



AstroDatatón 2024

01 de Agosto al 01 de Octubre de 2024

<http://dataton.inf.utfsm.cl>

Introducción

La Astro-Datatón 2024 es una iniciativa organizada por el Departamento de Informática de la UTFSM, en colaboración con científicos de la Uda, UV y PUCV, que abre una convocatoria para que estudiantes universitarios de pregrado y postgrado de todo Chile compitan resolviendo un problema de la Astrofísica aplicando técnicas de la Inteligencia Artificial. Los datos asociados a la competencia han sido generados específicamente para esta competencia siguiendo metodologías validadas en la literatura científica.

La iniciativa cuenta con el financiamiento de la Alianza Académica para la Transformación Digital UTFSM-Scotiabank y con el patrocinio del Centro Científico y Tecnológico de Valparaíso (CCTVal).

El evento será difundido desde el 15 de Julio de 2024 hasta al menos el 30 de agosto de 2024 en redes sociales y listas de correo de universidades chilenas como por ejemplo la lista de la Sociedad Chilena de Ciencias de la Computación (SCCC) y la lista de la Sociedad Chilena de Astronomía (SOCHIAS). El evento se divide en dos fases principales. En la Fase 1, los equipos trabajarán remota y asincrónicamente (online) para desarrollar y enviar soluciones preliminares a través de una plataforma dedicada. En la Fase 2, los equipos seleccionados presentarán sus soluciones ante un jurado de expertos que definirá a los ganadores.

¿Quiénes Pueden Participar?

La participación en la datatón está abierta a estudiantes de pregrado o postgrado, tanto chilenos como extranjeros, siempre que estén asociados a un centro de investigación, universidad u otro tipo de centro de educación superior chileno. En caso de ser necesario, los organizadores podrían requerir la acreditación de la condición de estudiante mediante un

certificado de alumno regular (con validez de más o menos un año) o una carta firmada por una autoridad académica/docente del centro, carrera o unidad de la que el estudiante forma parte.

Una condición fundamental para que un equipo pueda adjudicarse alguno de los premios materiales de la competencia, es que el equipo designe un representante que pueda recibir el premio en dinero a través de un banco dentro del territorio nacional.

Se admite explícitamente la participación de estudiantes en intercambio, siempre y cuando la condición anterior se verifique.

Se admite explícitamente la participación de estudiantes chilenos en el extranjero, siempre y cuando formen equipos con al menos un estudiante asociado a una universidad, centro de investigación, o centro de educación superior chileno. En caso de resultar finalista, el equipo de jueces requerirá la presencia de esta última persona en la final.

Se admite explícitamente la participación de mentores no estudiantes en un equipo, siempre y cuando formen equipos con al menos un estudiante asociado a una universidad, centro de investigación, o centro de educación superior chileno. En caso de resultar finalista, el equipo de jueces requerirá que la exposición final sea realizada por uno de los estudiantes.

Los organizadores se reservan el derecho de interpretar las reglas y resolver casos excepcionales pudiendo ser interrogados sobre los criterios empleados.

Fases y Metodología de Participación

Para participar, los estudiantes interesados deberán formar equipos de 1 a 3 personas e inscribirse hasta el 15 de agosto de 2024 siguiendo el procedimiento indicado en el sitio de la competencia (<http://dataton.inf.utfsm.cl>). La composición del equipo debe respetar las reglas del apartado anterior. La fecha de cierre de inscripciones podría ser modificada por los organizadores y en tal caso será comunicada a todos los participantes así como en la redes físicas o virtuales que éste estime pertinente.

Para inscribirse, se requerirá el nombre de los participantes, su afiliación (más reciente, actual o futura), y un nombre de fantasía que identifique al equipo. Estos datos podrían ser

publicados sólo para fines de organización de la competencia. En caso de que, para fines públicos, una persona prefiera sustituir su nombre registral por su nombre social debe contactar a los organizadores explicando el motivo.

La descripción detallada del problema a resolver así como los datos que los diferentes estudiantes inscritos deberán y podrán utilizar para construir sus soluciones estarán disponibles a partir del día 1 de Agosto. Estos datos serán denominados "conjunto de entrenamiento" y su publicación dará origen a la denominada "Fase 1". En la Fase 1, los equipos trabajarán remota y asincrónicamente (online).

Para evaluar la calidad de las soluciones se empleará un conjunto de casos para los cuales los participantes no conocen las respuestas o predicciones correctas. Los participantes deberán desarrollar una solución que permita generar automáticamente predicciones para estos casos denominados "casos de pruebas". El conjunto completo de casos de prueba estará disponible sólo una vez que concluya la Fase 1 y se anuncien los finalistas.

Para que los equipos obtengan feedback sobre el desempeño de las soluciones en todo el arco de duración de la Fase 1, se implementará un sitio web en la plataforma CodaBench (o equivalente). Esta plataforma permite que los equipos participantes envíen una posible solución al problema y que ésta se evalúe automáticamente. Los resultados correspondientes a la última solución enviada por cada equipo serán diariamente publicados en un tablero que mostrará para todos la clasificación de cada equipo según las métricas definidas.

Para la mantención de este tablero, denominado "tablero público" se usará sólo un 50% de los datos totales de pruebas. Además, existirá un número máximo total de envíos que un equipo puede hacer a la plataforma y un límite diario.

La primera semana de octubre se publicará un tablero final de resultados evaluando cada solución sobre el 100% de los datos de prueba correspondientes al desafío. Si un equipo hubiese enviado múltiples soluciones, se considerará sólo la última. Es responsabilidad de cada equipo decidir qué solución se debe evaluar (es decir, se deja como más reciente en el tablero público).

El tablero final permitirá seleccionar a tres equipos finalistas, dando inicio a la Fase 2. Durante esta fase, los finalistas serán invitados a preparar presentaciones que describan sus soluciones, presentaciones a exponer un día entre el 11 y el 18 de octubre (adicionalmente, se podrían

invitar algunos equipos adicionales para menciones honorosas). Un jurado que incluirá expertos de la academia y de la industria, además de representantes de la alianza UTFSM-Scotiabank, evaluará las soluciones mediante criterios cualitativos y cuantitativos documentados en una rúbrica. Si un equipo no quisiera exponer, se invitará a los equipos que ocupen las siguientes posiciones del tablero a sustituirlos en este proceso. Estas presentaciones se realizarán en modalidad presencial en el campus Casa Central de la UTFSM (Avenida España 1680, Valparaíso). Los organizadores podrían solicitar el código necesario para reproducir las predicciones finalistas. La falta de transparencia o reproducibilidad de las soluciones es causal de descalificación por parte de los organizadores o del jurado.

Premios

Los equipos que avancen hasta la Fase 2 serán candidatos a los siguientes premios:

1. Primer Lugar: \$2.000.000 de pesos chilenos.
2. Segundo Lugar: \$700.000 de pesos chilenos.
3. Tercer Lugar: \$300.000 de pesos chilenos.

La evaluación de los jueces y la comunicación de los resultados finales se realizará en una ceremonia pública y presencial de cierre del evento.

La rúbrica definitiva de evaluación será comunicada el día de selección de los finalistas.

Objetivos

- Promover entre los participantes el interés por el trabajo interdisciplinario, presentando un problema real que puede beneficiarse de la colaboración entre ingenieros y astrofísicos.
- Contribuir a desarrollar en los participantes la capacidad de resolver problemas reales de análisis de datos.

- Estimular el interés de estudiantes chilenos por aplicaciones éticas y responsables de la Ciencia de Datos y la Inteligencia Artificial en la industria y la sociedad chilena.
- Fomentar la formación de redes de intercambio y cooperación entre estudiantes chilenos en el área de Ciencia de Datos e Inteligencia Artificial.
- Contribuir a desarrollar nuevos enfoques para problemas aún abiertos en Ciencia de Datos e Inteligencia Artificial.

Características del Desafío

El problema a resolver en este desafío consiste en predecir la distribución de masa de un sistema astrofísico a partir de las deformaciones que produce en la apariencia de cuerpos situados detrás de él. Esto es interesante porque nos permite estudiar la distribución de materia oscura asociada al sistema, materia que no puede ser observada directamente con nuestros telescopios porque no emite luz. En este desafío específico, el sistema astrofísico es un cúmulo de galaxias y los cuerpos detrás de él son galaxias.

En términos prácticos, se requiere construir un sistema que acepte X como entrada y devuelva Y como salida. Si aproximamos las galaxias de fondo como elipses, la entrada X será un cubo con 3 "tajadas" o "slices". Cada una de las primeras dos tajadas será una matriz que representa una de las dos componentes de las elipses que aproximan las galaxias detrás del sistema astrofísico de interés. La tercera tajada será una matriz que representa el error de medida de la elipticidad observada de las galaxias. Por otro lado, la salida Y es una matriz cuyos valores representan la masa proyectada del sistema astrofísico de interés en una posición determinada del cielo.

La forma en píxeles de la entrada es $128 \times 128 \times 3$ y la forma de la salida es $128 \times 128 \times 1$. Para construir el sistema se dispone de un conjunto de "entrenamiento", es decir, muchos pares de la forma (X, Y) donde Y es la distribución de masa correcta para X . Sin embargo, el sistema se evaluará sobre un conjunto de pruebas, es decir, un conjunto de casos para los cuales los participantes sólo conocerán X .

Métricas de Evaluación

Como es usual en competencias de Machine Learning y Ciencia de Datos, el error de predicción de un modelo o solución se calculará como el promedio de los errores de predicción obtenidos sobre un conjunto de pruebas. Con el objetivo de excluir casos de sobre-ajuste¹ y soluciones triviales, es esencial que este conjunto sea diferente al conjunto de datos entregado a los participantes para construir (entrenar) los modelos.

Si $M[i,j]$ denota el valor correspondiente a la posición (i,j) en la representación matricial de un tensor bidimensional M , el error individual de una predicción P correspondiente a un caso de pruebas T , será calculado considerando las siguientes componentes.

1. Error Porcentual Absoluto Medio con Pesos (WMAPE). Una métrica ampliamente utilizada en problemas de pronóstico y análisis predictivo de datos continuos (como en esta competencia) es el error porcentual absoluto medio² (MAPE):

$$\text{MAPE} = \sum_{i,j} \frac{|P[i,j] - T[i,j]|}{|T[i,j]|}$$

A diferencia del error absoluto medio (MAE), definido como

$$\text{MAE} = \sum_{i,j} |P[i,j] - T[i,j]|,$$

el MAPE representa el error en forma de porcentaje respecto del valor real, lo que lo hace fácil de interpretar y comparar entre diferentes conjuntos de datos o modelos. Por otro lado, a diferencia del error cuadrático medio, también ampliamente utilizado, el MAE es menos sensible a errores extremos de predicción. Tanto MAE como MAPE miden el error de predicción pixel a pixel, son sensibles a la magnitud del error en cada pixel y asignan a cada uno de ellos la misma importancia. Esta última propiedad puede corregirse asignando un peso variable $W(i,j)$ a cada posición (i,j) . Para asignar un mayor

¹ Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

² <https://doi.org/10.1016/j.ijforecast.2015.12.003>

peso a las regiones de mayor densidad de masa, se adoptará la siguiente función $W(i,j)$ propuesta recientemente³:

$$W[i, j] = 1 + \frac{T[i, j]}{\max(T)}.$$

Con esta última definición, la primera medida de error a considerar es

$$\text{WMAPE} = \frac{\sum_{i,j} W[i, j] \frac{|P[i,j]-T[i,j]|}{|T[i,j]|}}{\sum_{i,j} W[i, j]},$$

donde la última normalización busca simplemente mantener la interpretación de la métrica MAPE como un porcentaje de error medio.

2. Error DICE (DICEE). El coeficiente de Sørensen-Dice⁴ (DICE) es una métrica ampliamente utilizada para medir el grado de superposición de dos tensores de datos binarios, problema que aparece por ejemplo al querer comparar la segmentación de una imagen predicha por un modelo respecto de la segmentación obtenida mediante un método automático⁵. Dados dos tensores binarios A y G donde G representa el tensor de referencia, el coeficiente DICE puede definirse como

$$\text{DICE} = \frac{\sum_{i,j} G[i, j]A[i, j]}{\sum_{i,j} G[i, j]A[i, j] + \alpha \sum_{i,j} (1 - G[i, j])A[i, j] + \beta \sum_{i,j} (1 - A[i, j])G[i, j]},$$

con

$$\alpha = \beta = 1/2.$$

No es difícil observar que, en el caso de tensores binarios, el numerador de la ecuación anterior determina el número de verdaderos positivos (TP), mientras que el denominador es equivalente a la suma de verdaderos positivos, falsos positivos (ponderados por el coeficiente α) y falsos negativos (ponderados por el coeficiente β).

Cambiando los valores de α y β es posible asignar un mayor peso a falsos positivos o a

³<https://doi.org/10.3847/1538-4357/ac3090>

⁴<https://doi.org/10.1371/journal.pone.0247751>

⁵<https://doi.org/10.1371/journal.pone.0267753>

falsos negativos. La incorporación de esta métrica a la competencia intenta determinar si el modelo predictivo reconoce correctamente las zonas de alta densidad de masa. Para obtener este efecto, los tensores que representan los mapas verdadero T y predicho P serán convertidos a tensores binarios G y A determinando si los valores en cada posición son mayores (1) o menores (0) al valor medio.

$$G[i, j] = (T[i, j] - \bar{T})_+$$

$$A[i, j] = (P[i, j] - \bar{P})_+$$

Dado que el coeficiente DICE está siempre entre 0 y 1, éste puede convertirse en una métrica a minimizar definiendo el error DICE:

$$\text{DICEE} = 1 - \text{DICE}$$

3. Error en Predicción de Picos de Masa. Estudiar la precisión con la que un método identifica las posiciones correspondientes al pico máximo (o a los picos máximos) de un determinado mapa de convergencia es un procedimiento ampliamente empleado en la literatura astrofísica. Para medir esta capacidad de modo más directo, se propone incorporar la siguiente métrica

$$\text{DPICOS} = \sum_{n=1}^N |H_n - H'_n| + |V_n - V'_n|,$$

donde H_n y V_n denotan las posiciones horizontal y vertical respectivamente (en unidades de pixel) correspondientes a la ubicación del n -ésimo pico (intensidad máxima) del mapa de convergencia real y H'_n y V'_n denotan las posiciones horizontal y vertical respectivamente (en unidades de pixel) correspondientes a la ubicación del n -ésimo pico del mapa de convergencia predicho. Para esta competencia se adoptará $N=3$ y se determinará la posición de los picos empleando funciones notables disponibles en el software astronómico LensTools⁶.

Para integrar las diferentes métricas y al mismo tiempo poder identificar en cada momento en qué aspectos una solución es mejor que otra, se propone asignar a cada equipo una clasificación (ranking) independiente por cada métrica. Si se tienen M competidores, el equipo

⁶ <https://doi.org/10.1016/j.ascom.2016.06.001>

con menor error de acuerdo a cada criterio obtendrá M puntos, el siguiente M-1 puntos, y así sucesivamente hasta que la solución con mayor error obtendrá sólo 1 punto. Para combinar los puntos obtenidos en los diferentes criterios de evaluación se utilizarán métodos con propiedades deseables y conocidas⁷ como el *criterio de Borda* (suma de los puntos obtenidos) o el *criterio de Copeland* (número de victorias en comparaciones de a pares).

Se utilizará el sistema de puntaje numérico para seleccionar 3 equipos finalistas. Los 3 equipos con el *puntaje más bajo* (primeros en el tablero) deberán exponer sus soluciones ante un jurado que definirá ganadores definitivos considerando tanto aspectos cuantitativos como cualitativos y una rúbrica (sistema de criterios y escalas).

Evaluación Cualitativa por parte del Jurado

Las presentaciones de los 3 equipos finalistas serán evaluadas por parte de cada miembro del jurado usando una rúbrica de 5 criterios con puntajes de 1 a 4.

Tabla 1. Criterios de evaluación cualitativa de los finalistas de la competencia.

Criterio	Descripción
Compleitud	Se explican todos los pasos relevantes para reproducir la solución, incluyendo el manejo y tratamiento de los datos realizados; la formulación o selección de un modelo predictivo; y la elección de una estrategia de entrenamiento y evaluación/validación.
Originalidad	Se enfatizan aspectos creativos e innovadores en la metodología utilizada, ya sea en el tratamiento de los datos, en la forma en que se obtienen las predicciones, en la forma de entrenar el modelo, o en la forma de evaluar la solución.
Correctitud Técnica	Todas las herramientas, modelos, o métodos que incluye la solución son empleados correctamente. Los datos se tratan, preparan, y utilizan apropiadamente, el modelo es adecuado para el problema, y las métricas de evaluación/validación utilizadas son coherentes con el objetivo de la competencia.

⁷ <https://doi.org/10.1016/j.mathsocsci.2007.04.004>

Claridad	El expositor transmite los aspectos relevantes de su solución de modo claro y conciso, sin perderse en aspectos superfluos, sabe hacia dónde va y de qué está hablando en cada momento. Se entiende bien todo el trabajo realizado.
Profundidad	El presentador motiva su metodología, razona sobre sus elecciones y sus consecuencias en la calidad de la solución, convence y deja ver un buen nivel de dominio y profundización del problema.

El puntaje otorgado por cada juez permitirá ordenar los equipos de primer lugar (mayor puntaje) a tercer lugar (menor puntaje). En caso de que dos o más equipos empaten, éstos obtendrán como ranking el promedio de los ranking que hubiesen obtenido si no hubieran empatado. Por ejemplo, si un juez asigna puntajes 12, 18, y 12 a los equipos 1, 2, y 3 respectivamente, estos equipos obtendrán rankings 2.5, 1, y 2.5 respectivamente.

Tabla 2. Puntajes para evaluación cualitativa de los finalistas de la competencia.

Puntaje 1	Puntaje 2	Puntaje 3	Puntaje 4
Insuficiente	Suficiente	Bueno	Excelente
Claramente no se satisface el criterio	El criterio se satisface poco o quedan dudas sobre su plena satisfacción.	El criterio se satisface con claridad pero hay detalles por mejorar.	El criterio se satisface plenamente. Es difícil esperar más.

A partir de las posiciones (1, 2, o 3) que los diferentes jueces asignen a cada equipo, se determinará la posición promedio de cada uno. El equipo por el mejor ranking promedio ocupará el primer lugar, el equipo con el segundo mejor ranking promedio ocupará el segundo lugar, y el equipo con el tercer mejor ranking promedio ocupará el tercer lugar.

Junto a la evaluación cualitativa anterior, el jurado podría re-evaluar cuantitativamente las soluciones enviadas por los equipos usando criterios que serán publicados antes del inicio de Fase 2.

Fechas Importantes

- Apertura de convocatoria e inscripciones: 15 de julio de 2024
- Cierre de inscripciones: 15 de agosto de 2024
- Apertura de la competencia: 1 de agosto de 2024
- Cierre de la competencia y confirmación de finalistas: 30 de septiembre de 2024
- Evaluación de los finalistas por parte del jurado: 11-18 de octubre de 2024
- Proclamación de resultados finales: 11-18 de octubre de 2024

Los organizadores se reservan el derecho de modificar plazos, fechas y modalidades de evaluación, informando por email a todos los participantes registrados y a través del sitio web de la competencia.