Surge así el concepto

de calibración de modelos, el cual consiste en determinar la habilidad de los métodos para generar predicciones que pueden ser interpretadas como las probabilidades correctas de pertenencia a las categorías a predecir [24, 16, 36, 35].

Para nuestro conocimiento, no hay estudios en la literatura sobre las características que los modelos cuánticos tienen en torno a la calibración, lo cual puede darse porque este tipo de modelos han sido desarrollados recientemente y, en consecuencia, es un área de creciente investigación.

Así, este trabajo busca responder la siguiente pregunta de investigación:

¿Cuáles son las características de los métodos de aprendizaje de máquina cuántico de interés en cuanto a la calibración de las probabilidades de predicción en diversas tareas de clasificación?

1.2. Objetivos

1.2.1. Objetivos generales

- Evaluar las propiedades de calibración en modelos de aprendizaje de máquina cuántico en distintas tareas de clasificación.

1.2.2. Objetivos específicos

- Establecer los conjuntos de datos y las métricas adecuadas para medir el nivel de calibración de los modelos a evaluar.
- Analizar las propiedades de los métodos de aprendizaje de máquina cuántico seleccionados en cuanto a calibración.
- Comparar el nivel de calibración de los métodos de aprendizaje de máquina cuántico y no cuánticos de interés para diversas tareas de clasificación

1.3. Contribuciones principales

Este trabajo presenta un análisis exploratorio y cuantitativo de la calibración de algunos modelos de aprendizaje de máquina cuántico, y su rendimiento en comparación a métodos tradicionales como *naive Bayes*, máquinas de soporte vectorial, regresión logística, árboles de decisión, árboles aleatorios, k vecinos más cercanos, y perceptrones multicapa. Para nuestro conocimiento, no hay investigaciones que cubran la calibración de modelos de aprendizaje de máquina cuántico. Esta investigación contribuye para identificar las ventajas o desventajas

que los modelos de aprendizaje de máquina cuántico tienen para diversas tareas de clasificación, pero en particular, en aplicaciones de sistemas críticos para la seguridad.

Así, obtenemos que los métodos cuánticos tienden a ser competitivos, tanto en exactitud como en las métricas de calibración, a comparación de los clasificadores de referencia en algunos casos. En especial, los clasificadores cuánticos tienden a tener mejores propiedades de calibración cuando el rendimiento en la tarea de clasificación es mejor. No obstante, hay clasificadores cuánticos que muestran un desempeño consistentemente peor en comparación de otros modelos.

1.4. Estructura de la tesis

El resto de este documento está organizado de la siguiente manera: el segundo capítulo contiene una revisión bibliográfica sobre los principales conceptos relacionados con el problema de interés. El tercer capítulo explica los clasificadores cuánticos explorados. El cuarto capítulo describe los conjuntos de datos utilizados en la experimentación. El quinto capítulo ilustra la configuración experimental y los principales resultados del proceso de experimentación. En el último capítulo se presentan las conclusiones y el trabajo futuro.

2. Antecedentes

2.1. Marco Teórico

En el presente capítulo resumimos las principales definiciones relacionadas al entendimiento de la calibración desde una perspectiva teórica en modelos de aprendizaje automático y estadístico usando ejemplos prácticos. Adicionalmente, resaltamos la importancia de la calibración en tareas de clasificación, y finalmente resumimos todos los trabajos relacionados.

2.1.1. Definición de calibración

Existen dos definiciones que hacen referencia a la calibración de modelos estadísticos y de aprendizaje de máquina. La primera definición, y la más común que se encuentra en la literatura sobre este concepto, hace referencia al entrenamiento de modelos para encontrar parámetros desconocidos o describir el comportamiento de un fenómeno natural [9, 26, 53].

Por ejemplo, existen métodos para determinar la relación entre diferentes variables en áreas como rendimiento de cultivos y análisis climático [46], hidrología [3], estudio de especies [47, 4, 9, 26], simulación de fenómenos naturales en geología [29] y medicina [17].

En segundo lugar, la calibración de un modelo estadístico o de aprendizaje automático usado para clasificación también se refiere al hecho que las probabilidades estimadas reflejen el comportamiento de algún evento en el estado real de la naturaleza [35], y en caso contrario, el modelo podría predecir la ocurrencia de un evento con mayor o menos confianza en comparación con la realidad [31].

Esta definición de calibración también se conoce como fiabilidad o validez, términos más utilizados en meteorología [5, 11, 19, 6]. Otros autores se refieren a este concepto como calibración de confianza o simplemente confianza de los modelos [10, 24].

El presente trabajo se centra en la segunda definición de calibración mencionada, ya que se desean analizar las propiedades de las probabilidades obtenidas a partir de los métodos de aprendizaje cuántico de interés frente a otros modelos, y las propiedades de calibración que

podrían tener dichos clasificadores.

Formalmente, en la literatura se encuentran tres definiciones que reflejan distintos niveles de calibración. En general, un estimador o modelo está perfectamente calibrado, calibrado marginalmente o por clase, o calibrado con confianza [38, 54].

Definición 2.1.1 (Calibración perfecta) *Un clasificador probabilístico de múltiples clases $\mathbf{g} : \mathbf{X} \rightarrow \Delta_k$ está perfectamente calibrado, o simplemente calibrado, si dado cualquier vector de predicciones $\mathbf{q} = (q_1, \dots, q_K) \in \Delta_k$, la proporción de las clases $\tilde{\mathbf{q}}$, y cualquier muestra \mathbf{x} que tiene el mismo vector de predicciones, entonces el vector de predicciones \mathbf{q} y el vector de proporciones $\tilde{\mathbf{q}}$ son iguales. Esto es*

$$\mathbb{P}(Y = i | \mathbf{g}(\mathbf{X}) = \mathbf{q}) = \tilde{q}_i, \quad \forall i \in 1, \dots, K. \quad (2-1)$$

donde $\Delta_k = \{(q_1, \dots, q_K) \in [0, 1]^k; \sum_{i=1}^K q_i = 1\}$, \mathbf{X} es un espacio de características, y Y es el espacio de la variable objetivo [38, 54, 24].

Definición 2.1.2 (Calibración por clases) *Un clasificador probabilístico de múltiples clases $\mathbf{g} : \mathbf{X} \rightarrow \Delta_k$ está calibrado por clase, o marginalmente calibrado, si para alguna clase i y probabilidad predicha q_i [38, 54] se tiene*

$$\mathbb{P}(Y = i | \mathbf{g}(\mathbf{X}) = q_i) = \tilde{q}_i, \quad i \in 1, \dots, K. \quad (2-2)$$

Definición 2.1.3 (Calibración con confianza) *Un clasificador probabilístico de múltiples clases $\mathbf{g} : \mathbf{X} \rightarrow \Delta_k$ está calibrado con confianza, si para cualquier constante $c \in [0, 1]$ [38, 54] se cumple*

$$\mathbb{P}(Y = \operatorname{argmax}(g(\mathbf{X})) | \max(g(\mathbf{X})) = c) = c. \quad (2-3)$$

Un ejemplo práctico del concepto de calibración se observa en la Figura 2-1. Imaginemos que observamos las características de una población de 10 pacientes donde 20% no tienen ninguna enfermedad, 30% sufre una enfermedad de los pulmones, y 50% tienen una enfermedad del corazón. El objetivo es entrenar un clasificador que prediga correctamente cual condición de las tres mencionadas una persona tiene. De acuerdo con las definiciones de calibración dadas, \mathbf{x} necesita generar el mismo vector de probabilidades \mathbf{q} , por lo cual, en el ejemplo suponemos que todos los individuos observados tienen características similares.

Para cumplir esa tarea, cuatro clasificadores diferentes, $g_1(\cdot)$, $g_2(\cdot)$, $g_3(\cdot)$ y $g_4(\cdot)$, son entrenados. Una vez entrenados, el primer clasificador genera un vector de probabilidades igual

a $(0.2, 0.3, 0.5)$ donde 0.2 representa la probabilidad de pertenecer a la clase de personas que no tienen ninguna enfermedad, 0.3 a las que tienen alguna enfermedad en los pulmones, y 0.5 los que tienen una del corazón. Si se predice que las personas pertenecen a la clase donde el clasificador arroja la probabilidad más grande, $g_1(\cdot)$ obtendría una exactitud de 0.5, pero dado que el vector de probabilidades predichas es igual a la distribución real de las condiciones no observados de la población, se dice que el clasificador $g_1(\cdot)$ está perfectamente calibrado.

En el segundo caso, $g_2(\cdot)$ genera un vector de probabilidades igual a $(0.2, 0.5, 0.3)$ que refleja una exactitud de 0.3, dado que predice que las personas tienen una enfermedad en el pulmón, pero solo 30% de las personas la tienen en realidad. Por otro lado, la probabilidad que las personas no tengan ninguna condición es de 0.2 que corresponde a la distribución real de la población, por lo tanto, se dice que $g_2(\cdot)$ es calibrado por clase porque para al menos una de las clases su probabilidad es igual a la realidad no observada.

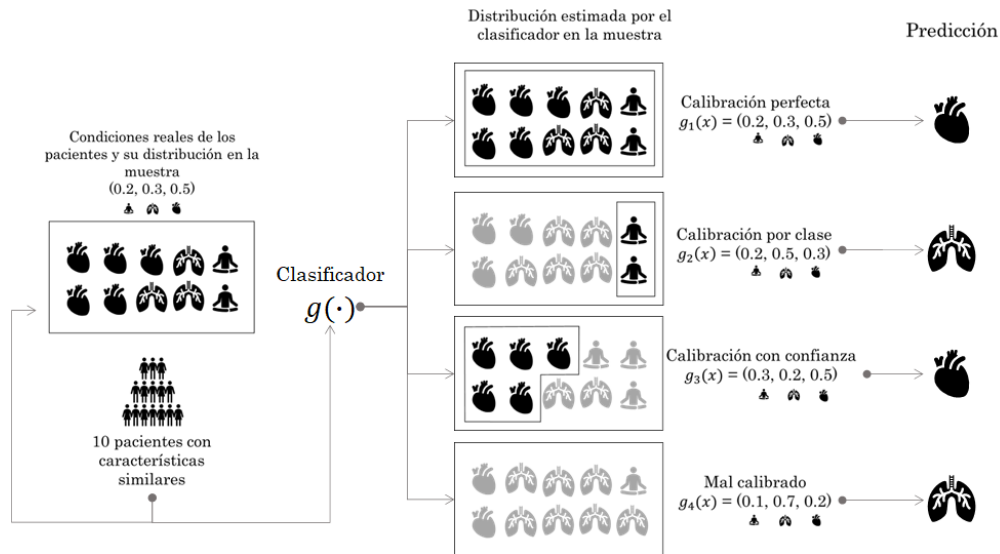


Figure 2-1.: Ejemplo de las distintas definiciones de calibración para cuatro modelos que tienen como objetivo clasificar si una persona es sana o tiene alguna enfermedad del corazón o del pulmón

Por otro lado, el clasificador $g_3(\cdot)$ tiene un vector de probabilidades igual a $(0.3, 0.2, 0.5)$ con una exactitud de 0.5 debido a las personas correctamente clasificadas en las que tienen alguna enfermedad del corazón. En contraste a $g_1(\cdot)$, $g_3(\cdot)$ no está perfectamente calibrado debido a que las probabilidades para las personas sanas y con alguna enfermedad del pulmón no son iguales a la distribución real. Sin embargo, la clase con la probabilidad más alta sería la de personas con alguna enfermedad del corazón, y dado que esta clase tiene la misma distribución que la realidad, se dice que el clasificador $g_3(\cdot)$ es calibrado con confianza.

Por último, podemos observar que el clasificador $g_4(\cdot)$ tiene una exactitud de 0.3 porque predice que las personas tienen una enfermedad pulmonar debido a que el vector de probabilidades (0.1, 0.7, 0.2), y cualquier distribución real de las clases no coincide con las probabilidades dadas por el clasificador, por lo tanto, se dice que $g_4(\cdot)$ no está calibrado. Este escenario puede ocurrir cuando los clasificadores entrenados tienden a estimar probabilidades cercanas a 1, y éstas no necesariamente reflejan la incertidumbre que el modelo podría tener en torno a los valores objetivo reales.

En el ejemplo anterior, hay un enfoque en el hecho de que los modelos reflejan la incertidumbre de las predicciones lo cual no necesariamente corresponde con una alta exactitud del clasificador. Por lo tanto, uno podría tener modelos que discriminen con mucha precisión los valores objetivo, pero los valores generados para clasificar las observaciones no pueden interpretarse correctamente como probabilidades de pertenencia. Por ejemplo, en [33] estudian la estimación de la incertidumbre de las predicciones a través de probabilidades calibradas en el aprendizaje en línea y en [1] se propone un método para obtener estimaciones fiables de la incertidumbre de las predicciones en imágenes de fondo de ojo.

2.1.2. Métricas de calibración

En general, las métricas utilizadas para medir y analizar el nivel de calibración de los modelos no son triviales. En el ejemplo del apartado anterior suponemos que tenemos la distribución real de la variable objetivo, pero en la práctica no tenemos acceso a esa información. Por ello, en la literatura existen diferentes propuestas de métricas de calibración, pero por el momento no ha habido consenso en la literatura científica sobre un marco para medir esta propiedad de los clasificadores [42, 38, 27, 25, 37].

Para comprender las diferentes métricas de calibración, definimos el error de calibración, que refleja qué tan lejos están las probabilidades generadas por un clasificador de la distribución esperada.

Definición 2.1.4 (Error de calibración) *Dado un clasificador probabilístico de múltiples clases $g : \mathbf{X} \rightarrow \Delta_k$ el error de calibración es dado por*

$$CE(g) = \left(\mathbb{E} [|g(\mathbf{X}) - \mathbb{E}(Y|g(\mathbf{X}))|^2] \right)^{1/2}, \quad (2-4)$$

donde Δ_k se define como en la ecuación 2-1 [54].

Como podemos apreciar la definición del error de calibración y la definición de modelo perfectamente calibrado (ecuación 2-1) son similares, donde $\mathbb{P}(Y = i | g(\mathbf{X}) = \mathbf{q})$ se reemplaza por $\mathbb{E}(Y | g(\mathbf{X}))$. Por lo tanto, si $CE(g) = 0$, entonces el clasificador g está perfectamente

calibrado, y si cambiamos el exponencial en la definición por p , se obtiene el ℓ_p error de calibración [54].

Los valores esperados en 2-4 son teóricos, por lo que se requiere una estimación de ellos. Como resultado aparece el error esperado de calibración (ECE por sus siglas en inglés), el cual se define en la siguiente expresión.

Definición 2.1.5 (Error esperado de calibración) *Dado un clasificador probabilístico de múltiples clases $\mathbf{g} : \mathbf{X} \rightarrow \Delta_k$ el error esperado de calibración es dado por*

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|, \quad (2-5)$$

donde n es el número de muestras, B_m es el conjunto de índices de las muestras cuyas predicciones caen en el intervalo $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$, $m \in 1, \dots, M$, M es el número de grupos que se construyen con las probabilidades predichas por $\mathbf{g}(\cdot)$, $|B_m|$ es la cardinalidad de B_m , $acc(B_m)$ es la exactitud promedio en B_m y $conf(B_m)$ es la confianza promedio en B_m [24].

A pesar de ser una de las métricas más utilizadas, el ECE tiene algunas fallas mencionadas en [54, 42, 37] que pueden causar una subestimación del error de calibración, algunas de las cuales están relacionadas con la construcción de las particiones B_m . Por esta razón, se sugiere usar remuestreo para generar una distribución de las estimaciones, usar pruebas de hipótesis basadas en una distribución empírica que suponga una calibración perfecta y calcular particiones dependientes de los datos en lugar de uniformes para superar estos problemas [54].

Adicionalmente, el ECE no tiene en cuenta el error de calibración de cada valor objetivo, lo que podría ocultar algunas propiedades importantes de los clasificadores para algunas clases. Para resolver este inconveniente y basado en 2-2, el cwECE se define como:

Definición 2.1.6 (Error de calibración esperado por clase) *Dado un clasificador probabilístico de múltiples clases $\mathbf{g} : \mathbf{X} \rightarrow \Delta_k$ el error esperado de calibración por clase es dado por*

$$cwECE = \frac{1}{K} \sum_{k=1}^K \sum_{m=1}^M \frac{|B_{m,k}|}{n} |acc(B_{m,k}) - conf(B_{m,k})|, \quad (2-6)$$

donde n es el número de muestras, $B_{m,k}$ es el conjunto de índices de las muestras cuyas predicciones caen en el intervalo $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$ para la clase k , $m \in 1, \dots, M$, M es el número de particiones, $|B_{m,k}|$ es la cardinalidad de $B_{m,k}$, $acc(B_{m,k})$ es la exactitud promedio

en $B_{m,k}$ para la clase k , y $\text{conf}(B_{m,k})$ es la confianza promedio en $B_{m,k}$ para la clase k [38]. Note que los intervalos I_m pueden no ser iguales para todas las clases, pero por simplicidad se toman iguales.

Otra métrica de calibración utilizada en la literatura es el error cuadrático medio de las probabilidades predichas y los valores reales, llamada *Brier score* [5], la cual se desarrolla en meteorología, pero puede ser aplicada a cualquier tarea de clasificación

Las métricas para cuantificar el error de calibración son importantes para tener una herramienta objetiva. Sin embargo, el diagrama de confiabilidad es una herramienta visual relevante en el estudio de calibración para cualquier tipo de clasificador. Algunas de sus ventajas se discuten en [6], así como su construcción. Este gráfico compara la exactitud y la confianza promedio en las particiones o particiones establecidas como en la ecuación (2-5), luego, cuanto más cerca estén los valores de la diagonal, mejor será la calibración del modelo.

Para ilustrar la construcción del diagrama de confiabilidad, podemos imaginar que tenemos un clasificador binario $g_b(\cdot)$ que fue entrenado para predecir si una persona tiene una enfermedad cardíaca o si está sana, y una muestra de 100 pacientes de un hospital y su estado (enfermo o sano). Entonces, en base a esta muestra queremos saber si el clasificador está calibrado o no usando el diagrama de confiabilidad. En primer lugar, categorizamos el espacio de las probabilidades pronosticadas en particiones, $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$ y $[0.8, 1]$, que es equivalente a B_m o $B_{m,k}$ en las definiciones 2-2 y 2-5. De todos los individuos analizados, 10 de ellos tienen una probabilidad de tener la enfermedad que se encuentran en el intervalo $[0.4, 0.6)$ según el clasificador (Figura 2-2).

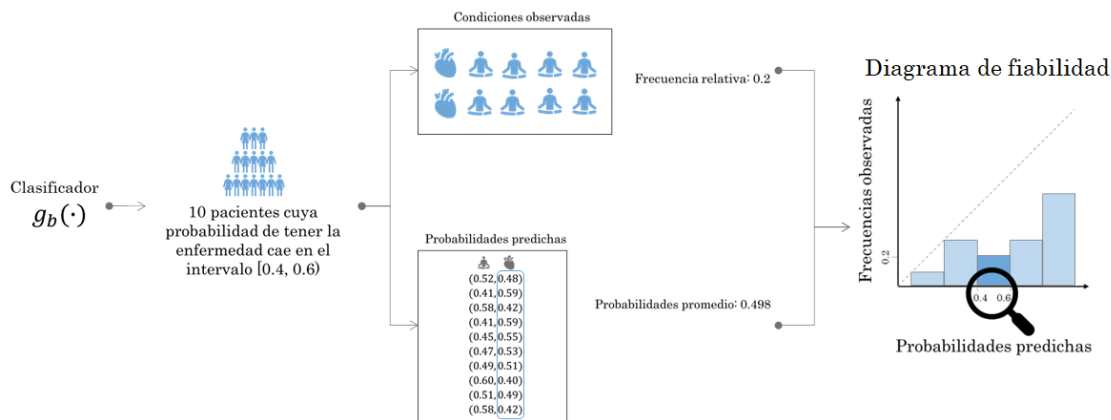


Figure 2-2.: Ilustración de la construcción del diagrama de fiabilidad en una tarea de clasificación binaria que busca identificar personas enfermas y sanas

Una vez identificados esos 10 pacientes, sus probabilidades de estar enfermos y la predicción del modelo, observamos que 2 de ellos tienen una cardiopatía, entonces la frecuencia obser-

vada es igual a 0.2 (2/10), y por otro lado, la probabilidad promedio de estar enfermo de los 10 individuos es igual a 0.498. Luego, graficamos en el eje x las probabilidades promedio y en el eje y las frecuencias observadas.

Si las barras tienden a estar sobre la diagonal, se dice que el clasificador tiene demasiada confianza en sus predicciones, es decir, el modelo genera probabilidades que muestran una confianza errónea en las predicciones que hace. Por el contrario, si tiende a generar probabilidades por debajo de la diagonal, se dice que el clasificador está por debajo de la confianza, como es el caso de este ejemplo, y finalmente, está calibrado si las barras están cerca de la diagonal.

En el ejemplo anterior, tomamos particiones igualmente espaciadas, sin embargo, la selección del esquema y número de particiones pueden hacer que se subestime el error de calibración [37], lo que hará que el diagrama de calibración también se vea afectado. Por esta razón, existen varios esquemas de particiones para evitar la subestimación del error de calibración que se utilizan para lograr una mejor medición de la calibración. Uno de ellos depende de los datos generando particiones de amplitudes variadas, el cual se ilustra en [54].

Es importante señalar que el diagrama presentado anteriormente se utiliza para un problema de clasificación binaria, por lo tanto, en [54] se propone un diagrama de calibración para un máximo de tres clases. Mientras que en [38] se proponen diagramas de fiabilidad individualmente para cada clase basados en la definición 2-2 donde las frecuencias relativas en la Figura 2-2 se reemplazan por la exactitud en los intervalos y las probabilidades promedio por las probabilidades medias de la clase k . También grafican diagramas de fiabilidad basados en la definición 2-3 de manera similar a los diagramas de clase pero usando el promedio de las probabilidades máximas.

Finalmente, para el conocimiento de los autores, no existe un procedimiento estándar para cuantificar la calibración de los modelos, por lo que algunos autores han propuesto marcos teóricos para realizar esta tarea [54, 55], métricas para la estimación del error de calibración que tratan de resolver algunos de los problemas mencionados anteriormente, y pruebas estadísticas para validar si un modelo está o no calibrado [42, 41, 37, 55, 54, 45]. Por lo tanto, esta es un área de investigación en continuo desarrollo y determinar si un clasificador está calibrado puede convertirse en una tarea desafiante.

2.1.3. Métodos de calibración

En secciones anteriores, la calibración se ha mencionado como una característica inherente de los modelos estadísticos o de aprendizaje automático. Sin embargo, existen herramientas desarrolladas para mejorar la calibración de los estimadores una vez han sido entrenados.

En [56] se comenta una discusión desarrollada por Provost y Domingos (2000), en la que mencionan que suavizar las probabilidades obtenidas por los árboles de decisión es una forma efectiva de mejorar la calibración del modelo; este método de suavizado se llama “*m-estimation*” y fue desarrollado por Cestnik (1990). Asimismo, en [56] se menciona un método llamado “*Curtailment*”, el cual se utiliza para corregir la calibración en árboles de decisión mediante el proceso de evitar ramas con muy pocas muestras de entrenamiento y en consecuencia, se obtienen probabilidades más calibradas.

Las máquinas vectoriales de soporte son un método de discriminación que en principio no genera probabilidades para cada clase objetivo. Sin embargo, en [48] se propone un método para obtener probabilidades a partir de las salidas de cualquier clasificador obtenido, las cuales tienden a ser calibradas [57].

Por otro lado, en [43] se propone un método llamado regresión isotónica, que calibra las predicciones y podría aplicarse a máquinas vectoriales de soporte, *naive Bayes*, y árboles de decisión. Este método es similar al propuesto en [48], pero la regresión isotónica es un enfoque más general ya que solo requiere una función monótona creciente [43]. Así, en [18] se propone un método de calibración aplicado a *random forest* usando las metodologías de [48, 43] en datos de litofacies. Para calibrar redes neuronales en clasificación de imágenes y traducción automática, el suavizado de etiquetas es útil para mejorar la precisión de estos modelos e implícitamente mejora la calibración [40].

En [31] se propone un método para mejorar la calibración de un modelo mediante muestreo de etiquetas en redes neuronales, el cual no afecta sustancialmente la precisión del modelo. Por su parte, en [20] se presenta la estructura de redes neuronales convolucionales para identificar la presencia de meteoritos en el cielo nocturno, obteniendo que luego de realizar un proceso de calibración, se mejora la precisión, cobertura y calibración de los modelos.

Es importante mencionar que algunos de los métodos para calibrar y medir la calibración se aplican en su totalidad para tareas de clasificación binaria, y no donde hay múltiples clases. Por ello, existen metodologías que generalizan el problema de calibración cuando existen varias categorías de respuesta [36], por ejemplo, en reconocimiento óptico de caracteres (OCR por sus siglas en inglés).

Finalmente, la mayoría de las metodologías anteriores se aplican a modelos de aprendizaje automático donde los estimadores se entrenan con una partición de los datos, sin embargo, esto no sucede en el aprendizaje automático en línea donde hay una gran cantidad de información y no es práctico entrenar a los modelos con grandes cantidades de datos o en caso de que el modelo deba actualizarse constantemente. Así, en [33] se presenta un método para

calibrar las probabilidades de modelos de clasificación.

2.1.4. Aprendizaje de máquina cuántico

Para nuestro conocimiento, en la literatura no existen investigaciones realizadas sobre el tema de calibración en métodos de aprendizaje automático cuántico. Esto puede deberse a los recientes avances en esta área aplicados a los modelos de clasificación, lo que significa que aún quedan por realizar más estudios sobre las ventajas o desventajas de los métodos cuánticos en el aprendizaje de máquina.

Sin embargo, se han propuesto métodos de aprendizaje de máquina cuántico que tienen un rendimiento similar o superior en comparación con los modelos convencionales. Un ejemplo de esto se presenta en [23], donde se realiza una generalización de redes tensoriales, que permite manejar estructuras complejas en aprendizaje supervisado y genera medidas de precisión que son competitivas con métodos ampliamente utilizados en aprendizaje automático.

De igual forma, existen aplicaciones de métodos cuánticos en el área de la medicina que obtienen resultados comparables con los modelos de clasificación tradicionales [50, 30], demostrando el potencial que existe en este campo para mejorar los resultados en términos de exactitud y precisión del estado del arte en tareas supervisadas.

Las matrices de densidad combinadas con la medición cuántica se pueden aplicar para resolver problemas de clasificación o regresión, teniendo la posibilidad de ser utilizadas con arquitecturas de redes neuronales, lo que permite aprovechar modelos de extracción de características profundas [28, 21]. Adicionalmente, estos métodos tienen la propiedad de estar contruidos en base a conceptos probabilísticos [28], lo que cobra relevancia en el estudio de la calibración. Además, en [52] se muestran resultados competitivos en comparación con los modelos de aprendizaje profundo en áreas como la medicina donde es relevante tener resultados bien calibrados. Una explicación particular de este tipo de clasificadores se proporciona en una sección posterior.

Finalmente, una perspectiva general sobre este tema se brinda en [51], donde se exponen las principales ideas en torno al aprendizaje de máquina cuántico, algoritmos contruidos, entre otros temas de interés en este campo de investigación.

2.2. Estudios comparativos

Teniendo en cuenta que el presente trabajo busca comparar el rendimiento de diferentes clasificadores cuánticos frente a modelos del estado del arte en aprendizaje automático, resultan de especial interés las investigaciones que se han realizado de tipo similar, en las que

se proponen métodos de calibración [37, 38, 57, 41], y se realiza el análisis de calibración de cierto tipo de modelos [24, 37, 56, 43, 54].

En [24] se lleva a cabo un estudio comparativo de los métodos de calibración en las redes neuronales modernas, que utiliza métricas como el error de calibración esperado (ECE por sus siglas en inglés), el error de calibración máximo (MCE por sus siglas en inglés) y log-verosimilitud negativa para evaluar la calibración. De igual forma, se menciona como causa de la mala calibración la falta de regularización en las redes modernas.

La calibración de los modelos de aprendizaje de máquina se analiza en [43], y se concluye que el *naive Bayes* hace una fuerte suposición de independencia sobre los datos, lo que provoca una falta de calibración en las predicciones. Mientras tanto, modelos como los *bagged trees* y algunas arquitecturas de redes neuronales generan probabilidades bien calibradas. Además, los clasificadores que se basan en la maximización del margen, como las máquinas de soporte vectorial, provocan distorsiones en las predicciones y, en consecuencia, problemas de calibración [43].

Por otra parte, en [44] se evalúan métricas para cuantificar la calibración, métodos para calibrar estimadores en procesamiento de lenguaje natural y estimar intervalos de confianza para el error de calibración.

Algunas de las similitudes entre las investigaciones que comparan diferentes métodos de calibración o analizan diversos clasificadores son las métricas de rendimiento aplicadas para cuantificar qué tan bien discrimina el modelo, las métricas para medir la calibración y los conjuntos de datos que se utilizan. Dado que en un apartado anterior se hizo especial hincapié en las métricas y formas de cuantificar la calibración, en este caso daremos prioridad al resto de similitudes.

Así, la Tabla 2-1 resume las principales métricas utilizadas para medir el desempeño de los modelos en términos de discriminación. Se observa que en varias de las investigaciones se utiliza el logaritmo de la función de pérdida (*log-loss* en inglés), la exactitud o el error de clasificación, y el error cuadrático medio o MSE por sus siglas en inglés, pero no existe una metodología estándar para medir la calibración debido a las diferentes métricas que se aplican.

Finalmente, algunos de los conjuntos de datos que se destacan en los estudios comparativos que evalúan métricas o métodos de calibración se resumen en la Tabla 2-2. La mayoría de las investigaciones que evalúan la calibración en redes neuronales utilizan la colección de imágenes del “*Canadian Institute For Advanced Research*” (CIFAR-10) [34], conjunto de datos ampliamente utilizado en la literatura sobre aprendizaje automático. Sin embargo, existe

Table 2-1.: Uso de métricas de desempeño de clasificadores en estudios comparativos relacionados con la calibración

Referencia	Log-loss	Error Cuadrático Medio/ Raíz del Error Cuadrático Medio (MSE-RMSE)	Error	Log-verosimilitud negativa	Exactitud/ ROC AUC
[24]			X	X	
[37]		X			
[38]	X				
[41]		X			X
[43]	X	X			
[54]				X	
[56]	X				
[57]		X	X		

una amplia variedad de conjuntos de datos que cubren una amplia variedad de contextos y tareas de clasificación binaria o de múltiples clases.

Adicionalmente, en [41, 38] se utilizan más de 20 conjuntos de datos diferentes que cubren tareas binarias y de múltiples clases, pero solo algunas de ellas se presentan en la tabla 2-2. A diferencia de los conjuntos de datos reales, en [41, 54] también aplican sus métodos con datos generados aleatoriamente para analizar comportamientos particulares relacionados con la calibración.

2.3. Importancia de la calibración de modelos

La importancia de estudiar la calibración de los modelos de aprendizaje de máquina surge en contextos donde el costo de la clasificación errónea y las decisiones que se toman con base en los resultados son sensibles y pueden generar grandes costos [57]. Por ejemplo, en [57] se menciona que si un modelo predice el número de personas que son susceptibles a que hagan una donación monetaria en función de sus características, tener probabilidades no calibradas puede identificar incorrectamente a las personas que pueden donar cantidades muy bajas, en lugar de personas que podrían donar grandes cantidades de dinero.

Por otro lado, existen modelos utilizados para diagnosticar en el campo de la dermatología que son tan buenos como los obtenidos con juicio de expertos [31]. Sin embargo, para generar confianza en este tipo de métodos, se deben obtener estimaciones adecuadas de la precisión del modelo. Por ello, analizar la calibración en el área de la medicina es vital para lograr la confianza de las personas y fomentar su uso para diagnosticar o apoyar las decisiones que se toman, además de mejorar la interpretación de métodos como las redes neuronales [31, 24].

Es importante aclarar que la calibración de un modelo no genera ningún inconveniente en

Table 2-2.: Descripción de datos usados en estudios comparativos relacionados con la calibración

Conjunto de datos	Tarea de clasificación	Referencia
Adult	Salario de personas (Binario)	[57, 41, 43]
Breast	Cáncer de mama (Binario)	[41]
Heart	Cardiopatías (Multiclase)	[41]
Dermatology	Enfermedades eritematoescamosas (Multiclase)	[38]
Glass	Identificación de tipos de vidrios (Multiclase)	[38]
Birds	Especies de pájaros (Multiclase)	[24]
Cars	Marcas de carros por año y modelo (Multiclase)	[24]
CIFAR-10	Objetos varios (Multiclase)	[24, 38, 37, 54]
CIFAR-100	Objetos varios (Multiclase)	[24, 38]
ImageNet	Imágenes naturales (Multiclase)	[24, 37]
KDD-98	Personas susceptibles de hacer donaciones (Binario)	[57, 56]
Pendigits	Dígitos escritos a mano (Multiclase)	[38, 57, 41]
Reuters	Temas de noticias (Multiclase)	[24]
SST Binary/Fine Grained	Reseñas de películas (Binario & Multiclase)	[24]
SVHN	Dígitos en direcciones de casas (Multiclase)	[24, 38]
TIC	Servicios de pólizas de seguro (Binario)	[57]
20 Newsgroup	Temas de noticias (Multiclase)	[24, 57]

cuanto a la precisión, es decir, al comparar un estimador calibrado con uno no calibrado, ambos pueden diferir en el nivel de calibración que tienen, pero al mismo tiempo, tener un rendimiento similar en términos de exactitud [31]. Por el contrario, si se aplica un método para calibrar, esto podría mejorar la capacidad discriminatoria del modelo [20].

2.4. Revisión de literatura

Un resumen del trabajo relacionado mencionado hasta ahora se presenta en la Tabla 2-3.

Finalmente, teniendo en cuenta los principales conceptos involucrados en nuestro trabajo, el campo de intersección entre el aprendizaje de máquina y la estadística, la calibración y la física cuántica es el área de interés en el presente trabajo como se muestra en la Figura 2-3.

En la Figura 2-3 los principales conceptos involucrados son el aprendizaje de máquina y la estadística, y la física cuántica. Sin embargo, ambos campos abarcan muchas áreas que no son de interés en el presente trabajo, en especial, temas de física cuántica. Mientras tanto, el concepto de calibración puede verse como un sub-campo del aprendizaje automático por-

Table 2-3.: Resumen de los antecedentes

Tema	Referencia	Descripción
Definición de calibración	[9, 26, 53, 46, 3] [47, 4, 29, 17]	Calibración se refiere a encontrar parámetros desconocidos de modelos
	[24, 33, 35, 10, 19, 1]	Calibración se refiere al reflejo de la ocurrencia de un evento a través de las probabilidades de un estimador
Métricas de calibración	[5, 11, 18]	<i>Brier score</i>
	[16, 43, 6, 40, 44]	Diagrama de fiabilidad
	[24, 33, 10, 40]	Error esperado de calibración o error máximo de calibración
Métodos de calibración	[56]	Suavizamiento de probabilidades en árboles de decisión
	[57, 48]	<i>Platt scaling</i>
	[43]	Regresión isotónica
	[40]	Suavizamiento de etiquetas
	[36]	Calibración de múltiples clases
	[31]	Muestreo de etiquetas
Estudios comparativos	[36]	Calibración en aprendizaje en línea
	[37, 38, 57, 41]	Comparación o propuestas de métodos de calibración
	[24, 37, 56, 43, 54, 26, 44]	Análisis de calibración en modelos de aprendizaje de máquina
Importancia de la calibración	[57, 32, 10]	Importancia en aplicaciones críticas para la seguridad
	[33]	Medición de la incertidumbre
	[24, 16, 31]	Interpretación de los resultados
	[20]	Mejora de la exactitud del modelo
Aprendizaje de máquina cuántico	[51]	Perspectiva general
	[50, 30]	Aplicaciones en medicina
	[23]	Modelos con redes tensoriales
	[28, 21, 52]	Proyectiva cuántica y matrices de densidad

que las definiciones encontradas en la revisión de la literatura están relacionadas con tareas supervisadas y optimización. Por último, nuestro trabajo estudia la calibración de clasificadores ampliamente utilizados en la literatura y los compara con modelos que aprovechan el formalismo matemático de la física cuántica.



Figure 2-3.: Diagrama de Venn para representar las principales áreas de estudio involucradas en el presente trabajo

3. Quantum Measurement Classification

Las ideas principales detrás de los modelos de aprendizaje de máquina cuántico de interés se abordan en este capítulo, tales como la medición cuántica, las matrices de densidad, las *random Fourier features*, y cómo se utilizan estos conceptos para resolver una tarea de clasificación.

Los métodos cuánticos utilizados en este trabajo se proponen y describen en detalle en [21, 28]. En primer lugar, en [28] los autores propusieron un método novedoso llamado “*Quantum Measurement Classification*” (QMC) para el aprendizaje supervisado que se basa en el formalismo matemático de la mecánica cuántica, en especial, en funciones de mapeo cuántico, matrices de densidad y medición cuántica. Además, este clasificador no requiere ninguna optimización de parámetros a diferencia de la mayoría de los modelos ampliamente utilizados en aprendizaje de máquina.

Posteriormente, en [21] se muestra la generalización de la teoría presentada en [28] relacionada a la estimación de la densidad usando funciones *kernel*, y cómo se podría usar para tareas de clasificación y regresión. Además, los autores presentan algunas propiedades del método QMC que permiten optimizar los parámetros de la matriz de densidad (QMC-SGD), y aplicar la descomposición de Schmidt a la matriz de densidad para reducir su dimensión y mejorar el tiempo de ejecución (QMC-D).

3.1. Quantum Measurement Classification

3.1.1. Mapeo de características cuánticas

Como se menciona en [28], el primer paso para entrenar los clasificadores QMC, QMC-SGD y QMC-D es realizar un mapeo cuántico sobre las características x y la variable objetivo y a los espacios cuánticos de Hilbert \mathcal{H}_x y \mathcal{H}_y para las n muestras, como se muestra en la siguiente expresión usando notación de Dirac

$$\begin{aligned} \psi : \mathcal{X} \times \mathcal{Y} &\longrightarrow \mathcal{H}_x \otimes \mathcal{H}_y \\ (x, y) &\longmapsto |\psi_x(x)\rangle \otimes |\psi_y(y)\rangle \end{aligned} \tag{3-1}$$

Algunas de las funciones de mapeo de características cuánticas descritas en [28] son la codificación softmax, la codificación *one-hot*, los estados comprimidos, los estados coherentes y las *random Fourier features*. Además, cada función de mapeo tiene diferentes hiperparámetros que conducen a resultados distintos, por lo que su selección es relevante.

En particular, las *random Fourier features* proporcionan una aproximación de un *kernel* invariante en un espacio de Hilbert explícito. Así, en [21] los autores muestran la relación entre la estimación de densidad *kernel* y QMC, mediante el uso de *random Fourier features* como método de aproximación del *kernel* Gaussiano.

3.1.2. Matrices de densidad

El segundo paso es resumir el estado cuántico de la muestra de entrenamiento a través del cálculo de la matriz de densidad ρ_{train} que representa la distribución conjunta $P(x, y)$ [28]. El cálculo de esta matriz de densidad se puede realizar de tres formas diferentes dependiendo de la incertidumbre (cuántica o clásica) que se quiera representar. Para los clasificadores entrenados en este trabajo se utilizó el cálculo de la matriz de densidad como estado mezclado.

Dependiendo del mapeo cuántico utilizado, las operaciones involucradas en el cálculo de la matriz de densidad son diferenciales, lo cual hace posible usar el algoritmo de gradiente descendente con una función de pérdida adecuada como log-verosimilitud negativa para optimizar los parámetros [21], en cuyo caso se denomina QMC-SGD.

Si bien este procedimiento no garantiza estimar la matriz de densidad real, ha mostrado un buen desempeño en comparación con el cálculo de la matriz de densidad sin optimización, y tiene la ventaja de poder ser entrenado en conjunto con otras arquitecturas de aprendizaje profundo, en cuyo caso se denomina QMDM.

Además, en [21] se muestra que la matriz de densidad podría expresarse usando descomposición de Schdmit de la siguiente forma

$$\rho_{train} = V^T \Lambda V, \tag{3-2}$$

donde $V \in \mathbb{R}^{r \times D}$, $\Lambda \in \mathbb{R}^{r \times r}$ es una matriz diagonal y $r < D$ es el rango reducido de la factorización, D es la dimensión del mapeo cuántico (*random Fourier features*). La ecuación 3-2 permite reducir el tiempo de cálculo de la densidad de una nueva muestra, debido al uso de $r < D$, este método se denomina QMC-D.

Vale la pena señalar que QMC-D no requiere el uso de la optimización de los parámetros de la matriz de densidad, sin embargo, en este trabajo QMC-D se refiere al uso de la descomposición descrita en 3-2 de la matriz de densidad en combinación con la optimización de sus

parámetros.

Finalmente, en este trabajo se utilizaron las *random Fourier features* como un mapeo cuántico para aproximar un *kernel* Gaussiano. Luego, como se describe en [21], los hiperparámetros de los modelos son la dimensión de las *random Fourier features* D , el parámetro de dispersión γ del *kernel* Gaussiano y el rango de la factorización de la matriz de densidad en el caso del clasificador QMC-D. Mientras tanto, los parámetros corresponden a los pesos y sesgos de las *random Fourier features*, y los componentes de la descomposición, $V \in \mathbb{R}^{r \times D}$ y $\lambda \in \mathbb{R}^r$, el vector con los elementos de la diagonal de Λ . Por lo tanto, los hiperparámetros de los clasificadores se encuentran a través de validación cruzada o un proceso iterativo, pero los parámetros se encuentran de manera óptima a través del gradiente descendente.

3.1.3. Medición cuántica

Para calcular el vector de probabilidad asociado a las clases objetivo y en función de las características x , se define un operador de predicción en $\mathcal{H}_X \otimes \mathcal{H}_Y$, que permite calcular una medida cuántica sobre la matriz de densidad ρ_{train} [28], dado por la expresión

$$\pi(x) = |\psi_X(x)\rangle\langle\psi_X(x)| \otimes Id_{\mathcal{H}_Y}, \quad (3-3)$$

donde $Id_{\mathcal{H}_Y}$ es el operador identidad en \mathcal{H}_Y . Al realizar la medición cuántica del operador $\pi(\cdot)$ sobre la matriz de densidad ρ_{train} y un vector x de características, se tiene la matriz

$$\rho' = \frac{\pi(x)\rho_{train}\pi(x)}{Tr[\pi(x)\rho_{train}\pi(x)]}, \quad (3-4)$$

que contiene la información del subsistema \mathcal{S}_X y \mathcal{S}_Y de forma conjunta.

3.1.4. Predicción

El vector de probabilidades se obtiene a partir de la traza parcial de la medición cuántica (3-4) respecto al subsistema \mathcal{S}_X , es decir, $\rho'_Y = Tr_X[\rho']$.

Los pasos anteriores se siguen con las muestras de entrenamiento para la estimación de la matriz de densidad y predicción del vector de probabilidades. De esta forma, para calcular el vector de probabilidades para una nueva muestra x^* una vez se calcula el ρ_{train} con los datos de entrenamiento, solo el mapeo cuántico (3-1), el operador cuántico (3-3), la medición cuántica (3-4), y la traza parcial que depende del ρ_{train} se deben aplicar como se menciona en [28].

3.2. Kernel Quantum Measurement

De forma similar a la propuesta en [21] en la que se evidencia la relación de las matrices de densidad y la medición cuántica con la estimación de densidad basada en *kernels*, dichos autores proponen un método de aprendizaje de medición cuántica basado en *kernels* o *Kernel Quantum Measurement* (KQM) en el cual hacen uso de las matrices de densidad para describir las relaciones entre variables x y y que puede ser visto como probabilístico y basado en *kernels*.

Este se sustenta en la descomposición de ρ_{train} que permite expresar el cálculo de las proyecciones en términos del producto punto de la representación de las muestras. Esto permite entonces usar el *kernel trick* en lugar de hacer el cálculo explícito con la representación dada por las *random Fourier features*, y por tanto, el *kernel* se convierte en un hiperparámetro más del modelo.

Lo anterior se obtiene debido a que las matrices de densidad se definen en un espacio asociado a una función *kernel*, por medio de la función de mapeo de características cuánticas como se menciona en [28, 21]. Ofreciendo así, un método que permite trabajar en altas dimensiones, y que se sustenta en el uso de las matrices de densidad como representaciones de funciones de probabilidad.

4. Conjuntos de datos

En este capítulo se describirán los diferentes conjuntos de datos públicos utilizados para la experimentación. Se seleccionaron seis conjuntos de datos: CIFAR-10, Adult, Breast, Dermatology, Glass, y Heart que se describen completamente en la Tabla 2-2. Entre estos conjuntos de datos hay diferentes escenarios: binario y de múltiples clases en el que se encuentra el conjunto de imágenes (CIFAR-10). Por lo tanto, el conjunto de datos Adult, Breast, y Heart tienen una variable objetivo binaria, y los otros conjuntos de datos son de múltiples clases. El conjunto de datos Heart se utilizó como binario porque ha sido la forma habitual en la que se ha utilizado a pesar de tener múltiples clases [15, 13, 2].

Todos los conjuntos de datos están disponibles públicamente y se recopilaron principalmente a través del Repositorio de Aprendizaje de Máquina de UCI [12] (Adult, Dermatology, Heart y Glass), mientras que los datos de Breast y CIFAR-10 se obtuvieron a través de los módulos de Python *Sklearn* y *Keras* [49, 8].

La Tabla 4-1 muestra el número de registros utilizados para entrenar, seleccionar los parámetros óptimos y probar el desempeño de los modelos para las diferentes tareas, es decir, las particiones de los datos en entrenamiento, validación y prueba. Solo CIFAR-10 contenía una partición preestablecida de entrenamiento y prueba, en cuyo caso agregamos solo una partición aleatoria en los datos de entrenamiento para validar los resultados. Para lograr un porcentaje similar de muestras en las particiones del conjunto de datos CIFAR-10 en los otros conjuntos de datos, para Adult, Breast, Dermatology y Heart, se generaron particiones aleatorias de 76,5 %, 13,5 % y 10 % y, por último, para Glass, debido a tener múltiples clases desequilibradas y para asegurar que todas las particiones tuvieran ejemplos de todas las categorías, se seleccionaron aleatoriamente particiones de 72.25 %, 12.75 % y 15 % para cada categoría.

La tarea de clasificación en cada conjunto de datos es diferente, e incluye problemas de diversas áreas. El conjunto de datos CIFAR-10 contiene imágenes a color de 32x32x3 píxeles donde la variable objetivo representa objetos como automóviles, barcos, aviones y animales, y el objetivo es tratar de identificar todos estos elementos. La tarea en el conjunto de datos Adult es intentar predecir si los ingresos de las personas en EE. UU. superan los \$50000 dolares por año según los datos del censo. Por otro lado, la clase objetivo del conjunto de datos Glass representa el tipo de vidrio que se usa en los vehículos, las ventanas y otros, lo

cual es de interés en las ciencias forenses. Por último, las tareas en los conjuntos de datos de Breast, Dermatology y Heart están relacionadas con condiciones médicas en las que los clasificadores calibrados se vuelven relevantes.

Table 4-1.: Descripción de los conjuntos de datos utilizados en la experimentación y el número de datos usados en cada una de las particiones. Las palabras “Train” y “Val” son usadas para referirse a las particiones de entrenamiento y validación en los conjuntos de datos

Datos	Particiones	Características	Clases objetivo y su distribución
CIFAR-10	Train: 45000 Val: 5000 Prueba: 10000	Imágenes a color 32x32x3 píxeles	Train: (4477, 4517, 4496, 4521, 4486, 4488, 4526, 4534, 4480, 4475) Val: (523, 483, 504, 479, 514, 512, 474, 466, 520, 525) Prueba: (1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000)
Adult	Train: 24998 Val: 4396 Prueba: 3257	numéricas: 5 categóricas: 9	Tr: (18870, 6038) Val: (3394, 1002) Prueba: (2456, 801)
Breast	Train: 435 Val: 77 Prueba: 57	numéricas: 30 categóricas: 0	Train: (169, 266) Val: (26, 51) Prueba: (17, 40)
Dermatology	Train: 273 Val: 49 Prueba: 36	numéricas: 1 categóricas: 33	Train: (85, 51, 50, 32, 39, 16) Val: (11, 7, 15, 9, 5, 2) Prueba: (15, 2, 6, 7, 4, 2)
Glass	Train: 149 Val: 30 Prueba: 35	numéricas: 10 categóricas: 0	Tr: (50, 54, 11, 9, 5, 20) Val: (9, 10, 3, 2, 2, 4) Prueba: (11, 12, 3, 2, 2, 5)
Heart	Train: 226 Val: 41 Prueba: 30	numéricas: 5 categóricas: 8	Tr: (120, 106) Val: (21, 20) Prueba: (19, 11)

En la Tabla 4-1 se muestra la distribución de cada clase para los conjuntos de datos en las particiones de entrenamiento, validación y prueba. Las distribuciones de Adult, Dermatology y Glass están desbalanceadas, y las demás tienen un conteo más equilibrado en las clases objetivo. Por lo tanto, los conjuntos de datos seleccionados reflejan los desafíos que se pueden observar en los datos de la vida real, es decir, clases desbalanceadas para tareas de múltiples clases o binarias con características numéricas y categóricas.

Es importante señalar que los valores faltantes no están presentes en todas las variables de los conjuntos de datos. Por lo general, solo unos pocos registros contienen al menos un valor faltante en una variable, y el conjunto de datos que tiene más de estos valores es Adult con 90 muestras. Por lo tanto, en todos los casos se ignoró esa información debido a los pocos registros que tenían ese problema, y no fue necesario aplicar ningún método de imputación.

5. Análisis de calibración

Este capítulo presenta el análisis de calibración para las tareas descritas y mencionadas en el capítulo anterior aplicando diferentes modelos del estado del arte en aprendizaje de máquina como líneas de base y comparando los resultados con los clasificadores de aprendizaje de máquina cuántico. Observamos que los modelos cuánticos tienden a discriminar las clases objetivo de forma similar a las líneas base, pero al mismo tiempo mantienen métricas de calibración más bajas.

5.1. Modelos de aprendizaje de máquina

Para la experimentación se utilizaron algunos modelos de aprendizaje automático ampliamente utilizados como regresión logística (LR), *naive Bayes* (NB), máquinas de soporte vectorial con *kernel* lineal (SVM-L) y Gaussiano (SVM-RBF), árboles de decisión (DT), *random forest* (RF), *k*-vecinos más cercanos (kNN) y perceptrón multicapa con una (MLP 1) y dos (MLP 2) capas densas cuyas propiedades de calibración se han estudiado en trabajos anteriores [43, 32, 56]. Adicionalmente, para la experimentación relacionada con redes neuronales, se utilizó LeNet5 debido a que en los antecedentes fue una de las redes más utilizadas para este tipo de experimentos [24, 38]. La arquitectura de LeNet5 utilizada es la propuesta en el artículo original [39].

Por otro lado, los modelos cuánticos aplicados son los descritos en el Capítulo 3, es decir, QMC, QMC-D, QMC-SGD, KQM y QMDM, los cuales tienen diferentes hiperparámetros que requieren ser optimizados para encontrar mejores soluciones. Además, es importante tener en cuenta que la función de mapeo cuántico utilizada en QMC, QMC-D, QMC-SGD y QMDM son las *random Fourier features*. Mientras tanto, en KQM esa función de mapeo se reemplaza por un *kernel* explícito, por ejemplo, Gaussiano (KQM-RBF), coseno (KQM-COS) o polinomial (KQM-POL).

5.2. Configuración experimental

5.2.1. Preprocesamiento

Como se mencionó anteriormente, los conjuntos de datos utilizados para la experimentación son los descritos en el Capítulo 4. Además, una vez descartados los valores faltantes,

se estandarizaron las variables numéricas utilizando la media y la desviación estándar para mejorar la convergencia de los modelos, dado que este preprocesamiento en los datos tiende a generar mejores resultados en términos de discriminación de las clases objetivo, mientras tanto, las variables categóricas se representaron en una forma de codificación *one-hot*.

El preprocesamiento de los datos se aplicó para los conjuntos de datos Adult, Breast, Dermatology, Glass y Heart. El conjunto de datos CIFAR-10 se normalizó dividiendo el valor de cada píxel por 255 y se restó la media. Además, durante el entrenamiento del modelo que se utilizó con CIFAR-10, se aplicó el aumento de datos utilizando desplazamientos aleatorios horizontales y verticales de 0.1 y giros horizontales para todas las imágenes de entrenamiento.

Cabe señalar que el preprocesamiento se realizó con los datos de entrenamiento, validación y prueba, y en caso de estandarización se utilizó la media y desviación estándar de los datos en la partición de entrenamiento, similar para la resta utilizando la media para el conjunto de datos CIFAR-10.

5.2.2. Búsqueda de hiperparámetros

La experimentación se dividió en dos partes según los objetivos. Los primeros experimentos buscan medir el efecto del uso de QMDM en capas superiores sobre una red neuronal en relación con la calibración del modelo aprovechando las propiedades de características de extracción que tienen las redes neuronales. En segundo lugar, exploramos la calibración de los modelos cuánticos QMC, QMC-D, QMC-SGD y KQM en comparación con regresión logística, *naive Bayes*, máquinas de soporte vectorial con funciones *kernel* lineal y Gaussiano, árboles de decisión, *random forest*, kNN y MLP.

Para el primer experimento, la arquitectura LeNet5 se entrenó en el conjunto de datos CIFAR-10 y, una vez que se registraron las métricas de interés, las dos capas densas antes de la salida se reemplazaron por una capa de QMDM donde la dimensión de las *random Fourier features* D , el parámetro de dispersión γ de esa función y la dimensión d de la factorización de la matriz de densidad fueron seleccionadas en base a la exactitud de validación de una cuadrícula de 25 combinaciones de parámetros aleatorios generados por valores de (2048, 1024, 512, 256), $(2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}, 2^{-8}, 2^{-9})$ y (128, 256, 512, 1024) respectivamente. Además, el tamaño del lote para entrenar las redes neuronales fue de 256, el número de épocas de 200 y una tasa de aprendizaje inicial de 0.001 que decayó en un factor de 0.1, 0.01, 0.001 y 0.0001 en las épocas 80, 120, 160 y 180 con un optimizador Adam usando la función categórica de pérdida de entropía cruzada.

Se utilizaron parámetros similares para el entrenamiento de la arquitectura LeNet5, pero la tasa de aprendizaje utilizada en este caso fue 0.005 y la función de pérdida utilizada fue el

error cuadrático medio como en [39]. Dada la aleatoriedad de la optimización de los pesos de la red neuronal, se repitió el experimento diez veces y se seleccionaron los mejores resultados de las repeticiones.

Por otro lado, los conjuntos de datos de Adult, Breast, Dermatology, Glass y Heart se utilizaron para entrenar los modelos cuánticos QMC, QMC-D, QMC-SGD y KQM, y para comparar los resultados con las líneas base: regresión logística, *naive Bayes*, máquinas de soporte vectorial, árboles de decisión, *random forest*, kNN y MLP. Así, para el QMC se exploraron los hiperparámetros D y γ de las *random Fourier features* entre 5 y 120, y 0.0001 y 3 respectivamente.

Luego de la selección de los mejores hiperparámetros de QMC, para la estrategia QMC-D se varió la dimensión de la descomposición d de la matriz de densidad, valor que se exploró entre 3 y $2 \cdot D$ en base a D y γ seleccionado previamente. A diferencia del modelo QMC, QMC-D requiere optimizar los valores de la estimación de la matriz de densidad. Para lograrlo, se utilizó el optimizador de Adam con una tasa de aprendizaje de 0.005 con la función de pérdida de entropía cruzada categórica y 20 épocas.

Para el modelo QMC-SGD, los hiperparámetros D y γ se exploraron en el rango de 5 a 85 y de 0.0001 a 5 respectivamente. Adicionalmente, se utilizó el optimizador de Adam con una tasa de aprendizaje entre 0.005 y 0.01, la función de pérdida de entropía cruzada categórica y un número de épocas que varía entre 10 y 50 según el conjunto de datos. Además, para cada combinación de hiperparámetros se repitió el experimento 5 veces debido a la aleatoriedad del proceso de optimización.

Finalmente, para el KQM con *kernel* Gaussiano se exploró el valor inicial de σ alrededor de 0.001 y 2.5, el *kernel* del polinomio usó el mismo rango para σ y el coeficiente α del polinomio tomó un valor inicial de 0.5, mientras que el *kernel* coseno no requirió hiperparámetros iniciales. Después de seleccionar los mejores hiperparámetros para cada *kernel*, se seleccionaron los resultados del mejor. Además, el KQM requiere como entrada el número inicial de muestras para entrenar el modelo y dependiendo de ese valor, la convergencia de los resultados es mejor o peor. Por lo tanto, dependiendo del conjunto de datos, el número de muestras exploradas varió en los siguientes valores (8, 16, 32, 64, 128, 256, 512). Finalmente, se utilizaron 50 épocas, un tamaño de lote de 256, una tasa de aprendizaje inicial de 0.05 y la función de pérdida de entropía cruzada categórica con el optimizador de Adam.

En la selección de los mejores hiperparámetros para entrenar los modelos cuánticos se utilizó la exactitud de validación, es decir, la combinación de los hiperparámetros que se seleccionaron fueron los que maximizaron esa métrica. Adicionalmente, para mejorar la exploración de los hiperparámetros se corrieron por primera vez los experimentos de los modelos cuánticos

utilizando un amplio rango de valores para identificar dónde se pueden encontrar las mejores soluciones, luego se ejecutó una segunda vez con un rango más restringido y estrecho.

Todos los resultados de los modelos cuánticos relacionados con el segundo tipo de experimentos se compararon con regresión logística, *naive Bayes*, máquinas de soporte vectorial, árboles de decisión, *random forest*, kNN y MLP. Para entrenar estos modelos, se realizó una validación cruzada con 10 particiones y los hiperparámetros explorados para cada modelo se resumen en la Tabla 5-1. Además, para el MLP se utilizó un optimizador Adam con una tasa de aprendizaje inicial de 0.001 que disminuye si no mejora la pérdida de entrenamiento.

Table 5-1.: Ajuste de hiperparámetros para clasificadores de referencia

Modelo	Hiperparámetros	Rango explorado
LR	Inverso de la fuerza de regularización (C)	0.001 - 5
NB	-	-
SVM-L	Parámetro de regularización (C)	0.001 - 5
SVM-RBF	Parámetro de regularización (C)	0.001 - 5
	Coficiente del <i>kernel</i> (Γ)	0.01 - 10
RF	Profundidad máxima de los árboles (max_depth)	2 - 50
	Número de árboles (n_est)	10 - 500
kNN	Número de vecinos (k)	2 - 50
	Función de pesos (w)	Uniforme & distancia
	Potencia de la distancia de Minkowski (p)	2 - 50
DT	Profundidad máxima de los árboles (max_depth)	2 - 50
MLP 1	Tamaño de la capa (n_1)	16-512
	Función de activación (act)	Relu & identidad
MLP 2	Tamaño de la primera capa (n_1)	16 - 256
	Tamaño de la segunda capa (n_2)	16 128
	Función de activación (act)	Relu & identidad

Para todos los conjuntos de datos, se entrenaron los clasificadores descritos en la Tabla 5-1. Sin embargo, debido al costo computacional, los modelos SVM y kNN no se entrenaron para el conjunto de datos Adult que tiene aproximadamente 25000 muestras de entrenamiento.

Por otro lado, en la Tabla 5-2 y 5-3 se describen los hiperparámetros seleccionados para cada conjunto de datos y clasificador para el experimento comparativo, mientras tanto, los hiperparámetros encontrados para el experimento LeNet-QMDM son D : 2048 y γ : 0.002. Los valores encontrados son bastante diferentes entre los clasificadores, lo que refleja la importancia de la búsqueda de hiperparámetros en los modelos de aprendizaje automático.

Table 5-2.: Hiperparámetros seleccionados para todos los clasificadores en los conjuntos de datos binarios Adult, Breast, y Heart

Classifier	Adult	Breast	Heart
DT	<i>max_depth</i> : 12	<i>max_depth</i> : 5	<i>max_depth</i> : 2
kNN	-	<i>k</i> : 10; <i>p</i> : 2; <i>w</i> : uniforme	<i>k</i> : 6; <i>p</i> : 2; <i>w</i> : uniforme
LR	<i>C</i> : 0.708	<i>C</i> : 0.501	<i>C</i> : 0.152
MLP 1	n_1 : 32; <i>act</i> : identidad	n_1 : 16; <i>act</i> : relu	n_1 : 32; <i>act</i> : relu
MLP 2	n_1 : 64; n_2 : 16; <i>act</i> : identidad	n_1 : 128; n_2 : 64; <i>act</i> : relu	n_1 : 128; n_2 : 16; <i>act</i> : identidad
NB	-	-	-
RF	<i>max_depth</i> : 23; <i>n_est</i> : 490	<i>max_depth</i> : 8; <i>n_est</i> : 490	<i>max_depth</i> : 5; <i>n_est</i> : 40
SVM	-	<i>C</i> : 1.768 <i>kernel</i> : lineal	<i>C</i> : 1.646; γ : 0.093 <i>kernel</i> : Gaussiano
QMC	<i>D</i> : 68; γ : 0.158	<i>D</i> : 23; γ : 0.158	<i>D</i> : 21; γ : 0.159
QMC-D	<i>D</i> : 68; γ : 0.158; <i>d</i> : 90	<i>D</i> : 23; γ : 0.158; <i>d</i> : 9	<i>D</i> : 21; γ : 0.159; <i>d</i> : 9
QMC-SGD	<i>D</i> : 15; γ : 0.222	<i>D</i> : 15; γ : 0.0001	<i>D</i> : 20; γ : 0.0001
KQM	<i>D</i> : 128; σ : 1.243 <i>kernel</i> : Gaussiano	<i>D</i> : 512; σ : 2.073 <i>kernel</i> : Gaussiano	α : 0.450; <i>D</i> : 128; σ : 1.250 <i>kernel</i> : polinomial

5.2.3. Métricas de calibración y desempeño

La salida de los modelos de referencia fueron directa o indirectamente las estimaciones de probabilidades para cada clase cuya medición de calibración es de interés. Sin embargo, la salida del modelo SVM es el resultado de la función de decisión que no varía en un rango particular. Por esta razón, en las tareas binarias los valores de la función de decisión se convirtieron en probabilidades utilizando la función logística mencionada por [48], y en las tareas de múltiples clases se utilizó la función softmax.

Para el entrenamiento de los diferentes modelos se utiliza la función de pérdida o la exactitud para seleccionar los mejores hiperparámetros y medir el poder discriminatorio de los clasificadores. Además, para el primer experimento donde se utilizó el conjunto de datos CIFAR-10 en cada época se midió el ECE (ecuación 2-5) y cwECE (ecuación 2-6), mientras tanto, para el QMC-SGD y en los modelos KQM se midió el ECE para explorar la relación entre la capacidad discriminatoria del modelo y su calibración.

En todos los experimentos, después del entrenamiento del clasificador, se midió el *Brier Score*, *log-loss* (logaritmo de la función de pérdida), exactitud, ECE con esquema de partición uniforme y dependiente de datos [54], y cwECE con esquema de partición uniforme que se midieron sobre el conjunto de prueba. Adicionalmente, se realizaron los diagramas de confiabilidad para el ECE y cwECE para explorar visualmente la calibración de los modelos.

Finalmente, los experimentos se ejecutaron en Python 3.8 por medio de Google Colab Pro. Los principales módulos usados para entrenar los modelos fueron *Sklearn* [49] y *Keras* [8], y para calcular las métricas de calibración se usó el modulo creado por [54] en combinación con

Table 5-3.: Hiperparámetros seleccionados para todos los clasificadores en los conjuntos de datos de múltiples clases Dermatology y Glass

Modelo	Dermatology	Glass
DT	<i>max_depth</i> : 8	<i>max_depth</i> : 5
kNN	<i>k</i> : 8; <i>p</i> : 2; <i>w</i> : distancia	<i>k</i> : 6; <i>p</i> : 2; <i>w</i> : distancia
LR	<i>C</i> : 0.758	<i>C</i> : 5.0
MLP 1	<i>n</i> ₁ : 64; <i>act</i> : relu	<i>n</i> ₁ : 512; <i>act</i> : relu
MLP 2	<i>n</i> ₁ : 256; <i>n</i> ₂ : 16; <i>act</i> : relu	<i>n</i> ₁ : 256; <i>n</i> ₂ : 64; <i>act</i> : relu
NB	-	-
RF	<i>max_depth</i> : 14; <i>n_est</i> : 55	<i>max_depth</i> : 44; <i>n_est</i> : 85
SVM	<i>C</i> : 1.415; <i>kernel</i> : lineal	<i>C</i> : 2.896; γ : 0.096 <i>kernel</i> : Gaussiano
QMC	<i>D</i> : 104; γ : 0.053	<i>D</i> : 84; γ : 0.316
QMC-D	<i>D</i> : 104; γ : 0.053; <i>d</i> : 147	<i>D</i> : 84; γ : 0.316; <i>d</i> : 150
QMC-SGD	<i>D</i> : 30; γ : 0.0001	<i>D</i> : 65; γ : 0.0001
KQM	α : -0.633; <i>D</i> : 32; σ : 0.341 <i>kernel</i> : polinomial	<i>D</i> : 128; σ : 0.601 <i>kernel</i> : Gaussiano

código propio que se encuentra en el repositorio <https://github.com/ghamayac/Calibration>.

5.3. Resultados

En esta sección se presentan los resultados de los experimentos descritos separados para las tareas de clasificación binaria y de múltiples clases con el fin de facilitar la lectura del documento. De esta forma, se presentan las distintas métricas de calibración y desempeño de los modelos, así como los diagramas de fiabilidad.

5.3.1. Clasificación binaria

Métricas de calibración y desempeño

Para comparar objetivamente los clasificadores, se utilizaron las métricas de calibración y la exactitud. Por lo tanto, en la Tabla 5-4 para el conjunto de datos Adult se presentan estas métricas para los distintos modelos. Se puede apreciar que el clasificador *random forest* tiene la exactitud más alta en la partición de prueba, pero el cwECE y el ECE con el esquema uniforme y dependiente de los datos no son los más bajos entre los demás modelos. Sin embargo, el clasificador MLP con una capa tiene una precisión similar en la partición de prueba y refleja mejores métricas de calibración.

Por otro lado, el modelo logístico, *naive Bayes*, MLP con dos capas y árboles de decisión tienen métricas de exactitud similares y casi idénticas, pero no muestran el mismo comportamiento para la calibración (Tabla 5-4). De esta forma, MLP y la regresión logística casi

comparten las mismas métricas de calibración, mientras que *naive Bayes* y los árboles de decisión no están bien calibrados. En particular, se observa que a pesar de tener un buen desempeño en la tarea de clasificación, el clasificador *naive Bayes* es el que refleja el peor comportamiento en cuanto a la calibración lo cual se evidencia con los valores más altos en el *Brier score*, *log-loss*, y las estimaciones del error de calibración.

En la Tabla 5-4 los clasificadores cuánticos no reflejan los mejores resultados en términos de precisión, de hecho, ninguno de los modelos tiene un mejor desempeño que los modelos de referencia. Sin embargo, KQM, QMC-SGD y QMC-D tienen una exactitud cercana y tienen una exactitud en la partición de prueba más baja que el mejor clasificador en aproximadamente solo 0.2. Además, QMC-SGD y KQM tienen una de las métricas más bajas en las estimaciones del error de calibración, incluso KQM tiene mejores métricas de calibración que la regresión logística. Así, se tiene que estos modelos cuánticos están bien calibrados. Finalmente, QMC y QMC-D no evidencian un buen desempeño en la calibración y el poder discriminatorio del modelo.

Table 5-4.: Métricas de desempeño y calibración en la partición de prueba para el conjunto de datos Adult y los distintos modelos entrenados

Modelo	Brier score	Log-loss	Exactitud entrenamiento	Exactitud validación	Exactitud prueba	cwECE	ECE dependiente de datos	ECE uniforme
QMC	0.311	0.476	0.758	0.773	0.754	0.118	0.118	0.120
QMC-D	0.218	0.353	0.891	0.858	0.845	0.043	0.041	0.038
QMC-SGD	0.196	0.310	0.864	0.854	0.857	0.015	0.012	0.019
KQM	0.191	0.299	0.864	0.859	0.859	0.023	0.017	0.011
DT	0.203	0.730	0.878	0.880	0.861	0.028	0.024	0.025
LR	0.196	0.307	0.851	0.856	0.861	0.017	0.013	0.012
MLP 1	0.197	0.308	0.851	0.859	0.865	0.019	0.017	0.017
MLP 2	0.196	0.307	0.852	0.858	0.861	0.017	0.013	0.013
NB	0.902	13.927	0.851	0.856	0.861	0.454	0.450	0.449
RF	0.184	0.295	0.937	0.937	0.871	0.031	0.027	0.023

En el conjunto de datos Breast, la mayoría de los clasificadores tienen una precisión superior a 0.90, sin embargo, para QMC y QMC-D ese valor es inferior a 0.78. En particular, KQM, regresión logística, *naive Bayes*, SVM, MLP con una capa y QMC-SGD tienen casi la misma exactitud en la partición de prueba, pero el modelo KQM tiene ligeramente los mejores resultados en esa métrica, y a pesar de no ser el que tiene las estimaciones más bajas en el error de calibración, *Brier score* y *log-loss*, tiene métricas competitivas con respecto a los mejores modelos (Tabla 5-5). Adicionalmente, entre los clasificadores mencionados con la misma exactitud en la partición de prueba, el QMC-SGD y MLP con una capa son los que tienen estimaciones más bajas en el cwECE, ECE con el esquema de partición dependiente de datos y el *Brier score*.

A diferencia de los resultados para el conjunto de datos Adult, en este caso el modelo *naive Bayes* no tiene los peores resultados en términos de calibración (Tabla 5-5). El QMC y QMC-D además de que tienen los resultados más bajos de exactitud, están entre los modelos con estimaciones más altas del error de calibración, el *Brier score* o *log-loss*.

En general, la exactitud de algunos de los métodos cuánticos es más alta a comparación de algunos modelos de referencia lo que no ocurre en el conjunto de datos Adult, pero el modelo QMC tiene la peor exactitud entre todos los clasificadores nuevamente y sus métricas de calibración no son las mejores, lo cual tiene sentido teniendo en cuenta que este modelo no optimiza ningún parámetro y da los resultados que se obtienen directamente de la matriz de densidad (Tabla 5-5).

Table 5-5.: Métricas de desempeño y calibración en la partición de prueba para el conjunto de datos Breast y los distintos modelos entrenados

Modelo	Brier score	Log-loss	Exactitud entrenamiento	Exactitud validación	Exactitud prueba	cwECE	ECE dependiente de datos	ECE uniforme
QMC	0.344	0.523	0.717	0.701	0.754	0.178	0.144	0.224
QMC-D	0.316	0.478	0.998	0.857	0.772	0.155	0.133	0.146
QMC-SGD	0.039	0.086	0.993	0.935	0.982	0.029	0.025	0.032
KQM	0.063	0.106	0.982	0.987	0.987	0.058	0.052	0.050
DT	0.160	2.454	0.993	1.000	0.912	0.072	0.072	0.080
kNN	0.073	0.114	0.975	0.987	0.947	0.044	0.037	0.016
LR	0.043	0.084	0.979	0.987	0.982	0.056	0.051	0.033
MLP 1	0.041	0.077	0.993	0.987	0.982	0.033	0.031	0.035
MLP 2	0.046	0.154	1.000	1.000	0.965	0.027	0.027	0.023
NB	0.075	0.283	0.979	0.987	0.982	0.042	0.042	0.027
RF	0.051	0.090	1.000	1.000	0.965	0.065	0.037	0.019
SVM	0.098	0.220	0.989	0.987	0.982	0.161	0.161	0.161

Por último, en la Tabla 5-6 se muestra que para el conjunto de datos Heart la estrategia KQM refleja los mejores resultados en la exactitud en la partición de prueba y las métricas de calibración, a pesar de no ser la mejor se mantienen competitivas en comparación con los demás clasificadores. Sin embargo, SVM y kNN tienen la segunda mejor exactitud, donde el kNN también tiene algunas de las mejores métricas de calibración.

La regresión logística, *naive Bayes*, *random forest*, MLP y QMC-SGD tienen la cuarta mejor exactitud en el conjunto de prueba, sin embargo, no todos los clasificadores tienen un comportamiento similar en su calibración. Así, la regresión logística y MLP con una capa tienen los mejores resultados en las métricas de calibración (Tabla 5-6). Por el contrario, entre estos modelos, QMC-SGD no está bien calibrado porque el *Brier score*, la *log-loss* y las estimaciones del error de calibración no son de las más bajas. Vale la pena señalar que *naive Bayes* tiene un buen desempeño en general, y sus métricas de calibración también son

unas de las más bajas, aunque este modelo no evidencia buenas propiedades de calibración como se menciona en [56].

Al igual que los conjuntos de datos anteriores, el clasificador QMC no refleja un buen desempeño en las métricas de calibración ni en la exactitud. Por otro lado, a pesar de tener un buen desempeño en la tarea de clasificación en particular, el modelo QMC-SGD en este caso no tiene métricas competitivas en calibración (Tabla 5-6).

Table 5-6.: Métricas de desempeño y calibración en la partición de prueba para el conjunto de datos Heart y los distintos modelos entrenados

Modelo	Brier score	Log-loss	Exactitud entrenamiento	Exactitud validación	Exactitud prueba	cwECE	ECE dependiente de datos	ECE uniforme
QMC	0.414	0.603	0.814	0.829	0.700	0.145	0.121	0.310
QMC-D	0.366	0.658	0.951	0.829	0.767	0.231	0.206	0.266
QMC-SGD	0.200	0.365	0.881	0.854	0.867	0.152	0.121	0.194
KQM	0.180	0.320	0.841	0.927	0.927	0.189	0.193	0.202
DT	0.277	0.431	0.761	0.854	0.700	0.132	0.117	0.132
kNN	0.139	0.238	0.872	0.878	0.900	0.094	0.094	0.094
LR	0.175	0.304	0.867	0.878	0.867	0.123	0.158	0.211
MLP 1	0.177	0.275	0.907	0.902	0.867	0.122	0.103	0.173
MLP 2	0.196	0.310	0.881	0.854	0.867	0.192	0.097	0.133
NB	0.199	0.290	0.867	0.878	0.867	0.108	0.108	0.090
RF	0.187	0.322	0.929	0.902	0.867	0.148	0.128	0.195
SVM	0.209	0.371	0.929	0.951	0.900	0.214	0.183	0.264

En general, en los conjuntos de datos binarios, los modelos cuánticos tienen resultados competitivos en la calibración según la tarea de clasificación. En particular, QMC-SGD tiende a tener resultados superiores en la exactitud en comparación con algunos de los clasificadores de referencia, y solo en el conjunto de datos de Heart (Tabla 5-6) no refleja un modelo calibrado. Adicionalmente, el clasificador KQM tiene un comportamiento similar, pero en este caso depende mucho de la función *kernel* utilizada donde se observa que los *kernel* Gaussiano y polinomial tienen los mejores resultados, además de estar calibrados. No obstante, QMC y QMC-D no reflejan los mismos resultados, y en particular, para los tres conjuntos de datos binarios, el QMC tiene un rendimiento deficiente en la exactitud y las métricas de calibración.

Análisis visual de calibración

Para ilustrar y comparar algunas de las propiedades de la calibración de los clasificadores cuánticos con los modelos de referencia, se usaron los diagramas de fiabilidad basados en el ECE usando el esquema de particiones uniforme y el cwECE como se menciona en [38]. De esta forma, para comparar los métodos cuánticos se seleccionaron los modelos de referencia que tienden a tener un comportamiento mejor de calibración y exactitud, esto es, regresión

logística, perceptrón de dos capas y *random forest*.

De esta forma, en la Figura 5-1 observamos para el QMC en el conjunto de datos Adult, que las muestras en las que el clasificador no tiene mucha confianza (probabilidades menores a 0.8) tienden a tener poca exactitud, mientras que para las probabilidades más altas tiende a estar por encima de la diagonal, lo que indica un modelo demasiado confiado en este segmento. En contraste, cuando se aplica QMC-D todas las barras en el diagrama de fiabilidad tienden a estar más cerca de la diagonal, incluso para aquellas confianzas cercanas al 0.5 donde el clasificador puede tener la mayor incertidumbre sobre las predicciones (Figura 5-1).

Adicionalmente, con base en el diagrama de fiabilidad de la Figura 5-1 cuando se aplica la estrategia QMC-SGD, las probabilidades tienden a estar más cerca de la diagonal en comparación con QMC y QMC-D. A su vez, las respectivas métricas de calibración son más bajas, lo que indica que algunos de los problemas que el clasificador QMC-D puede tener en relación con la calibración, como las predicciones con exceso de confianza o exactitud, se resuelven con la estrategia QMC-SGD.

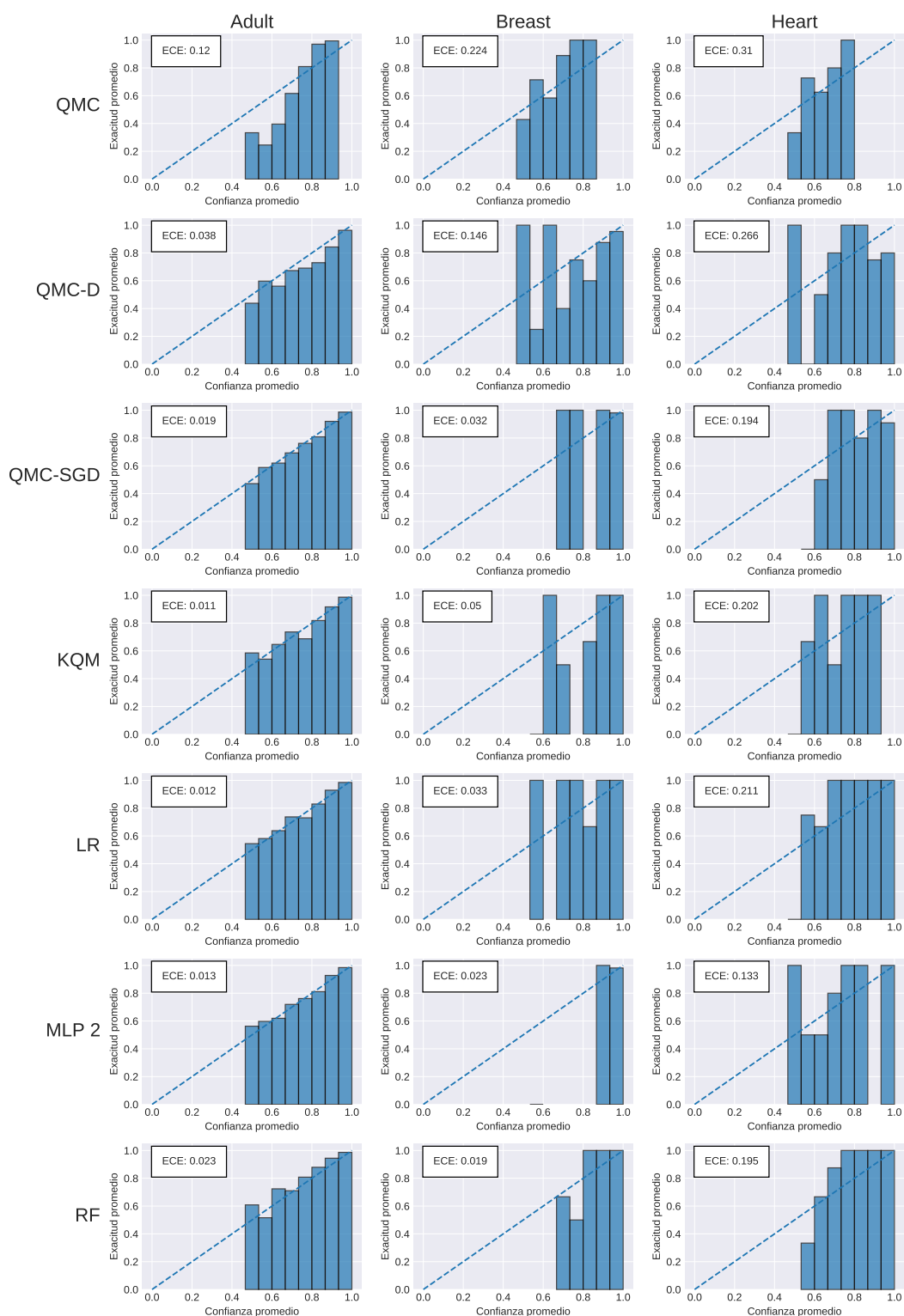
Lo anterior se espera debido al proceso de optimización que se lleva a cabo en el QMC-D y no en el QMC que toma los cálculos directos de la matriz de densidad, por lo tanto, dicha optimización mejora algunas de las propiedades de calibración en la estrategia QMC. Si bien, no en todos los conjuntos de datos se observa el mismo comportamiento, esto puede deberse a las pocas muestras que suelen tener en la partición de prueba algunos conjuntos de datos, lo que dificulta el cálculo de los diagramas de fiabilidad.

En la Figura 5-1 se aprecia que para el modelo KQM en el conjunto de datos Adult, el diagrama de fiabilidad es muy parecido al del QMC-D, aunque el ECE para el KQM es más bajo incluso muy cercano al QMC-SGD, reflejando así un mejor comportamiento en cuanto a la calibración.

Al comparar los diagramas de fiabilidad presentados en la Figura 5-1 para los modelos de línea base, se observa que en general tienden a estar muy cerca de la diagonal, y sus ECE también tienden a ser bajos a comparación de los métodos cuánticos. No obstante, tanto el clasificador KQM como el QMC-SGD tienen métricas similares e incluso inferiores, y visualmente, los diagramas de fiabilidad son parecidos y cercanos a la diagonal.

Por otro lado, para el conjunto de datos Breast, se aprecia que los diagramas de fiabilidad de la Figura 5-1 tienden a ser muy irregulares y con confianzas promedio superiores a 0.7 para la mayoría de modelos. El ECE para dichos modelos es similar, a diferencia de QMC y QMC-D que tienen errores de calibración altos y sus diagramas de calibración son más dispersos, evidenciando un rango más variado de confianza en sus estimaciones (Figura 5-1).

Figure 5-1.: Diagramas de fiabilidad basados en el ECE usando el esquema de particiones uniforme para los conjuntos de datos binarios (Adult, Breast, y Heart) en cada columna y algunos de los clasificadores en cada fila (QMC, QMC-D, QMC-SGD, KQM, regresión logística, MLP con 2 capas, y *random forest*)

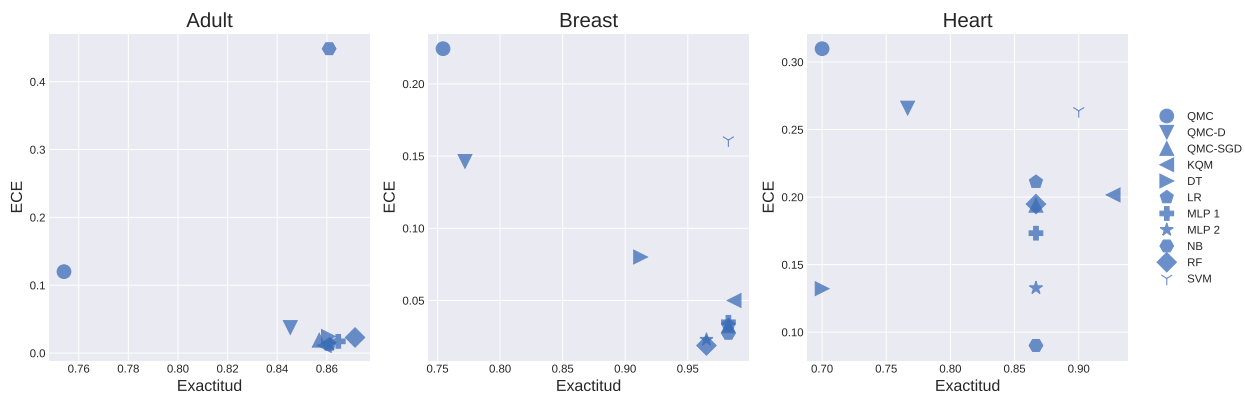


De forma similar, en la Figura 5-1 se observan los diagramas de fiabilidad para el conjunto de datos Heart, donde el *random forest* y el QMC son los que evidencian visualmente un mejor comportamiento de calibración, dado que se encuentran cercanos a la diagonal, y no muestran una concentración de confianza promedio alrededor de 1. Adicionalmente, estos dos modelos son los que reflejan algunas de los ECE más bajos entre todos los clasificadores.

Al analizar los diagramas de fiabilidad para el método QMC y los tres conjuntos de datos binarios (primera fila Figura 5-1), se tiene que aunque el ECE para dicho modelo no siempre es el más bajo entre los clasificadores, sí tiende a presentar un comportamiento cercano a la diagonal. Esto, en contraste con el resto de clasificadores que tienden a tener confianzas promedio muy irregulares y concentradas.

Finalmente, en la Figura 5-2 se presenta el diagrama de dispersión entre el ECE usando el esquema de particiones uniforme y la exactitud en el conjunto de prueba. Se aprecia que para los conjuntos de datos Adult y Breast, la mayoría de modelos presentan un error de calibración esperado y exactitud similar, y los clasificadores QMC y QMC-D tienden estar más alejados.

Figure 5-2.: Diagrama de dispersión entre el ECE usando el esquema de particiones uniforme y la exactitud en el conjunto de prueba para los conjuntos de datos binarios (Adult, Breast, y Heart) y los modelos entrenados



Por el contrario, la mayoría de resultados para el conjunto de datos Heart se encuentran distribuidos a lo largo del diagrama de dispersión, evidenciando así mayor variabilidad en el ECE y exactitud como se muestra en la Figura 5-2. No obstante, varios modelos cuentan con una exactitud alta, a pesar de no tener un error esperado de calibración bajo para este conjunto de datos.

5.3.2. Clasificación de múltiples clases

Métricas de calibración y desempeño

Los resultados de la comparación de los clasificadores de referencia y los métodos cuánticos para el conjunto de datos Dermatology se muestran en la Tabla 5-7, donde QMC-D, la regresión logística, *naive Bayes*, SVM y el *random forest* tienen los mejores resultados en la exactitud en la partición de prueba, y el clasificador *naive Bayes* refleja una calibración perfecta debido al valor cero aproximado en el *Brier score*, *log-loss*, cwECE, ECE con esquema uniforme y dependiente de datos. En particular, QMC-D y la regresión logística muestran buenos resultados en las métricas de calibración. A pesar de que KQM no tiene los mejores resultados, la exactitud es competitiva en comparación con los mejores, y su calibración es buena de acuerdo con el *Brier score*, *log-loss*, cwECE y ECE.

Por otro lado, como se muestra en la Tabla 5-7, el MLP, el árbol de decisión y el QMC-SGD no reflejan los peores resultados en términos de exactitud, incluso los clasificadores MLP con una o dos capas tienen algunos de los resultados más bajos en las métricas de calibración. Este no es el caso de los modelos kNN y QMC que muestran malos resultados en exactitud y calibración. Note que varias métricas en la Tabla 5-7 son exactamente iguales para los diferentes modelos, esto puede ocurrir debido a la pequeña cantidad de muestras en la partición de prueba del conjunto de datos.

Table 5-7.: Métricas de desempeño y calibración en la partición de prueba para el conjunto de datos Dermatology y los distintos modelos entrenados

Modelo	Brier score	Log-loss	Exactitud entrenamiento	Exactitud validación	Exactitud prueba	cwECE	ECE dependiente de datos	ECE uniforme
QMC	0.388	0.813	0.941	0.939	0.889	0.161	0.417	0.442
QMC-D	0.061	0.145	1.000	0.980	0.972	0.036	0.083	0.067
QMC-SGD	0.061	0.121	1.000	0.959	0.944	0.028	0.070	0.056
KQM	0.061	0.111	1.000	0.959	0.959	0.023	0.072	0.061
DT	0.111	1.921	0.996	0.980	0.944	0.019	0.054	0.058
kNN	0.082	0.144	0.989	0.980	0.917	0.036	0.104	0.104
LR	0.059	0.160	1.000	0.980	0.972	0.041	0.098	0.098
MLP 1	0.057	0.091	1.000	1.000	0.944	0.018	0.052	0.033
MLP 2	0.070	0.100	1.000	1.000	0.944	0.018	0.051	0.051
NB	0.000	0.000	1.000	0.980	0.972	0.000	0.000	0.000
RF	0.074	0.170	1.000	1.000	0.972	0.044	0.129	0.097
SVM	0.178	0.411	0.989	0.959	0.972	0.100	0.291	0.291

Como se presenta en la Tabla 5-8 para el conjunto de datos de Glass, el mejor clasificador con base en la exactitud son los árboles de decisión que también tienen las métricas de calibración más bajas. No obstante, la mayoría de los modelos de referencia tienen la segunda mejor precisión en la partición de prueba con un comportamiento similar en calibración.

En este caso, los modelos cuánticos no reflejan los mejores resultados en la tarea de clasificación, pero el KQM aún logra dar resultados competitivos. Por otro lado, como se muestra en la Tabla 5-8, QMC-SGD tiene una exactitud baja, y no muestra métricas de calibración muy altas, mientras que QMC tiene un rendimiento deficiente en exactitud y calibración, y QMC-D, tiene una exactitud baja y no tiene las métricas de calibración más altas.

Table 5-8.: Métricas de desempeño y calibración en la partición de prueba para el conjunto de datos Glass y los distintos modelos entrenados

Modelo	Brier score	Log-loss	Exactitud entrenamiento	Exactitud validación	Exactitud prueba	cwECE	ECE dependiente de datos	ECE uniforme
QMC	0.438	0.952	0.893	0.800	0.743	0.122	0.210	0.212
QMC-D	0.245	0.573	1.000	0.900	0.800	0.071	0.111	0.111
QMC-SGD	0.204	0.533	0.993	0.867	0.857	0.071	0.109	0.127
KQM	0.290	0.847	0.926	0.900	0.900	0.057	0.137	0.136
DT	0.057	0.987	1.000	1.000	0.971	0.010	0.029	0.029
kNN	0.170	0.270	1.000	1.000	0.857	0.050	0.110	0.110
LR	0.163	0.614	1.000	0.967	0.943	0.053	0.160	0.171
MLP 1	0.150	0.333	1.000	0.967	0.943	0.050	0.140	0.162
MLP 2	0.105	0.683	1.000	1.000	0.943	0.030	0.037	0.063
NB	0.146	0.275	1.000	0.967	0.943	0.033	0.071	0.055
RF	0.175	0.375	0.993	1.000	0.943	0.077	0.216	0.216
SVM	0.240	0.487	0.987	0.967	0.914	0.098	0.240	0.316

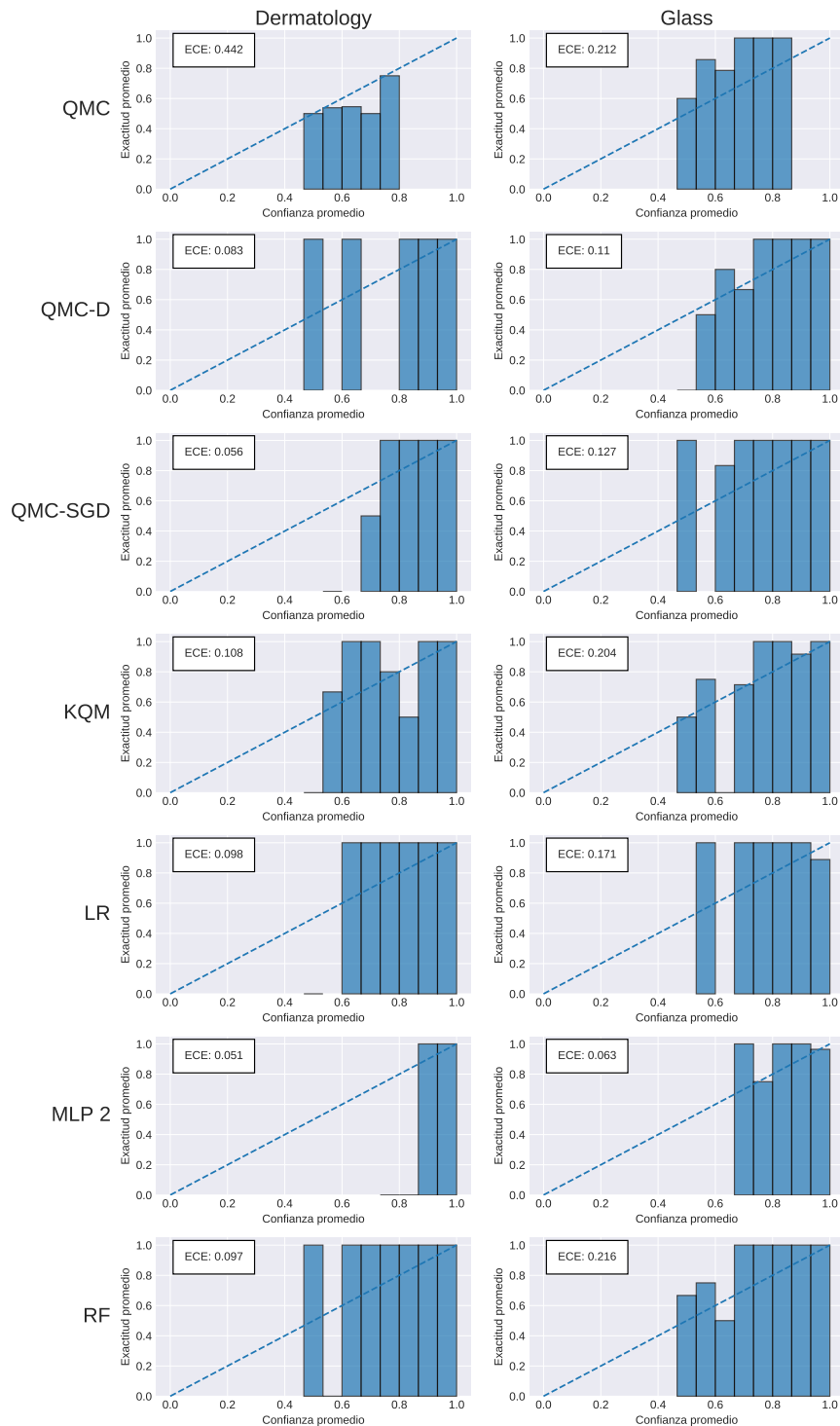
En general, los modelos cuánticos tienden a tener un mejor desempeño en Dermatology a comparación del conjunto de datos Glass, incluso para el primero, QMC-D tiene los mejores resultados en exactitud y está en los clasificadores con métricas más bajas de calibración. Sin embargo, este modelo tiende a mostrar un comportamiento volátil debido a que para una de las tareas de múltiples clases tiene un buen desempeño, pero en la otra no lo tiene, y lo mismo ocurre con las tareas binarias.

Por otro lado, KQM-Coseno y QMC muestran consistentemente un pobre desempeño en calibración y exactitud, mientras que KQM con *kernel* polinomial o Gaussiano tienden a reflejar buenos resultados tanto en calibración como en exactitud. De igual forma, el QMC-SGD a pesar de no estar en los mejores de exactitud siempre, muestra un comportamiento competitivo y una buena calibración.

Análisis visual de calibración

En la Figura 5-3, para el conjunto de datos Glass no hay una gran diferencia entre los diagramas de fiabilidad de QMC y QMC-D basados en ECE, porque para ambos clasificadores las probabilidades mayores a 0.8 tienden a estar sobre la diagonal indicando mala calibración.

Figure 5-3.: Diagramas de fiabilidad basados en el ECE usando el esquema de particiones uniforme para los conjuntos de datos de múltiples clases (Dermatology y Glass) en cada columna y algunos de los clasificadores en cada fila (QMC, QMC-D, QMC-SGD, KQM, regresión logística, MLP con 2 capas, y *random forest*)



Sin embargo, para QMC-D, KQM y *random forest* las probabilidades inferiores a 0.8 están más cerca de la diagonal que las de QMC, lo que refleja también un ECE más bajo que ilustra una mejor calibración.

A pesar que en este caso, el MLP con dos capas es el modelo que tiene un ECE más bajo entre los clasificadores presentados en la Figura 5-3, visualmente no refleja un diagrama de fiabilidad que se encuentre cercano a la diagonal en todos los casos. Incluso, la exactitud promedio para la mayoría de barras es muy cercana a uno, mientras que su confianza no necesariamente está cercana a estos valores. Por esta razón, es importante hacer un análisis de calibración de los diagramas de fiabilidad apoyado con métricas que puedan cuantificar el error de calibración.

A diferencia del MLP con dos capas, el KQM tiene el segundo ECE más bajo y su diagrama de fiabilidad, como se presenta en la Figura 5-3, evidencia un comportamiento más cercano a la diagonal y que visualmente refleja una mejor calibración.

Por otro lado, los diagramas de fiabilidad para el conjunto de datos Glass que se encuentran en la Figura 5-3 evidencian un comportamiento muy irregular en todos los casos, a excepción del QMC que presenta confianzas promedio cercanas a sus exactitudes, lo cual sumado al hecho que tiene el segundo ECE más bajo, indica una buena calibración del modelo.

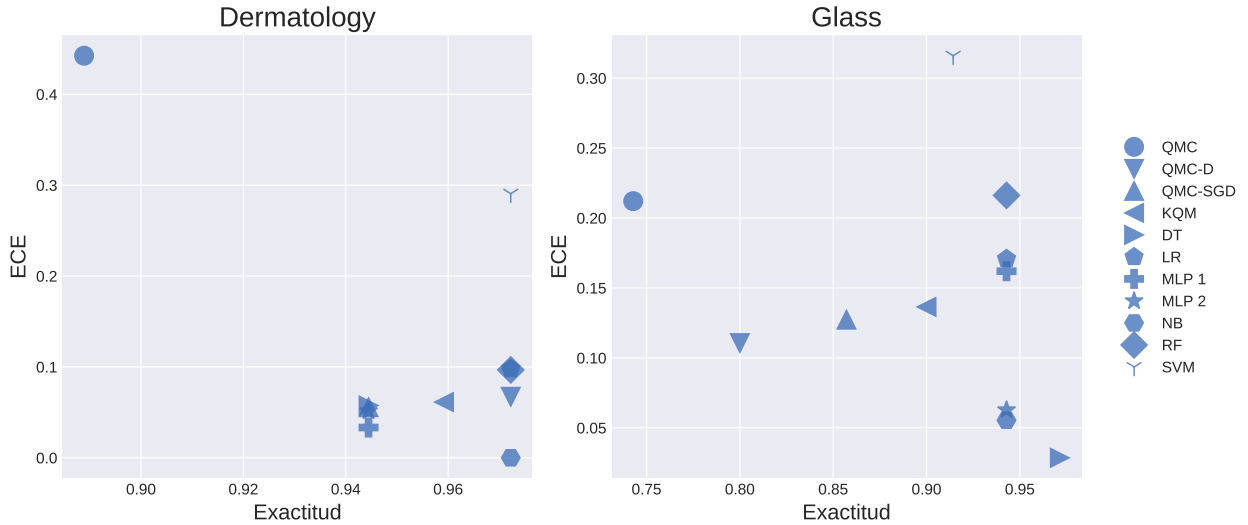
Mientras tanto, los diagramas de fiabilidad de los tres clasificadores de referencia presentados en la Figura 5-3 visualmente no presentan un buen comportamiento de calibración, ya que la mayoría de confianzas corresponden con exactitudes promedio cercanas a 1, a pesar de que el MLP de dos capas tiene el ECE más bajo de los modelos comparados en la gráfica.

Dicha característica se observa tanto para el conjunto de datos Dermatology como Glass, dado que estos tres clasificadores de referencia gráficamente no presentan valores cercanos a la diagonal.

Los diagramas de fiabilidad del QMC para ambos conjuntos de datos de múltiples clases evidencian un comportamiento cercano a la diagonal (primera fila Figura 5-3), similar a lo analizado en los datos binarios (Figura 5-1). No obstante, para Glass se tiene que la exactitud tiende a ser mayor a la confianza promedio para dicho modelo.

Por último, en la Figura 5-4 se aprecia el diagrama de dispersión del ECE con esquema uniforme y la exactitud en el conjunto de prueba, donde los resultados para el conjunto de datos Dermatology se encuentran más concentrados en exactitudes altas y un error de calibración esperado bajo para la mayoría de modelos, incluyendo los clasificadores cuánticos, solo QMC y SVM se alejan de la aglomeración de puntos.

Figure 5-4.: Diagrama de dispersión entre el ECE usando el esquema de particiones uniforme y la exactitud en el conjunto de prueba para los conjuntos de datos de múltiples clases (Dermatology y Glass) y los modelos entrenados



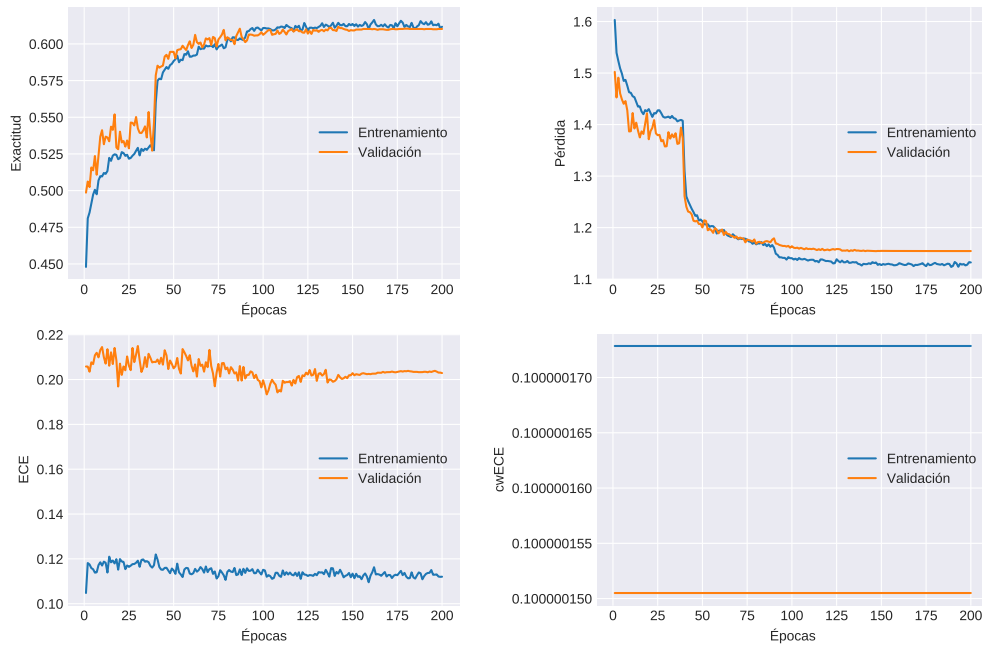
En contraste, para el conjunto de datos Glass se aprecia una mayor dispersión al comparar la exactitud y el ECE con esquema uniforme (Figura 5-4). En particular, es interesante notar que a excepción de QMC, la exactitud para los modelos cuánticos no supera 0.90 y todos los clasificadores de referencia están por encima, pero el ECE para los modelos cuánticos QMC-D, QMC-SGD y KQM tienen una mejor calibración.

Lo anterior indica que los modelos de aprendizaje de máquina cuántico utilizados a pesar de que en no todos los casos brindan las mejores métricas de exactitud, tienen buenas propiedades de calibración cuando se comparan con modelos ampliamente utilizados en la literatura de aprendizaje de máquina.

5.3.3. LeNet5-QMDM

Observamos en la Figura 5-5 las curvas de aprendizaje para el ECE usando el esquema uniforme y el cwECE para el modelo LeNet5-QMDM, donde ambas métricas muestran un comportamiento aproximadamente constante, a pesar de aumentar la capacidad discriminativa de la red a medida que aumentan las épocas. Lo cual indica que la adición de la capa cuántica no necesariamente ayuda a mejorar la calibración de la red del clasificador, aunque curvas similares se evidencian para el entrenamiento del modelo LeNet5 sin la adición de la capa cuántica (Apéndice A).

Figure 5-5.: Curvas de aprendizaje para la partición de entrenamiento y validación para la exactitud (izquierda-superior), función de pérdida (derecha-superior), ECE con esquema uniforme (izquierda-inferior) y cwECE (derecha-inferior) en el conjunto de datos CIFAR-10 con el clasificador LeNet5-QMDM



Por otro lado, en la Figura 5-5 las curvas de aprendizaje para la exactitud y pérdida para el modelo LeNet5-QMDM, muestran una forma estándar y convergencia, con los resultados de validación ligeramente por encima o debajo de la curva de entrenamiento respectivamente.

Las métricas de calibración para la estrategia LeNet5-QMDM no muestran una mejora en la calibración, incluso esas métricas (*Brier score*, *log-loss*, *cwECE*, ECE con esquema uniforme y dependiente de datos de ECE) son ligeramente más altas (Tabla 5-9).

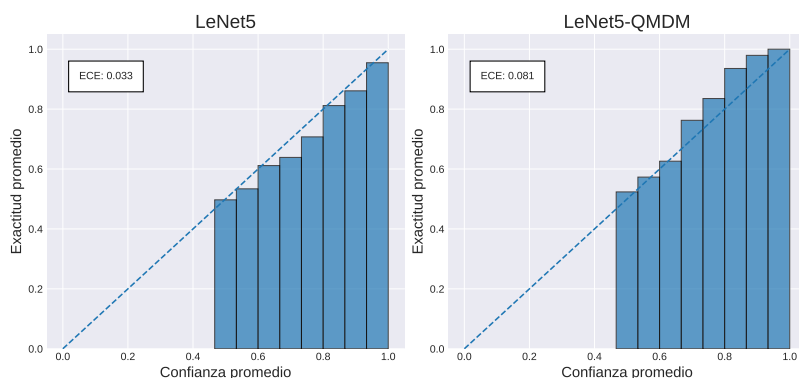
Table 5-9.: Métricas de desempeño y calibración en la partición de prueba para el conjunto de datos CIFAR-10 y los modelos LeNet5 y LeNet5-QMDM

Modelo	Brier score	Log-loss	Exactitud entrenamiento	Exactitud validación	Exactitud prueba	cwECE	ECE dependiente de datos	ECE uniforme
LeNet5	0.520	1.415	0.652	0.614	0.617	0.012	0.034	0.034
LeNet5-QMDM	0.523	1.153	0.639	0.610	0.613	0.020	0.084	0.081

De forma similar, en la Figura 5-6 se muestran los diagramas de fiabilidad basados en el ECE usando el esquema de particiones uniforme para ambas arquitecturas, donde el modelo LeNet5-QMDM presenta valores por encima de la diagonal, lo cual coincide con las métricas

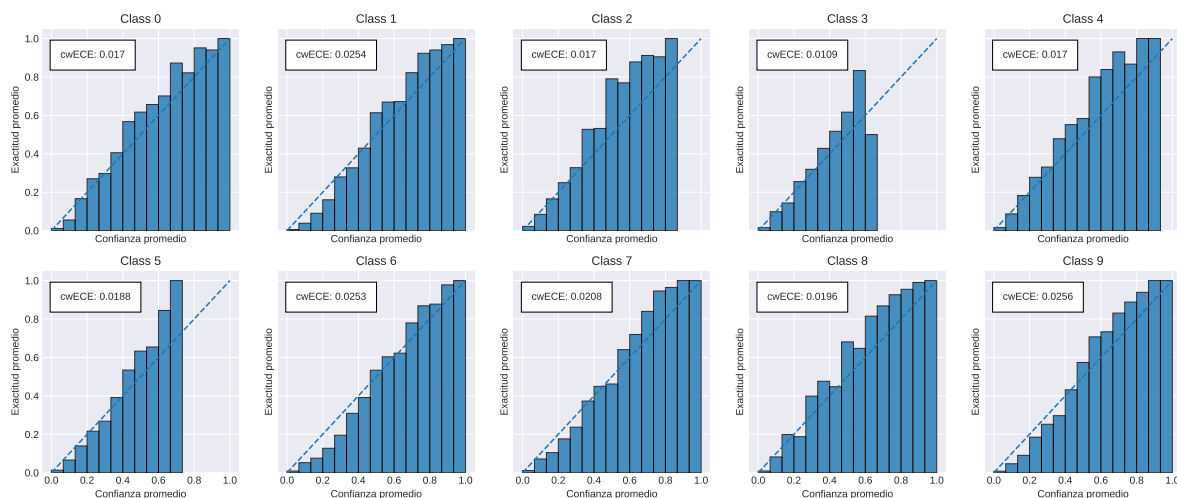
de calibración de la Tabla 5-9, ya que la arquitectura LeNet5 sin la capa cuántica de QMDM muestra un mejor comportamiento en cuanto a calibración.

Figure 5-6.: Diagramas de fiabilidad basados en el ECE usando el esquema de particiones uniforme para los modelos LeNet5 y LeNet5-QMDM en el conjunto de imágenes CIFAR-10



Al comparar los diagramas de fiabilidad por clases de LeNet5-QMDM y LeNet5 en la Figura 5-7, se observa un aumento en la exactitud promedio de la mayoría de las predicciones de confianza (superior a 0.6), lo que indica un modelo con falta de confianza para la mayoría de las clases, es decir, un modelo cuyas probabilidades tienden a ser menores que la exactitud promedio del fenómeno. No obstante, cuando el modelo refleja una baja probabilidad de clasificación se aprecia que la exactitud que refleja es casi idéntica, evidenciando así que el modelo es capaz de brindar probabilidades que representan la confianza que el clasificador tiene en la predicción.

Figure 5-7.: Diagramas de fiabilidad para las 10 clases del conjunto de datos CIFAR-10 con el clasificador LeNet5-QMDM



Finalmente, para este caso no se presentan los diagramas de fiabilidad basados en el ECE con esquema uniforme dado que para ambas arquitecturas de redes neuronales se aprecian comportamientos muy similares. Lo cual sucede también para los diagramas de fiabilidad por clases, pero dado que este conjunto de datos cuenta con varias muestras y varias clases, es interesante ver la utilidad de esta herramienta de análisis de calibración.

5.4. Discusión

En la exploración de la calibración de los clasificadores cuánticos, se observa que algunos de estos modelos pueden alcanzar resultados competitivos o mejores en términos de exactitud a comparación de algunos de los modelos de referencia en tareas binarias o de múltiples clases, además de contar con diagramas de fiabilidad y métricas de calibración que indican clasificadores bien calibrados.

Por lo tanto, las probabilidades predichas para todas las clases objetivo de estos métodos podrían interpretarse correctamente como la confianza que esos modelos tienen en sus predicciones, lo cual es relevante en aplicaciones críticas para la seguridad y ofrece una herramienta para interpretar los resultados de los clasificadores que en algunos casos no son intuitivos.

En particular, QMC-SGD y KQM con *kernel* Gaussiano y polinomial muestran resultados más consistentes en las diferentes tareas de clasificación en comparación con QMC y QMC-D, donde este último presenta un comportamiento más volátil en su calibración y propiedades discriminatorias. Por el contrario, QMC no evidencia resultados muy competitivos en cuanto a la exactitud y en varios casos a las métricas de calibración, pero en los diagramas de fiabilidad tiende a mostrar resultados que se encuentran cerca a la diagonal, lo cual indica que no necesariamente tiene malas propiedades de calibración.

Lo anterior se puede presentar porque QMC-SGD y KQM a comparación de QMC-D, no reducen la dimensionalidad del conjunto de datos al realizar la factorización de la matriz de densidad, lo cual puede acelerar la ejecución de los modelos, pero puede llevar a perder información útil para la detección de las clases objetivo. Mientras que QMC a pesar de no aplicar la factorización de la matriz de densidad, no realiza un proceso de optimización de los parámetros involucrados en el modelo, lo cual mejora la habilidad discriminatoria y como se observó en los resultados, las propiedades de calibración.

De esta forma, tanto los conjuntos de datos que tienen muy pocos registros en la partición de prueba, la dificultad inherente a las tareas de clasificación y las fallas que pueden tener las herramientas utilizadas para medir la calibración, pueden ser algunos de los factores que influyen en la variabilidad de los resultados de los modelos. Causando por ejemplo, que modelos como QMC-SGD que tiende a tener un buen desempeño en la tarea de clasificación

para los distintos conjuntos de datos, para Heart no tenga el mismo desempeño competitivo.

Para el experimento de la red neuronal LeNet5 en combinación con QMDM, aunque las propiedades de calibración de la arquitectura original no mejoran al agregar la capa cuántica QMDM, se tiene que el desempeño en la tarea de clasificación no empeora y se mantiene aproximadamente constante en ambos modelos. De esta forma, es importante analizar si esto sucede para el conjunto de imágenes CIFAR-10 y la arquitectura LeNet5 en particular, o por el contrario, para otros conjuntos de datos y arquitecturas se puede evidenciar un comportamiento distinto.

Finalmente, la volatilidad en los resultados de los modelos entre conjuntos de datos para tareas binarias y de múltiples clases, a parte de deberse a la dificultad inherente que tiene la tarea de clasificación, se puede presentar por el poco número de registros que se tienen en las particiones de prueba, lo cual puede causar problemas en la estimación del error de calibración como se menciona en [54, 42, 37].

6. Conclusiones y trabajo futuro

El análisis de calibración en aprendizaje de máquina es relevante en la aplicación de clasificadores entrenados para problemas de la vida real donde se desea tener conocimiento de la incertidumbre que tienen los modelos en sus predicciones, por ejemplo, aplicaciones críticas para la seguridad, diagnóstico de enfermedades, conducción autónoma, entre otros. Además, el aprendizaje de máquina cuántico ha ganado relevancia en los últimos años debido a los resultados competitivos que ha logrado en comparación con los modelos de ampliamente utilizados y la mejora del costo de tiempo de algunos de los algoritmos [50, 30, 52].

Este trabajo presenta un análisis y exploración de las propiedades de calibración que tienen los modelos cuánticos descritos en [28, 21, 52] y su comparación con modelos de clasificación ampliamente usados en tareas binarias y de múltiples clases. En general, los resultados obtenidos por los métodos cuánticos tienden a ser competitivos con los modelos de referencia, e incluso en algunos casos, tanto la exactitud como las métricas de calibración son superiores. En particular, cuando los clasificadores cuánticos tienden a tener mejores propiedades de calibración, el desempeño en la tarea de clasificación es mejor. No obstante, hay clasificadores como QMC y KQM con función de *kernel* coseno que muestran consistentemente un peor desempeño en comparación con los otros métodos y QMC-D que en algunos de los conjuntos de datos tiene una alta exactitud (Adult y Dermatology), pero no necesariamente está calibrado.

En contraste, KQM con *kernel* Gaussiano y polinomial, y QMC-SGD tienen los mejores resultados entre los clasificadores cuánticos incluso superando en el conjunto de datos Breast y Heart la exactitud de los métodos de referencia y al mismo tiempo mantienen propiedades buenas de calibración. Sin embargo, para la estrategia KQM, la selección de la función *kernel* afecta fuertemente los resultados, ya que en algunos casos el *kernel* Gaussiano presenta un buen desempeño en comparación con el polinomial, y viceversa, tanto para las métricas de calibración, desempeño y diagrama de fiabilidad, lo que indica la importancia de una buena representación de los datos.

A pesar de que con los clasificadores QMC y QMC-D no se obtienen métricas competitivas de calibración o discriminación en la mayoría de los casos, observamos que la calibración del QMC mejora al aplicar la estrategia QMC-D. En comparación con estos enfoques, QMC-SGD tiende a mejorar aún más las propiedades de calibración de los dos métodos anteriores

mientras genera mejores métricas de discriminación.

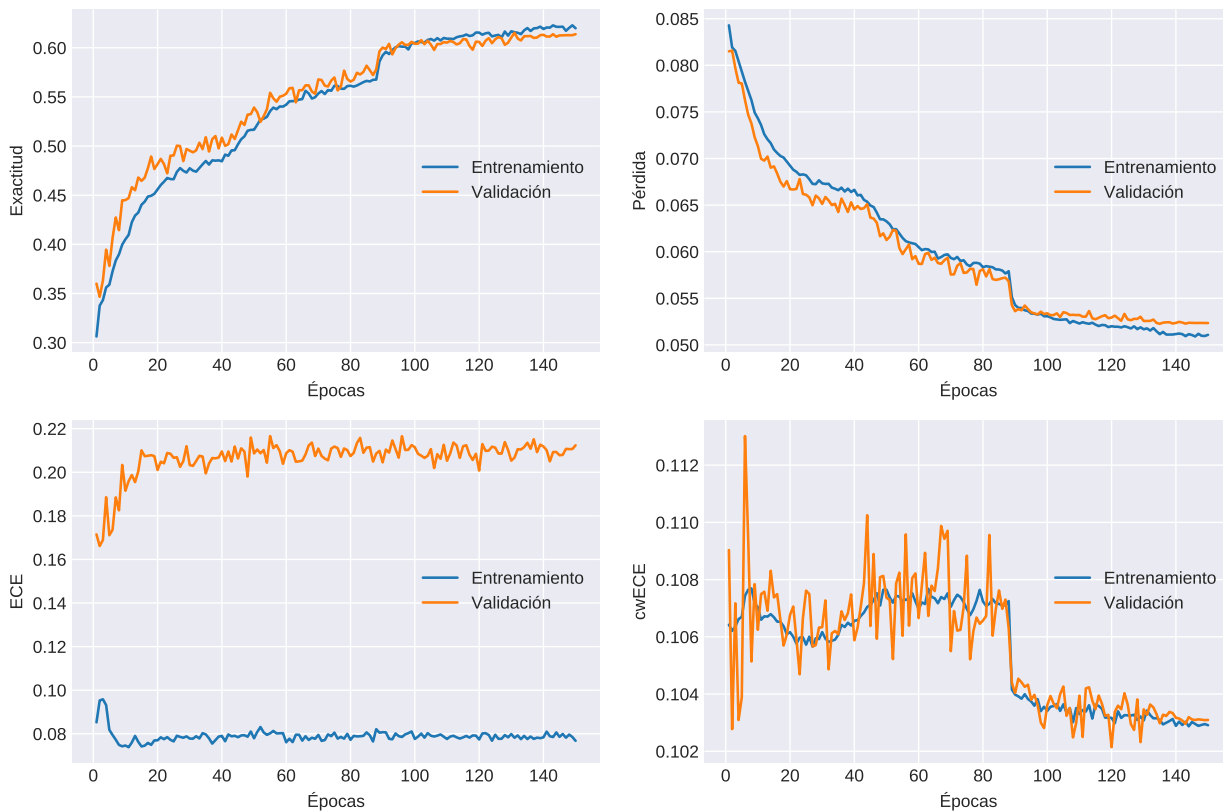
En cuanto al experimento realizado con LeNet5 no se evidencia que la capa cuántica QMDM mejore las métricas de calibración de la red neuronal, incluso al analizar la calibración por clase, se tiene que empeora en la mayoría de casos. No obstante, la exactitud se mantiene aproximadamente igual entre ambas arquitecturas, indicando que a pesar de que la capa de QMDM no mejora las propiedades de calibración, no empeora el desempeño en la tarea de clasificación de la red original.

Como trabajo futuro, debido a la sensibilidad de los diagramas de confiabilidad con respecto a los tamaños de muestra, se pueden proponer experimentos similares a los realizados, pero con conjuntos de datos con más muestras para identificar mejor las diferencias que pueden existir entre QMC, QMC-D, QMC-SGD, y KQM. Mientras tanto, para los experimentos de redes neuronales se podría entrenar la arquitectura LeNet con otros conjuntos de datos reales y arquitecturas profundas como las probadas en [24], esto debido a la volatilidad que se observó en los experimentos comparativos en el clasificador QMC-D que es equivalente al combinado con la red neuronal en el conjunto de datos CIFAR-10 (LeNet-QMDM).

A. Apéndice: Curvas de aprendizaje

En este apéndice se presenta la Figura para la exactitud, función de pérdida, ECE usando el esquema dependiente de los datos, y cwECE en las curvas de aprendizaje para la arquitectura entrenada LeNet5 con el conjunto de imágenes de CIFAR-10.

Figure A-1.: Curvas de aprendizaje para la partición de entrenamiento y validación para la exactitud (izquierda-superior), función de pérdida (derecha-superior), ECE con esquema uniforme (izquierda-inferior) y cwECE (derecha-inferior) en el conjunto de datos CIFAR-10 con el clasificador LeNet5



Bibliografía

- [1] Ayhan, Murat S. ; Berens, Philipp: Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. **In:** *Medical Imaging with Deep Learning (MIDL)*, 2018, S. 1–9
- [2] Aha, D ; Kibler, Dennis: Instance-based prediction of heart-disease presence with the Cleveland database. *University of California* 3 (1988), Nr. 1, S. 3–2
- [3] Bastola, S. ; Ishidaira, H. ; Takeuchi, K.: Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe. *Journal of Hydrology* 357 (2008), Nr. 3-4, S. 188–206
- [4] Beaudouin, R. ; Monod, G. ; Ginot, V.: Selecting parameters for calibration via sensitivity analysis: An individual-based model of mosquitofish population dynamics. *Ecological Modelling* 218 (2008), Nr. 1-2, S. 29–48
- [5] Brier, Glenn W.: Verification of forecast expressed in terms of probability. *Monthly Weather Review* 78 (1950), jan, Nr. 1, S. 1–3. – ISSN 0027–0644
- [6] Bröcker, Jochen ; Smith, Leonard A.: Increasing the reliability of reliability diagrams. *Weather and Forecasting* 22 (2007), jun, Nr. 3, S. 651–661. – ISSN 08828156
- [7] Biamonte, Jacob ; Wittek, Peter ; Pancotti, Nicola ; Rebentrost, Patrick ; Wiebe, Nathan ; Lloyd, Seth: Quantum machine learning. *Nature* 549 (2017), Nr. 7671, S. 195–202
- [8] Chollet, François ; others. Keras. <https://keras.io>. 2015
- [9] Charoenpanyanet, A.: Modeling Anopheles mosquito density spatial and seasonal variations using remotely sensed imagery and statistical methods. *International Journal of Geoinformatics* 13 (2017), Nr. 1, S. 35–47
- [10] Carneiro, G. ; Zorron Cheng Tao Pu, L. ; Singh, R. ; Burt, A.: Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical Image Analysis* 62 (2020)
- [11] Degroot, M. D. ; Fienberg, Stephen E. The comparison and evaluation of forecasters. mar 1982

-
- [12] Dua, Dheeru ; Graff, Casey. UCI Machine Learning Repository. 2017
- [13] Demiroz, G ; Govenir, HA ; Ilter, N: Learning differential diagnosis of erythematous diseases using voting feature intervals. *Artificial Intelligence in Medicine* 13 (1998), Nr. 3, S. 147–165
- [14] Diego Hernando, Useche R.: Quantum measurement learning for medical image classification. (2022)
- [15] Detrano, Robert ; Janosi, Andras ; Steinbrunn, Walter ; Pfisterer, Matthias ; Schmid, Johann-Jakob ; Sandhu, Sarbjit ; Guppy, Kern H. ; Lee, Stella ; Froelicher, Victor: International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology* 64 (1989), Nr. 5, S. 304–310
- [16] Dormann, Carsten F.: Calibration of probability predictions from machine-learning and statistical models. *Global Ecology and Biogeography* 29 (2020), apr, Nr. 4, S. 760–765. – ISSN 14668238
- [17] Deng, Y. ; Yang, B.-R. ; Luo, J.-W. ; Du, G.-X. ; Luo, L.-P.: DTI-based radiomics signature for the detection of early diabetic kidney damage. *Abdominal Radiology* 45 (2020), Nr. 8, S. 2526–2531
- [18] Feng, Runhai: Improving uncertainty analysis in well log classification by machine learning with a scaling algorithm. *Journal of Petroleum Science and Engineering* 196 (2021). – ISSN 09204105
- [19] Foster, Dean P. ; Vohra, Rakesh V.: Asymptotic calibration. *Biometrika* 85 (1998), Nr. 2, S. 379–390. – ISSN 00063444
- [20] Galindo, Y. ; De Cicco, M. ; Quiles, M.G. ; Lorena, A.C.: *Monitoring Night Skies with Deep Learning*. Bd. 1332. 2020 460–468 Seiten. – ISBN 9783030638191
- [21] González, Fabio A. ; Gallego, Alejandro ; Toledo-Cortés, Santiago ; Vargas-Calderón, Vladimir: Learning with Density Matrices and Random Features. *arXiv preprint arXiv:2102.04394* (2021)
- [22] Gallego-Mejia, Joseph ; Bustos-Brinez, Oscar ; Gonzalez, Fabio: InQMAD: Incremental Quantum Measurement Anomaly Detection. *arXiv preprint arXiv:2210.05061* (2022)
- [23] Glasser, Ivan ; Pancotti, Nicola ; Ignacio Cirac, J.: From Probabilistic Graphical Models to Generalized Tensor Networks for Supervised Learning. *IEEE Access* 8 (2020), S. 68169–68182. – ISSN 21693536
- [24] Guo, Chuan ; Pleiss, Geoff ; Sun, Yu ; Weinberger, Kilian Q. On calibration of modern neural networks. 2017

- [25] Gupta, Kartik ; Rahimi, Amir ; Ajanthan, Thalaiyasingam ; Mensink, Thomas ; Sminchisescu, Cristian ; Hartley, Richard: Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800* (2020), 6
- [26] Guan, Yawen ; Sampson, Christian ; Tucker, J. D. ; Chang, Won ; Mondal, Anirban ; Haran, Murali ; Sulsky, Deborah: Computer Model Calibration Based on Image Warping Metrics: An Application for Sea Ice Deformation. *Journal of Agricultural, Biological, and Environmental Statistics* 24 (2019), Nr. 3, S. 444–463. – ISSN 15372693
- [27] Ghoshal, Biraja ; Tucker, Allan: On calibrated model uncertainty in deep learning. *arXiv preprint arXiv:2206.07795* (2022)
- [28] González, Fabio A. ; Vargas-Calderón, Vladimir ; Vinck-Posada, Herbert: Classification with Quantum Measurements. *Journal of the Physical Society of Japan* 90 (2021), Nr. 4, S. 044002
- [29] Higdon, Dave ; Gattiker, James ; Williams, Brian ; Rightley, Maria: Computer model calibration using high-dimensional output. *Journal of the American Statistical Association* 103 (2008), Nr. 482, S. 570–583. – ISSN 01621459
- [30] Iliyasa, Abdullah M. ; Faticah, Chastine: A quantum hybrid PSO combined with fuzzy k-NN approach to feature selection and cell classification in cervical cancer detection. *Sensors (Switzerland)* 17 (2017), Nr. 12. – ISSN 14248220
- [31] Jensen, M.H. ; Jørgensen, D.R. ; Jalaboi, R. ; Hansen, M.E. ; Olsen, M.A.: *Improving uncertainty estimation in convolutional neural networks using inter-rater agreement*. Bd. 11767 LNCS. 2019 540–548 Seiten. – ISBN 9783030322502
- [32] Jiang, Xiaoqian ; Osl, Melanie ; Kim, Jihoon ; Ohno-Machado, Lucila: Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association* 19 (2012), mar, Nr. 2, S. 263–274. – ISSN 10675027
- [33] Kuleshov, Volodymyr ; Ermon, Stefano: Estimating uncertainty online against an adversary. **In:** *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, S. 2110–2116
- [34] Krizhevsky, Alex ; Hinton, Geoffrey ; others: Learning multiple layers of features from tiny images. (2009)
- [35] Kuhn, Max ; Johnson, Kjell: *Applied predictive modeling*. Bd. 26. Springer, 2013
- [36] Kuleshov, Volodymyr ; Liang, Percy: Calibrated structured prediction. **In:** *Advances in Neural Information Processing Systems* Bd. 2015-Janua. Bd. 2015-Janua, 2015. ISSN 10495258, S. 3474–3482

-
- [37] Kumar, Ananya ; Liang, Percy S. ; Ma, Tengyu: Verified uncertainty calibration. *Advances in Neural Information Processing Systems* 32 (2019)
- [38] Kull, Meelis ; Perello Nieto, Miquel ; Kängsepp, Markus ; Silva Filho, Telmo ; Song, Hao ; Flach, Peter. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. 2019
- [39] LeCun, Yann ; Bottou, Léon ; Bengio, Yoshua ; Haffner, Patrick: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86 (1998), Nr. 11, S. 2278–2324
- [40] Müller, R. ; Kornblith, S. ; Hinton, G.: When does label smoothing help? **In:** *Advances in Neural Information Processing Systems* Bd. 32. Bd. 32, 2019
- [41] Naeini, Mahdi P. ; Cooper, Gregory ; Hauskrecht, Milos: Obtaining well calibrated probabilities using bayesian binning. **In:** *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015
- [42] Nixon, Jeremy ; Dusenberry, Michael W. ; Zhang, Linchuan ; Jerfel, Ghassen ; Tran, Dustin: Measuring Calibration in Deep Learning. 2 (2019), Nr. 7
- [43] Niculescu-Mizil, Alexandru ; Caruana, Rich: Predicting good probabilities with supervised learning. **In:** *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, 2005. – ISBN 1595931805, S. 625–632
- [44] Nguyen, Khanh ; O’Connor, Brendan: Posterior calibration and exploratory analysis for natural language processing models. **In:** *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015. – ISBN 9781941643327, S. 1587–1598
- [45] Posocco, Nicolas ; Bonnefoy, Antoine: Estimating Expected Calibration Errors. **In:** *International Conference on Artificial Neural Networks* Springer, 2021, S. 139–150
- [46] Peleg, K.: Fast fourier transform based calibration in remote sensing. *International Journal of Remote Sensing* 19 (1998), Nr. 12, S. 2301–2315
- [47] Pearce, Jennie ; Ferrier, Simon: Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* 133 (2000), Nr. 3, S. 225–245. – ISSN 03043800
- [48] Platt, John ; Others: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10 (1999), Nr. 3, S. 61–74

- [49] Pedregosa, F. ; Varoquaux, G. ; Gramfort, A. ; Michel, V. ; Thirion, B. ; Grisel, O. ; Blondel, M. ; Prettenhofer, P. ; Weiss, R. ; Dubourg, V. ; Vanderplas, J. ; Passos, A. ; Cournapeau, D. ; Brucher, M. ; Perrot, M. ; Duchesnay, E.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), S. 2825–2830
- [50] Sergioli, Giuseppe ; Militello, Carmelo ; Rundo, Leonardo ; Minafra, Luigi ; Torrisi, Filippo ; Russo, Giorgio ; Chow, Keng L. ; Giuntini, Roberto: A quantum-inspired classifier for clonogenic assay evaluations. *Scientific Reports* 11 (2021), Nr. 1, S. 2830. – ISBN 0123456789
- [51] Schuld, Maria ; Sinayskiy, Ilya ; Petruccione, Francesco: An introduction to quantum machine learning. *Contemporary Physics* 56 (2015), Nr. 2, S. 172–185. – ISSN 13665812
- [52] Toledo-Cortés, Santiago ; Useche, Diego H. ; González, Fabio A.: Prostate Tissue Grading with Deep Quantum Measurement Ordinal Regression. *arXiv preprint arXiv:2103.03188* (2021)
- [53] Torres-Meza, M.d.J. ; Báez-González, A.D. ; Maciel-Pérez, L.H. ; Quezada-Guzmán, E. ; Sierra-Tristán, J.S.: GIS-based modeling of the geographic distribution of *Quercus emoryi* Torr. (Fagaceae) in México and identification of significant environmental factors influencing the species' distribution. *Ecological Modelling* 220 (2009), Nr. 24, S. 3599–3611
- [54] Vaicenavicius, Juozas ; Widmann, David ; Andersson, Carl ; Lindsten, Fredrik ; Roll, Jacob ; Schön, Thomas: Evaluating model calibration in classification. **In:** *The 22nd International Conference on Artificial Intelligence and Statistics* PMLR, 2019, S. 3459–3467
- [55] Widmann, David ; Lindsten, Fredrik ; Zachariah, Dave: Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems* 32 (2019)
- [56] Zadrozny, Bianca ; Elkan, Charles: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Icml* (2001), S. 1–8. ISBN 1–55860–778–1
- [57] Zadrozny, Bianca ; Elkan, Charles: Transforming classifier scores into accurate multi-class probability estimates. **In:** *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002. – ISBN 158113567X, S. 694–699