APRENDIZAJE AUTOMATICO CUANTICO APLICADO A LA CLASIFICACION DE DIABETES

1st Jordan Piero Borda Colque
Ingeniería Estadística e Informática

Puno, Perú

2nd Juan Kenyhy Hancco Quispe
Ingeniería Estadística e Informática

Puno, Perú

Abstract Quantum Machine Learning (QML) aims to show that it has advantages over classical machine learning methods. Only details at the moment some quantum-inspired methods achieve small increases and in the medium term, several experimental cases of hybrid quantum computing are promising for the future (does not take into account the results related to optimization only with classical quantum algorithms). Current quantum computers are noisy and have few qubits to test It is difficult to justify the current quantum advantage and potential of the QML approach. Research shows that we can achieve better classical encoding and better quantum sorting performance. Linear discriminant analysis (LDA) was obtained in the data preprocessing step. the result of the VQC algorithm implemented with Pennylane shows a performance improvement in the balanced accuracy technique (LDA) outperforms the underlying classical classification.

Keywords quantum machine learning; quantum data encoding; classical encoding; dimensionality reduction

I. Introducción

El aprendizaje automático (ML) es hoy una herramienta líder para resolver muchos desafíos en varios sectores, que incluyen: nivel de crédito [Provenzano .Provenzano .2020], análisis de fraude [Tiwari, Mehta, Sakhuja, Kumar SinghTiwari .2021], recomendación de producto [Rohde, Bonner, Dunlop, Vasile KaratzoglouRohde .2018], y previsión de la demanda [Masini, Medeiros MendesMasini .2020], Entre otros casos de uso ampliamente investigados. Bajo esta premisa, la investigación sobre las propiedades de la computación cuántica aplicada al ML se ha expandido rápidamente en los últimos años. Esto se debe a que los beneficios comprobados pueden ser muy útiles en todas las industrias.

El progreso reciente de estas exploraciones en Quantum Machine Learning (QML) Mishra et al. [2021] en el campo cuántico ML se ha expandido rápidamente en los últimos años, ya que puede haber una ventaja confiable útil entre la industria. Avances recientes en esta investigación en aprendizaje automático cuántico (QML).

El enfoque en las tecnologías cuánticas plantea un desafío para determinar si QML ¿Dará una ventaja sobre el aprendizaje automático clásico? La unidad real es ruidoso, lo que significa que la profundidad o el funcionamiento continuo de la puertas logicas son limitados. un qubit pierde el entrelazamiento y pierden información. Estos dispositivos hacen que NISQ [PreskillPreskill2018] y uso limitado de algoritmos cuánticos o híbridos útiles [Callison ChancellorCallison Chancellor2022].

Algunos casos se comercializan y el efecto es alentador, Algunas empresas se embarcan en el viaje de aprendizaje automático cuántico. Un ejemplo es CaixaBank (Español Bank), que desarrolla y prueba modelos QML utilizando el sistema cuántico Pennylane. Definir un modelo de puntuación para la evaluación de riesgos. [CaixaBank. EuromoneyCaixaBank. Euromoney2022]

Uno de los mayores desafíos para obtener resultados confiables sigue siendo la entrada/salida concepto y número de qubits buenos disponibles. IBM esta a la cabeza en esto, pero todavía no es lo suficiente como para usar miles o millones de qubits depende del tipo de problema a resolver. Para que las técnicas QML sean prácticas en un entorno comercial, deben superar las limitaciones de los pequeños números de qubits y crear una forma de utilizar grandes conjuntos de datos [Dalzell, Harrow, Koh La PlacaDalzell .2020].

En este artículo, abordamos el problema de entrada comparando diferentes procesos de preprocesamiento y métodos clasificadores en conjuntos de datos pequeños y grandes con un objetivo binario. el objetivo es determinar una arquitectura específica para el preprocesamiento, reducción de la dimensionalidad del estructura del conjunto de datos, la forma de codificación y el clasificador correspondiente. Demostramos que usar el análisis discriminante lineal (LDA) dentro de la fase de preprocesamiento es mejor que el Análisis de Componentes Principales (PCA) cuando el conjunto de datos posee

una número de características. Generalizamos este enfoque estudiando el efecto de LDA en la codificación de qubits [Haug, Self KimHaug .2021]

II. DATASETS

La selección de datos en esta investigación tiene como objetivo la clasificación de personas con Diabetes. Nosotros usamos el conjunto de datos extraído de Kaggle en un archivo CSV.

Pima Indians Diabetes Database

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

Este conjunto de datos proviene originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es predecir de forma diagnóstica si un paciente tiene o no diabetes, en función de ciertas medidas de diagnóstico incluidas en el conjunto de datos. Se impusieron varias restricciones a la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de herencia indígena pima.

III. METODOS QUE SE APLICARON

A. Reducción de dimensionalidad

PCA es una de las principales estructuras de reducción de dimensionalidad en la exploración de datos QML clásica. [Mensa, Sahin, Tacchino, Barkoutsos TavernelliMensa .2022].

Esta técnica reduce las propiedades y las comprime en N variables para ajustarse a un conjunto de N qubits que se pueden usar para realizar algoritmos de clasificación usando circuitos cuánticos basados en puertas. Este método se usa comúnmente para transformaciones lineales no supervisadas y para encontrar la varianza máxima en datos de alta dimensión. PCA reduce el tamaño al examinar las correlaciones entre varias características, crea ejes ortogonales o componentes principales y utiliza la dirección de mayor variación como un nuevo subespacio.

Hay muchas alternativas a PCA, una de las cuales tiene un gran impacto cuando se trata de problemas de distribución cuántica. LDA es un método supervisado que considera etiquetas de clase al reducir la dimensionalidad. LDA intenta identificar subespacios funcionales que maximicen la separación de clases. LDA funciona calculando el vector d-dimensional medio de las etiquetas de cada clase y construyendo las matrices de varianza dentro de cada clase y entre clases.

Como se mencionó anteriormente, PCA y LDA son técnicas de transformación lineal que descomponen matrices en valores propios y vectores propios. PCA no considera etiquetas de clase no supervisadas, pero LDA sí. Ambas técnicas se aplican al preprocesamiento de datos tradicional en QML. Usando un número limitado de qubits y funciones relacionadas, demostramos las ventajas de LDA sobre PCA.

B. Metricas

Tabla 1. Puntuaciones utilizadas para evaluar distintos clasificadores y sus correspondientes ecuaciones. Dónde TP, FP, TN y FN son verdadero positivo, falso positivo, verdadero negativo y falso positivo, respectivamente.

Metrica	Ecuación
Precisión	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2x(Precisiónx Recall)}{Precisión+Recall}$

Tabla 1.

C. Backends

Los backends y simuladores de computadoras cuánticas no son una decisión fácil cuando se necesita una iteración rápida sobre grandes conjuntos de datos. Los modelos de aprendizaje automático generalmente requieren una serie de ajustes e iteraciones antes de entrar en producción o producción real (remediación). Para QML, los desafíos son los mismos, pero el ecosistema de hardware es diferente. Los algoritmos cuánticos se pueden ejecutar en simuladores (simulaciones informáticas cuánticas completas y ruidosas) y en dispositivos reales.

1) Simuladores: El uso de simuladores cuánticos nos permite principalmente probar y evaluar los resultados en posibles escenarios de computación cuántica del mundo real y, en general, nos brinda la capacidad de ejecutar hasta 50 qubits utilizando computadoras clásicas. En el caso de este experimento, solo usamos el simulador Qiskit Aer y el simulador qubit predeterminado de Pennylane [Bergholm .Bergholm .2018].

Qiskit Air es un simulador de alto rendimiento para Qiskit Terra que proporciona un modelo ruidoso altamente personalizable para estudiar computación cuántica en el dominio NISQ. El kernel está diseñado en C++ para brindar velocidad e incluye elementos desde los simuladores en línea de alto rendimiento de IBM hasta un simulador local que también está escalado para ejecutarse en su computadora portátil o servidor.

In the case of Pennylane's default qubit, is a simple state-vector qubit simulator designed in Python with JAX, Autograd, Torch, and Tensorflow. This simulator is recommended by Pennylane for optimizations with a reduced number of qubits or when stochastic expectation values are going to be used.

IV. ALGORITMOS

A. Modelos de Machine Learning

Los problemas de clasificación son parte del dominio del aprendizaje supervisado y es por eso que usamos varios algoritmos clásicos en esta subárea de ML para establecer un punto de referencia frente al enfoque híbrido cuántico-clásico.

B. Regresión logística

Este método es uno de los más simples para el problema de clasificación binaria. [CramerCramer2002] El modelo es entrenar para aprender los parámetros de la ecuación lineal

$$\hat{k}^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \beta_n x_n^i \quad (1)$$

donde β_n son los coeficientes de regresión lineal x_n, son las características de la (i) muestra. La regresión lineal falla en la tarea de clasificación es por eso que la regresión lineal la formulación de regresión está incrustada en una función logística como la ecuación (1). Para calcular una probabilidad.

$$P(y^i = 1) = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \beta_n x_n^i)} \quad (2)$$

P son la probabilidad de que la etiqueta y para la muestra i corresponda al valor 1.La probabilidad se calcula para cada muestra (el modelo aprende los coeficientes correspondientes) y el umbral de probabilidad se fija en 0,5 para separar la resultado binario. Si $P(y^i = 1) < 0.5$ la etiqueta correspondiente es 0, si $P(y^i = 1) \geq 0.5$ la etiqueta es 1. La logística el método de regresión requiere que muchas muestras sean estables para aproximar eficientemente los parámetros β_n

C. Árbol de clasificación y regresión

Un árbol de decisiones (CART) es una especie de gráfico binario en el que el siguiente niño se basa en la decisión anterior. la base del árbol es la raíz, y luego se crean dos ramas llamadas división (como ”sí” o ”no”). Una estructura de árbol se construye por sucesivas decisiones hasta que se alcanza la última llamada hoja. Este tipo de técnica es simple pero propensa al sobreajuste. Están potentes algoritmos capaces de ajustar conjuntos de datos complejos. El proceso de aprendizaje se realiza con el criterio de Gini o entropía (Eq. 3).

$$H_i = \sum_{k=1}^{10} P_(i,k) log_2 P_(i,k) \quad (3)$$

donde i es el i^{th} nodo, $P_(i,k)$, es la probabilidad de la categoria k.

D. Naive Bayes

El algoritmo Naïve Bayes o NB es una versión más simple del teorema de Bayes (Eq. 4)

$$P(A|B) = \frac{P(A|B).P(A)}{P(B)} \quad (4)$$

Donde A y B son eventos, $P(A|B)$ es la probabilidad de que A dada B sea verdadera, $P(B|A)$ es la probabilidad de B dado que A es verdadero, $P(A)$ y $P(B)$ son las probabilidades independientes de A y B respectivamente. En el caso del clasificador DS, las probabilidades son condicionalmente independientes. Redujo significativamente el cálculo y lo convirtió en un problema tratable.

E. k-Nearest Neighbors

Los k-vecinos más cercanos (k-NN) es un algoritmo simple basado en la distancia no paramétrico. La hipótesis es que un punto similar estará cerrado en un espacio n-dimensional. Un punto será codificado y posicionado por cálculo de distancia (por ejemplo, distancia euclidiana). Luego, el algoritmo toma los k vecinos más cercanos y calcula el promedio de las clases para predecir la clase correspondiente para ese nuevo punto.

V. Quantum Machine Learning Models

Combinando computación cuántica (QC) y aprendizaje automático (ML).El campo emergente del aprendizaje automático cuántico (QML). Esta nueva área de investigación proporciona una mejor capacidad de respuesta de dos maneras, lo que aumenta el rendimiento de algoritmos ML o experimentos cuánticos o ambos. Uso de recursos cuánticos para aumentar Machine Learning En términos de velocidad y/o rendimiento, los investigadores pueden ganar Soluciones alternativas y/o más precisas.

A. Quantum Kernel

En este estudio, usamos una estructura similar a SVC, pero la nuestra está alimentada por un núcleo cuántico [Guo WengGuo Weng2022]. El mecanismo de la función kernel cuántica es similar al convencional, pero su implementación se basa en la superposición cuántica y los estados entrelazados. Además, en el caso de los núcleos cuánticos, el valor de salida depende estadísticamente de probabilidad, por lo que algunos investigadores llaman a este método kernel de probabilidad [SchuldSchuld2021].

B. Quantum Encoding

La codificación cuántica es el proceso desde los datos clásicos hasta la representación cuántica. Hay muchas formas de procesar datos clásicos y crear representaciones útiles. En este estudio, utilizamos el mapa de características cuánticas (Qiskit ZZFeatureMap) y la codificación de ángulos (Pennylane) QSVC y VQC respectivamente.

C. Angle Encoding

La codificación angular es un proceso de codificación de información clásica mediante rotaciones. La información clásica está representada por ángulos de rotación en puertas correspondientes y puede ser escrito como Ecuación:

$$|x\rangle = \bigotimes_i^n R(x_i)|0^n\rangle,$$

donde R son puertas de rotación como Rx, Ry y Rz. La codificación de ángulo se utiliza cuando la dimensión del vector de características x es igual al número de qubits.

D. WorkFlow

Presentamos el WorkFlow que utilizamos a lo largo de este estudio para comparar el algoritmos seleccionados (clásicos y cuánticos). El conjunto de algoritmos se aplicó a los datos. representación generada por métodos de reducción de dimensionalidad. El WorkFlow está compuesto de cuatro pasos:

1) : Cargue los datos y aplique un Exploratory Data Análisis. El objetivo es limpiar los datos y normalizarlos con un buen formato para el método de reducción de dimensionalidad.

2) : Reducción de dimensionalidad: PCA y LDA se utilizan para reducir el número de características a dos dimensiones comprimidas. PCA se utilizo con dos componentes. LDA se usó en un conjunto de datos dividido. Cada media parte se redujo con un componente por LDA.

3) : Codificación cuántica: los datos clásicos se codifican en una representación cuántica mediante mapas de características cuánticas. Este paso solo se usó para algoritmos cuánticos.

4) : Modelos aplicados: el conjunto seleccionado de algoritmos (ML y QML) se aplica a los datos codificación (clásica o cuántica) y evaluados a través de las mismas métricas [Tabla 1].

VI. RESULTADOS

En esta seccion obtenemos los resultados de la aplicacion de los modelos clasicos de clasificacion como LR, CART, KNN Y NB, con la constrastacion de los modelos cuanticos de machine learning QVC Y QSVC aplicados a la clasificacion de diabetes.

Tabla 2.Modelos clásicos (LR, KNN, CART, NB y SVM) y cuánticos (QSVC, VQA) aplicados a el conjunto de datos diabetes utilizando la reducción de dimensionalidad de PCA.

	Precision(%)	Recall(%)	f1-Score(%)	Balanced Accuracy(%)
LR	75	66.67	66.49	80.54
K-NN	66.67	66.67	64.07	74.95
CART	76.33	76.33	75.09	83.42
NB	76.33	75	72.67	84.7
QSVC	76.33	75	72.67	**83.45**
VQC	**83.33**	55.56	66.67	73.23

Tabla 2.Modelos clásicos (LR, KNN, CART, NB y SVM) y cuánticos (QSVC, VQA) aplicados a el conjunto de datos diabetes utilizando la reducción de dimensionalidad de LDA.

	Precision(%)	Recall(%)	f1-Score(%)	Balanced Accuracy(%)
LR	70.92	41.42	51.13	65.91
K-NN	54.82	50.08	51.53	64.65
CART	48.26	51.14	48.26	61.51
NB	69.4	42.78	51.87	66.33
QSVC	62.5	55.56	**58.92**	**72.94**
VQC	41.1	**63.64**	50	64.58

figure 1.Comparación de métricas entre VQC y QSVC usando LDA y PCA aplicadas a la clasificacion de diabetes

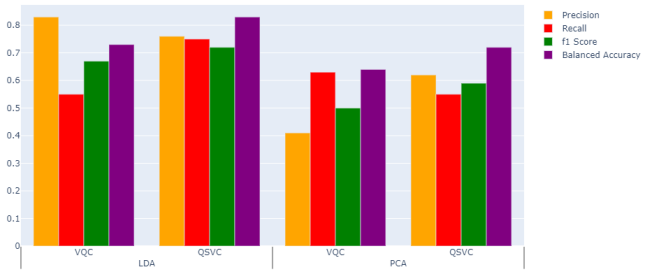
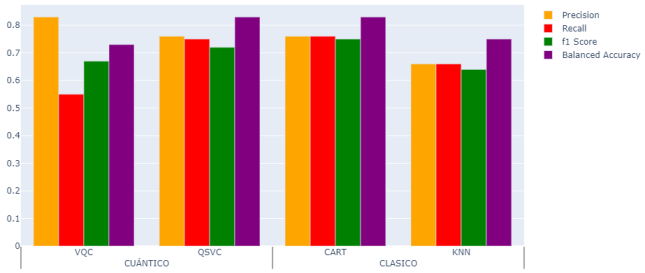


figure 2.Comparación de métricas de VQC y QSVC con CART y KNN (mejores algoritmos clásicos) con la aplicación de LDA aplicado a la clasificacion de diabetes



VII. DISCUSION Y CONCLUSIONES

Mostramos que una computadora cuántica puede extraer información más significativa de los datos clásicos y aproveche los resultados de la clasificación solo usando algunas dimensiones. Como se postula en [Maria SchuldMaria Schuld2022], la ventaja cuántica no necesita ser medido por la capacidad de vencer a los modelos clásicos de ML, pero puede considerarse como una mejor información técnica de extracción. Los pocos números de qubits de las computadoras cuánticas actualmente accesibles obligar a los investigadores a buscar nuevas alternativas. Programas clásicos de reducción de dimensionalidad, como PCA o LDA, son útiles para comprimir conjuntos de datos clásicos de alta característica (más de 100) en un número que se puede utilizar con una computadora cuántica. Aquí, probamos con dos dimensiones para dos qubits. LDA muestra resultados más prometedores para tareas de aprendizaje automático supervisado con computadoras cuánticas. La prevalencia de LDA bajo PCA no se exploró en este documento, pero se explorará en el futuro para comprender cómo LDA proporciona una mejor representación de datos para la codificación qubit.

REFERENCES

[Bergholm .Bergholm .2018] https://doi.org/10.48550/arxiv.1811.04968Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Ahmed, S., Ajith, V.Killoran, N. 2018. PennyLane: Automatic differentiation of hybrid quantum-classical computations. Pennylane: Automatic differentiation of hybrid

quantum-classical computations. arXiv. https://arxiv.org/abs/1811.04968 10.48550/ARXIV.1811.04968

[CaixaBank. EuromoneyCaixaBank. Euromoney2022] caixabankCaixaBank. Euromoney. Caixabank. euromoney. 2022. https://www.euromoney.com/reprints.

[Callison ChancellorCallison Chancellor2022] PhysRevA.106.010101Callison, A. Chancellor, N. 2022Jul. Hybrid quantum-classical algorithms in the noisy intermediate-scale quantum era and beyond Hybrid quantum-classical algorithms in the noisy intermediate-scale quantum era and beyond. Phys. Rev. A106010101. https://link.aps.org/doi/10.1103/PhysRevA.106.010101 10.1103/PhysRevA.106.010101

[CramerCramer2002] concretoCramer, J. 200212. The Origins of Logistic Regression The Origins of Logistic Regression Tinbergen Institute Discussion Papers 02-119/4. Tinbergen Institute. https://ideas.repec.org/p/tin/wpaper/20020119.html

[Dalzell, Harrow, Koh La PlacaDalzell .2020] Dalzell2020howmanyqubitsareDalzell, AM., Harrow, AW., Koh, DE. La Placa, RL. 202005. How many qubits are needed for quantum computational supremacy? How many qubits are needed for quantum computational supremacy? Quantum4264. https://doi.org/10.22331/q-2020-05-11-264 10.22331/q-2020-05-11-264

[Guo WengGuo Weng2022] guo2022whereGuo, M. Weng, Y. 2022. Where can quantum kernel methods make a big difference? Where can quantum kernel methods make a big difference? https://openreview.net/forum?id=NoE4RfaOOa

[Haug, Self KimHaug .2021] https://doi.org/10.48550/arxiv.2108.01039Haug, T., Self, CN. Kim, MS. 2021. Large-scale quantum machine learning. Large-scale quantum machine learning. arXiv. https://arxiv.org/abs/2108.01039 10.48550/ARXIV.2108.01039

[Maria SchuldMaria Schuld2022] https://doi.org/10.48550/arXiv.2203.01340Maria Schuld, NK. 2022. Is quantum advantage the right goal for quantum machine learning? Is quantum advantage the right goal for quantum machine learning? arXiv. https://arxiv.org/abs/2203.01340 10.48550/arXiv.2203.01340

[Masini, Medeiros MendesMasini .2020] https://doi.org/10.48550/arxiv.2012.12802Masini, RP., Medeiros, MC. Mendes, EF. 2020. Machine Learning Advances for Time Series Forecasting. Machine learning advances for time series forecasting. arXiv. https://arxiv.org/abs/2012.12802 10.48550/ARXIV.2012.12802

[Mensa, Sahin, Tacchino, Barkoutsos TavernelliMensa .2022] https://doi.org/10.48550/arxiv.2204.04017Mensa, S., Sahin, E., Tacchino, F., Barkoutsos, PK. Tavernelli, I. 2022. Quantum Machine Learning Framework for Virtual Screening in Drug Discovery: a Prospective Quantum Advantage. Quantum machine learning framework for virtual screening in drug discovery: a prospective quantum advantage. arXiv. https://arxiv.org/abs/2204.04017 10.48550/ARXIV.2204.04017

[PreskillPreskill2018] Preskill2018quantumcomputinginPreskill, J. 201808. Quantum Computing in the NISQ era and beyond Quantum Computing in the NISQ era and beyond. Quantum279. https://doi.org/10.22331/q-2018-08-06-79 10.22331/q-2018-08-06-79

[Provenzano .Provenzano .2020] https://doi.org/10.48550/arxiv.2008.01687Provenzano, AR., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A.Nordio, C. 2020. Machine Learning approach for Credit Scoring. Machine learning approach for credit scoring. arXiv. https://arxiv.org/abs/2008.01687 10.48550/ARXIV.2008.01687

[Rohde, Bonner, Dunlop, Vasile KaratzoglouRohde .2018] DBLP:journals/corr/abs-1808-00720Rohde, D., Bonner, S., Dunlop, T., Vasile, F. Karatzoglou, A. 2018. RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. CoRRabs/1808.00720. http://arxiv.org/abs/1808.00720

[SchuldSchuld2021] https://doi.org/10.48550/arxiv.2101.11020Schuld, M. 2021. Supervised quantum machine learning models are kernel methods. Supervised quantum machine learning models are kernel methods. arXiv. https://arxiv.org/abs/2101.11020 10.48550/ARXIV.2101.11020

[Tiwari, Mehta, Sakhuja, Kumar SinghTiwari .2021] DBLP:journals/corr/abs-2108-10005Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J. Singh, AK. 2021. Credit Card Fraud Detection using Machine Learning: A Study Credit card fraud detection using machine learning: A study. CoRRabs/2108.10005. https://arxiv.org/abs/2108.10005