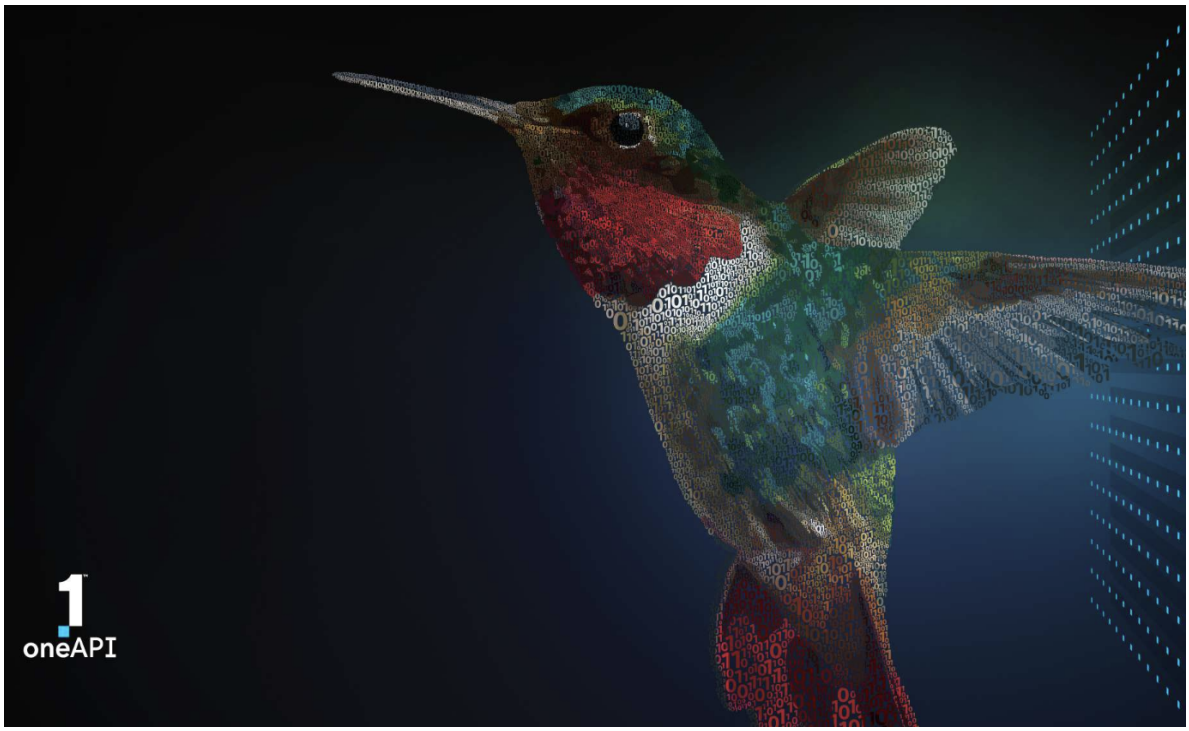
ARQUITECTURA DE COMPUTADORAS 2

# OneAPI

## DESARROLLO DE UNA ARQUITECTURA HETEROGÉNEA, PARA APLICACIONES DE ALTO RENDIMIENTO




## Introducción

Cada vez está más claro que el futuro de la informática no es un chip para todo, sino muchos chips para muchas cosas. Pat Gelsinger, director ejecutivo de Intel, imaginó recientemente un “mar de aceleradores” entre los cuales los clientes podrían seleccionar, mezclar y combinar según sus necesidades específicas.

Suena bastante bien, ¿verdad? A menos que usted sea el desarrollador de software encargado de crear código personalizado para cada nuevo tipo de chip (o incluso simplemente probar un nuevo tipo de chip).

## Conceptos fundamentales para entender que es oneAPI

### Intel Corporation

Intel Corporation, comúnmente conocida como Intel, es una de las principales empresas de tecnología a nivel mundial. Fundada en 1968, Intel se ha destacado en la fabricación de microprocesadores, chips y otros componentes esenciales para computadoras y dispositivos electrónicos. Sus productos y tecnologías son ampliamente utilizados en computadoras personales, servidores, dispositivos móviles, centros de datos, sistemas embebidos y una variedad de aplicaciones tecnológicas.

Intel es conocida por su serie de microprocesadores, como la familia Intel Core para computadoras personales y la serie Intel Xeon para servidores y estaciones de trabajo. La empresa también se ha involucrado en el desarrollo de tecnologías emergentes, como la inteligencia artificial, el internet de las cosas (IoT) y la computación cuántica.

Además de sus productos de hardware, Intel también ofrece software y soluciones relacionadas con la nube, la seguridad informática y la conectividad. A lo largo de su historia, Intel ha sido un líder en la industria de la tecnología y ha contribuido significativamente al avance de la informática y la innovación tecnológica a nivel global.

### ¿Qué es la Computación heterogénea en informática?

La Computación heterogénea se refiere a sistemas que usan más de un tipo de procesador. Estos son sistemas que ganan en rendimiento no justo por añadir el mismo tipo de procesadores, sino por añadir procesadores distintos.

Normalmente, incorporan capacidades de procesado especializadas para realizar tareas particulares.

## Algunos de los XPUs existentes

En informática, el término "XPU" se utiliza para referirse a unidades de procesamiento especializadas o aceleradores diseñados para tareas específicas. Estos aceleradores pueden ser utilizados para acelerar ciertas cargas de trabajo y son comunes en entornos de cómputo heterogéneo, donde se utilizan junto con la unidad central de procesamiento (CPU) principal. Algunos ejemplos de XPUs incluyen:

### GPU (Unidad de Procesamiento Gráfico)

Las GPU están diseñadas principalmente para tareas relacionadas con gráficos, como renderización de videojuegos, modelado 3D y cálculos científicos de alto rendimiento (por ejemplo, en aprendizaje profundo y aprendizaje automático).

### TPU (Unidad de Procesamiento Tensorial)

Las TPU son aceleradores desarrollados por Google específicamente para el procesamiento de redes neuronales y tareas de aprendizaje profundo. Están optimizadas para operaciones matriciales y de tensor, lo que las hace ideales para tareas de IA.

### FPGA (Arreglo de Compuertas Programable en Campo)

Las FPGA son dispositivos lógicos reconfigurables que permiten a los desarrolladores crear hardware personalizado para tareas específicas. Pueden utilizarse en una amplia gama de aplicaciones, desde la criptografía hasta la aceleración de algoritmos de procesamiento de señales.

Estos son solo algunos ejemplos de XPUs que existen en informática. La computación heterogénea, que combina varios tipos de XPUs junto con las CPUs, se utiliza ampliamente para acelerar tareas específicas y mejorar el rendimiento en una variedad de aplicaciones, desde juegos hasta inteligencia artificial y análisis de datos. El uso de XPUs depende de la carga de trabajo y las necesidades específicas de procesamiento de cada aplicación.

## Arquitecturas involucradas

### Escalar

cargas de trabajo complejas que funcionan mejor en una **CPU**.

### Vectorial

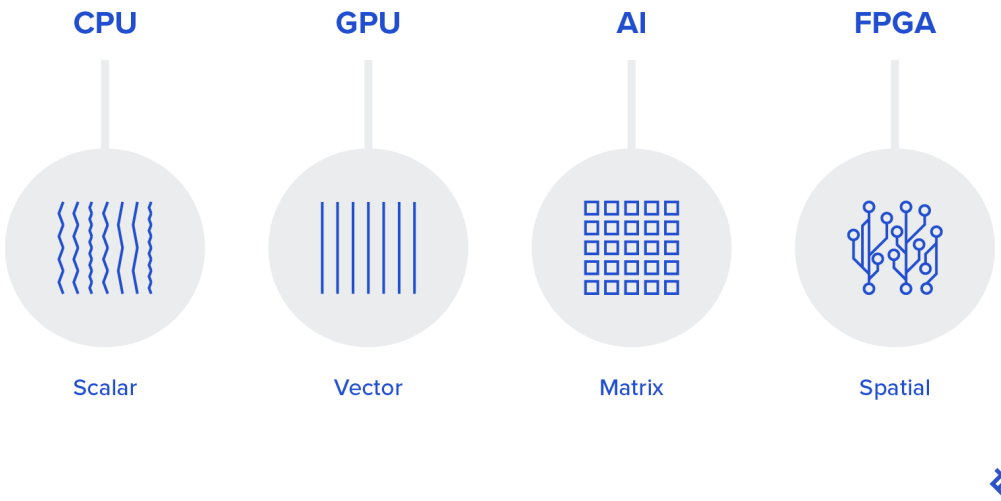
cargas de trabajo que se pueden descomponer en vectores de instrucciones o vectores de elementos de datos y acelerar ejecutando el código en un procesador vectorial como una **GPU**.

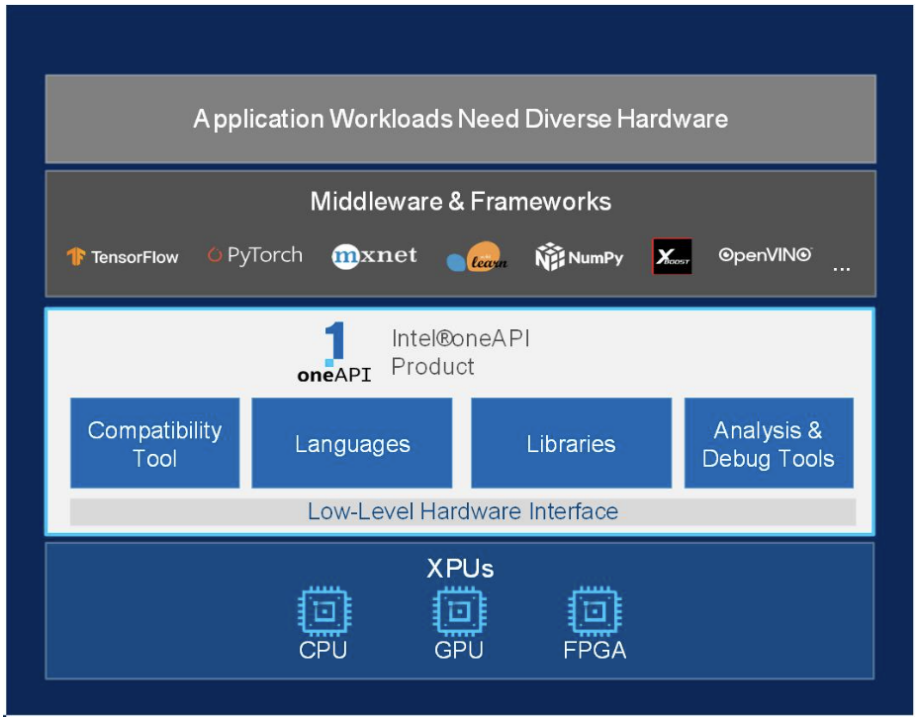
### Matricial

cargas de trabajo, incluidas **IA y ML,** que realizan muchos cálculos matriciales y se ejecutan mejor en chips especializados de IA/ML.

### Espacial

cargas de trabajo que requieren procesadores únicos y especializados construidos mejor sobre la marcha usando **FPGA** para satisfacer las necesidades computacionales de la carga de trabajo específica. Configurados adecuadamente, los FPGA pueden ejecutar cargas de trabajo escalares, vectoriales y matriciales, pero quizás no tan rápido o tan eficientemente como los procesadores especializados.






## Intel oneAPI

**oneAPI** , es promocionado por **Intel** como un modelo de programación único y unificado que tiene como objetivo simplificar el desarrollo en diferentes arquitecturas de hardware: CPU, GPU, FPGA, aceleradores de IA y más.

Intel ahora está comprometida con una estrategia de “primero el software” y espera que los desarrolladores se den cuenta. La gran idea detrás de oneAPI es permitir el uso de una plataforma para una variedad de hardware diferente, por lo que **los desarrolladores no tendrían que usar diferentes lenguajes, herramientas y bibliotecas cuando codifican para CPU y GPU.** Con oneAPI, la caja de herramientas y el código base serían los mismos para ambos.

Para que esto sea posible, **Intel desarrolló Data Parallel C++ (DPC++)** como una alternativa de código abierto a los lenguajes propietarios que normalmente se utilizan para programar hardware específico (por ejemplo, GPU o FFPGA).

Se supone que este nuevo lenguaje de programación ofrece la productividad y familiaridad de C++ al tiempo que incluye un compilador para implementar en diferentes objetivos de hardware.

Data Parallel C++ también incorpora SYCL de Khronos Group para admitir el paralelismo de datos y la programación heterogénea. Actualmente, Intel ofrece acceso gratuito a su DevCloud , lo que permite a los ingenieros de software probar sus herramientas y jugar con oneAPI y DPC++ en la nube sin muchos problemas.

## Intel oneAPI Toolkits

Cajas de herramientas (Toolkits) especializadas para satisfacer las necesidades de diferentes áreas de desarrollo.

### 1. Intel oneAPI Base Toolkit:

Este es el conjunto de herramientas base que proporciona compiladores, bibliotecas y utilidades esenciales para el desarrollo de software en una variedad de dominios.

### 2. Intel oneAPI HPC Toolkit

Diseñado para la informática de alto rendimiento (HPC), incluye herramientas y bibliotecas para la programación paralela y distribuida, así como optimizaciones para aceleradores y procesadores Intel.

### 3. Intel oneAPI AI Analytics Toolkit

Orientado a la inteligencia artificial (IA) y análisis de datos, proporciona herramientas y bibliotecas para el desarrollo de aplicaciones de aprendizaje profundo y análisis de datos de alto rendimiento.

### 4. Intel oneAPI IoT Toolkit

Centrado en el Internet de las cosas (IoT), este toolkit ofrece herramientas para el desarrollo de aplicaciones IoT que pueden aprovechar el rendimiento y la eficiencia energética de los procesadores Intel.

### 5. Intel oneAPI Rendering Toolkit

Diseñado para gráficos y visualización, este toolkit proporciona bibliotecas y herramientas para el desarrollo de aplicaciones de renderización y simulación.

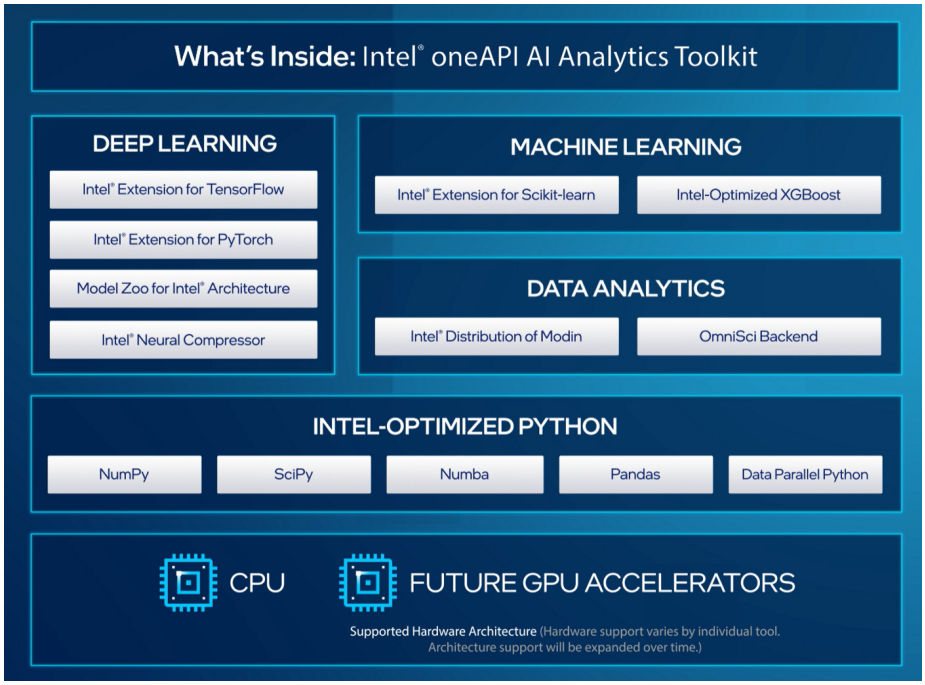
### 6. Intel oneAPI Threading Building Blocks (TBB)

Aunque no es un toolkit independiente, TBB es una biblioteca importante para la programación paralela y de múltiples hilos que se utiliza en combinación con otros toolkits.

Es importante tener en cuenta que Intel continúa desarrollando y actualizando su suite de herramientas oneAPI, por lo que pueden haberse agregado nuevos toolkits o se han realizado cambios. Se recomienda visitar el sitio web oficial de Intel para obtener información actualizada sobre los toolkits oneAPI disponibles.

## Intel® AI Analytics Toolkit (AI Kit)

### Kit de herramientas de análisis de IA Intel® (kit de IA)



Logre un rendimiento de extremo a extremo para cargas de trabajo de IA impulsadas por oneAPI

El kit de IA ofrece a los científicos de datos, desarrolladores de IA e investigadores herramientas y frameworks de Python* familiares para acelerar los procesos de análisis y ciencia de datos de un extremo a otro en la arquitectura Intel®.

Los componentes se crean utilizando bibliotecas oneAPI para optimizaciones informáticas de bajo nivel. Este conjunto de herramientas maximiza el rendimiento desde el preprocesamiento hasta el aprendizaje automático y proporciona interoperabilidad para el desarrollo eficiente de modelos.

Con este kit de herramientas, usted puede:

- Ofrecer alto rendimiento, entrenamiento de aprendizaje profundo en Intel® XPU e integrar una inferencia rápida en su flujo de trabajo de desarrollo de IA con tecnología Intel® optimizada,frameworks de aprendizaje profundo para TensorFlow* y PyTorch*, modelos previamente entrenados y herramientas de baja precisión.
- Lograr una aceleración directa para el preprocesamiento de datos y los flujos de trabajo de aprendizaje automático con paquetes Python de computación intensiva, Modin*, scikit-learn* y XGBoost.

Obtenga acceso directo a análisis y optimizaciones de IA de Intel para garantizar que su software funcione sin problemas.

## Acelerando scikit-learn para análisis de datos y aprendizaje automático

Tanto scikit-learn como Intel Extensión para Scikit-learn son parte del conjunto integral de herramientas y recursos de desarrollo de aprendizaje automático e inteligencia artificial Intel®.

## Intel® Extensión para Scikit-learn*

Intel® Extensión para Scikit-learn* acelera sin problemas sus aplicaciones scikit-learn para CPU y GPU Intel en configuraciones de uno y varios nodos.

Este paquete de extensión parchea dinámicamente los estimadores de scikit-learn mientras mejora el rendimiento de sus algoritmos de aprendizaje automático.

La extensión es parte del **Intel® AI Analytics Toolkit (AI Kit)** que brinda flexibilidad para usar herramientas de aprendizaje automático con sus paquetes de AI existentes.

Usando scikit-learn con esta extensión, puedes:

- Acelerar el entrenamiento y la inferencia hasta **100 veces c**on la precisión matemática equivalente.
- Continuar utilizando la API scikit-learn de código abierto.
- Habilitar y deshabilitar la extensión con un par de líneas de código o en la línea de comando.
- Acelere los algoritmos de scikit-learn (sklearn) reemplazando los estimadores existentes con versiones aceleradas matemáticamente equivalentes (ver algoritmos soportados).
- Ejecute en su elección de CPU compatible con x86 o GPU Intel porque las aceleraciones están impulsadas por la Biblioteca de análisis de datos Intel® oneAPI (oneDAL).
- Elige cómo aplicar las aceleraciones:
- Parchee todos los algoritmos compatibles desde la línea de comando sin cambios de código.
- Agregue dos líneas de código para parchear todos los algoritmos compatibles en su secuencia de comandos Python.
- Especifique en su script parchear solo los algoritmos seleccionados.
- Parchee y elimine parches globalmente de su entorno para todos los usos de scikit-learn.

## Validación del uso de Intel® Extension for Scikit-learn


Utilizando Intel® Extensión para Scikit-learn*, mostraré en **google colaboratory,** cómo se aceleran los algoritmos de Scikit-learn de manera perfecta con la instalación de un paquete pip y dos líneas de código ejecutando un cuaderno de ejemplo del repositorio de oneAPI.

El ejemplo ejecutado fue extraido de https://github.com/intel/scikit-learn-intelex/blob/master/examples/notebooks/random_forest_yolanda.ipynb .

Existe un error conocido en google colaboratory que evita la ejecución del código, para lo cual tuve que realizar los siguientes pasos para configurar correctamente el ambiente y que dicho cuaderno pueda ser ejecutado.

```
! python -m pip install --upgrade pip

! python -m pip install scikit-learn-intelex
```

```
import sys

import os

import site

sys.path.append(os.path.join(os.path.dirname(site.getsitepackages()[0]),
"site-packages"))
```

Y finalmente para habilitar la extensión ejecutamos el siguiente código :

```
from sklearnex import patch_sklearn
```

```
patch_sklearn()
```

Una vez importado, podremos utilizar sus diferentes funcionalidades optimizadas.

Ejemplo, una vez instalada la extensión ( **Intel® Extension** )

```
from sklearn.ensemble import RandomForestRegressor
```

Para cancelar las optimizaciones, usamos unpatch_sklearn y volvemos a importar la clase RandomForestRegressor.

```
from sklearnex import unpatch_sklearn

unpatch_sklearn()
```

La extensión Intel® para parches de Scikit-learn afecta el rendimiento de una funcionalidad específica de Scikit-learn. Consulte la lista de algoritmos y parámetros admitidos para obtener más detalles (https://intel.github.io/scikit-learn-intelex/latest/algorithms.html).

En los casos en que se utilizan parámetros no compatibles, el paquete recurre al Scikit-learn original. Si el parche no cubre sus escenarios, envíe un problema en GitHub.

## Supported Algorithms

Applying Intel® Extension for Scikit-learn* impacts the following scikit-learn algorithms:

### on CPU

#### Classification

| Algorithm | Parameters | Data formats |
|---|---|---|
| *SVC* | All parameters are supported | No limitations |
| *NuSVC* | All parameters are supported | No limitations |
| *RandomForestClassifier* | All parameters are supported except:<br>• `warm_start` = *True*<br>• `cpp_alpha` != *0*<br>• `criterion` != *'gini'* | Multi-output and sparse data are not supported |
| *KNeighborsClassifier* | • For `algorithm` == *'kd_tree'*:<br>all parameters except `metric` != *'euclidean'* or *'minkowski'* with `p` != *2*<br>• For `algorithm` == *'brute'*:<br>all parameters except `metric` not in [*'euclidean'*, *'manhattan'*, *'minkowski'*, *'chebyshev'*, *'cosine'*] | Multi-output and sparse data are not supported |
| *LogisticRegression* | All parameters are supported except:<br>• `solver` not in [*'lbfgs'*, *'newton-cg'*]<br>• `class_weight` != *None*<br>• `sample_weight` != *None* | Only dense data is supported |

Puedes encontrar el ejemplo usando **google colaboratory** realizado en mi repositorio :

https://github.com/anissval/oneAPI/blob/main/Intel%20AI%20Analytics%20Toolkit/random_forest_yolanda.ipynb

### (MSE) – Error cuadrático medio

MSE básicamente mide el error cuadrado promedio de nuestras predicciones. Para cada punto, calcula la diferencia cuadrada entre las predicciones y el objetivo y luego promedia esos valores.

**Cuanto mayor sea este valor, peor es el modelo**. Nunca es negativo, ya que estamos cuadrando los errores de predicción individuales antes de sumarlos, pero sería cero para un modelo perfecto.

### Resultados de la comparación de la métrica MSE de Scikit-learn parcheado y original

**Métrica MSE de Scikit-learn parcheado**: 83.60018104634611 -> **mejor modelo**

Métrica MSE de Scikit-learn sin parches: 83.80131297814816

Relación de métricas: 0,9975998952205618

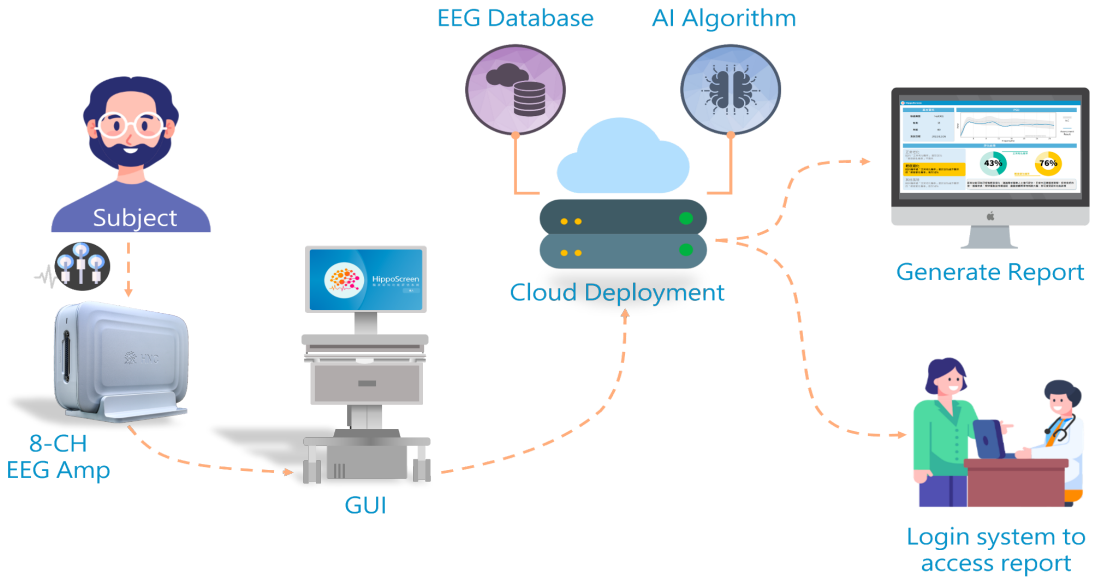
Original Scikit-learn time: **4453.85 s**

Intel® extension for Scikit-learn time: **2081.74 s**

**Obtén aceleración en 2,1 veces.**

## Caso de éxito, uso de herramientas Intel oneAPI y los frameworks de IA

### HippoScreen y oneAPI mejoran las pruebas de salud mental




**HippoScreen** aprovecha las herramientas Intel oneAPI y los frameworks de IA para optimizar los modelos de deep-learning que ayudan a diagnosticar la depresión.

Utilizando el Intel® AI Analytics Toolkit y el Intel® oneAPI Base Toolkit, HippoScreen mejoró la eficiencia y los tiempos de construcción de los modelos de deep-learning utilizados en su sistema de inteligencia artificial (IA) de ondas cerebrales. Estas mejoras permitieron a HippoScreen ampliar las aplicaciones de su sistema a una gama más amplia de afecciones y enfermedades psiquiátricas.

A nivel mundial, se estima que el 5 % de los adultos sufren de depresión. No existe un procedimiento de diagnóstico único para la depresión y, aunque algunos casos pueden diagnosticarse clínicamente, la mayoría de las evaluaciones dependen de las autodescripciones subjetivas de los pacientes. Para superar este problema, junto con el estigma generalizado que rodea a la depresión, HippoScreen desarrolló el sistema Stress EEG Assessment (SEA), que ayuda a los médicos a diagnosticar con mayor precisión los trastornos mentales. El sistema incluye un amplificador de electroencefalograma (EEG)

para la recopilación de datos y el procesamiento de señales, una interfaz gráfica de usuario para el control del proceso de prueba y un algoritmo de IA para el análisis de datos. SEA analiza señales de ondas cerebrales de 90 segundos y proporciona un índice de evaluación objetivo y cuantificable que busca representar numéricamente la probabilidad de que un individuo esté sufriendo de depresión.

Para mejorar la eficacia de los algoritmos y la precisión de los diagnósticos, al mismo tiempo que se reducen los tiempos de entrega de resultados de diagnóstico críticos al personal médico, HippoScreen aprovechó las optimizaciones de las herramientas de análisis de Intel y los frameworks de IA. Utilizando la herramienta de análisis Intel® VTune™ Profiler, la empresa alcanzó el máximo rendimiento y la mínima utilización de la CPU con un número de hilos de cinco. Además, el desempeño mejoró 2 veces, lo que permite a la empresa identificar y resolver rápidamente problemas de sobresuscripción de subprocesos.

Intel® Optimization for PyTorch e Intel® Extension for Scikit-learn, junto con los algoritmos propios de HippoScreen, analizaron las características de los datos del sistema EEG que culminaron en un factor de decisión único y dieron lugar a mejoras de desempeño 2,4 veces mayor.

"En HippoScreen hemos podido aprovechar las optimizaciones de software de Intel® Extension for Scikit-learn e Intel® Optimization for PyTorch para acelerar 2,4 veces los tiempos de compilación de los modelos de IA en nuestro sistema personalizado de análisis de ondas cerebrales EEG", afirma Daniel Weng, director de tecnología de HippoScreen NeuroTech Corp.

## Más apertura, más opciones, más rendimiento

### ¿Qué sigue en el viaje cohete de oneAPI? Más ambición, por supuesto.

Por parte del desarrollador, oneAPI ha estado soportando más lenguajes además de C++, como Python, Java y Julia. En cuanto al hardware, los próximos kits de herramientas Intel oneAPI 2023 obtienen soporte para las arquitecturas CPU, GPU y FPGA más recientes y futuras de Intel, e incluyen herramientas para facilitar la conversión de código propietario a código multiarquitectura.

En cuanto al rendimiento, se prevé hacer que oneAPI sea aún más inteligente, capaz de equilibrar la carga entre muchas CPU, GPU u otros aceleradores, no solo dentro de una máquina, sino entre bastidores y bastidores de ellas.

“Queremos que los usuarios se sientan cómodos manteniendo su software, incluso si necesitan dirigirse a nuevos clientes o nuevo hardware. Estamos creando la infraestructura que permite la diversidad de opciones en el ecosistema y esperamos que los clientes elijan Intel”.

## Una carrera dedicada a "pensar en paralelo"

A medida que llegue al mercado una nueva generación de supercomputadoras del tamaño de la palma de la mano, como la serie Intel® Data Center GPU Max, y las propias supercomputadoras superen la exaescala y entren en niveles de rendimiento en la escala zetta, la microgestión del hardware podría volverse simplemente imposible. Petersen aconseja a los desarrolladores "pensar en paralelo".

Como podemos observar, OneAPI ayuda al progreso al facilitar el desarrollo de aplicaciones de alto rendimiento que aprovechan al máximo la potencia de hardware heterogéneo. Esto tiene un impacto positivo en una amplia gama de industrias y campos, lo que conduce a avances tecnológicos y científicos significativos.

Con todo el avance alcanzado hasta el día de hoy solo queda ver todas las mejoras a futuro que esta tecnología y las industrias que hagan uso de ella nos brindaran.

## Referencias

https://github.com/intel/scikit-learn-intelex/blob/master/examples/notebooks/random_forest_yolanda.ipynb

https://github.com/intel/scikit-learn-intelex

https://stackoverflow.com/questions/67274928/unable-to-install-scikit-learn-intelex-in-colab

https://medium.com/@kazithaque22/intel-oneapi-for-heterogeneous-computing-ba0be34e72a2

https://www.intel.la/content/www/xl/es/newsroom/news/oneapi-2023-tools-maximize-value-intel-hardware.html#gs.515nwv

https://www.intel.la/content/www/xl/es/newsroom/news/hipposcreen-oneapi-improve-mental-health-testing.html

https://www.intel.com/content/www/us/en/developer/tools/oneapi/toolkits.html#gs.58pkh0

https://www.toptal.com/c-plus-plus/intel-oneapi-dpc-plus-plus

https://indico.cern.ch/event/878418/sessions/338513/attachments/2008185/3354564/Introduction_Intel_oneAPI_2020-03-24.pdf