Master en Ciencia de Datos

Visualización de Datos
Proyecto Final

Kevin Martín Chinea

1. Justificación del conjunto de datos

Kaggle se trata de una plataforma la cual basa su funcionamiento en el mundo de los datos. En esta se ofrecen una serie de desafíos y competencias para que personas especializadas en la ciencia de datos y aprendizaje automático puedan demostrar sus habilidades. Además, proporciona otras características como una amplia gama de recursos educativos, tutoriales y herramientas para ayudar a los usuarios a ampliar sus conocimientos en proyectos reales. También hay que destacar que cualquier persona en esta plataforma puede mostrar su trabajo haciendo que se compartan ideas, recursos y conocimiento para la mejora individual y de la propia comunidad.

Aparte de los conjuntos de datos compartidos por las organizaciones, esta plataforma genera metadatos, datos que se utilizan para describir y contextualizar otros datos. Dentro de estos podemos encontrar: los equipos que participan en las competiciones, las fechas de creación de los datos, quién ha proporcionado estos datos, las relaciones existentes entre diversos usuarios, entre otros muchos.

2. Relevancia del conjunto de datos en su contexto

Por lo descrito hasta ahora, se ha seleccionado el conjunto de datos Meta Kaggle para trabajar en esta práctica, destacando el potencial que ofrece los metadatos de esta plataforma (la cual se puede considerar una de las más importante en la ciencia de datos) para determinar aspectos como: métricas de evaluación para proyectos de ciencia de datos, influencia de equipos/usuarios o la evaluación de los temas de interés a lo largo del tiempo, entre muchos otros. Características que son, o pueden ser, de especial interés especialmente para alguien que quiere profundizar en la ciencia de datos o simplemente mantenerse al tanto de la evolución de esta.

Como se puede intuir, los propietarios de este conjunto de datos es el propio Kaggle el cual publica una parte de los datos derivados de su base de datos (procesados, transformados y filtrados, tanto verticalmente (columnas), como horizontalmente (filas)). Por esto, podemos asumir que la fuente de datos es fiable y que está libre de errores metodológicos durante su captura y procesamiento.

Dentro de la metodología de obtención de los datos se ha respetado cualquier información de tipo privada de los usuarios definiendo parte del conjunto de datos información exclusivamente pública de la plataforma. A nivel de actualización tiene una frecuencia prevista diaria, haciendo que estos estén continuamente actualizados y siempre públicos mediante una licencia CC BY-NC-SA 4.0. Licencia característica por permitir su utilización a otros, reutilizar y desarrollar su trabajo de forma no comercial, siempre que le den crédito y las nuevas creaciones estén bajo los mismos términos.

3. Complejidad del conjunto de datos.

El conjunto de datos se encuentra formado por 32 ficheros en formato CSV: CompetitionTags, Competitions, DatasetTags, DatasetTaskSubmissions, DatasetTasks, DatasetVersions, DatasetVotes, Datasets, Datasources, EpisodeAgents, Episodes, ForumMessageVotes, ForumMessages, ForumTopics, Forums, KernelLanguages, KernelTags, KernelVersionCompetitionSources, KernelVersionDatasetSources, KernelVersionKernelSources, KernelVersions, KernelVotes, Kernels, Organizations, Submissions, Tags, TeamMemberships, Teams, UserAchievements, UserFollowers, UserOrganizations, y Users.

Dependiendo del fichero con el que trabajemos tendrá una información u otra. A continuación podemos ver un listado con algunos de los más significativos para el estudio que vamos a llevar a cabo y la información que contienen:

- ***CompetitionTags.csv:*** Tags de las competiciones.
- ***Competitions.csv:*** Registros de las competiciones existentes.
- ***DatasetTags.csv:*** Tags asociados a los conjuntos de datos publicados.
- ***Datasets.csv:*** Conjunto de datos existentes en la plataforma.
- ***UserFollowers.csv:*** Relaciones entre los usuarios.
- ***UserOrganizations.csv:*** Relación de pertenencia de un usuario a una organización específica.
- ***Users.csv:*** Listado de los usuarios existentes con su información.
- ***DatasetVotes.csv:*** Votos positivos de los usuarios al dataset.
- ***DatasetVersions.csv:*** Distintas versiones de los conjuntos de datos existentes.

Además, vemos que ciertos ficheros, como por ejemplo el *UserOrganization.csv*, *DatasetTags.csv* ó *CompetitionTags.csv*, tienen estructuras formadas por tres claves en las que se relacionan dos conceptos (por ejemplo, usuario y organización). Esto es debido al origen de los datos, que como ya se ha explicado en apartados anteriores, proviene de la propia base de datos SQL de Kaggle significando que este fichero es la relación entre dos tablas.

Algunos de los ficheros con los que vamos a trabajar son los siguientes:

CompetitionTags.csv:

- *Id*: Entero identificador del registro.
- *CompetitionId*: Identificador de la competición.
- *TagId*: Identificador del tag.

Competition.csv:

- *Id*: Identificador de la competición.
- *Slug*: Nombre sin espacios.
- *Title*: Título de la competición.

- *Subtitle*: Subtítulo de la competición.
- *HostSegmentTitle*: Título del segmento hospedado.
- *ForumId*: Identificador del foro asociado.
- *OrganizationId*: Identificador de la organización que gestiona la competición.
- *CompetitionTypeId*: Identificador del tipo de competición.
- *HostName*: Nombre del hospedador.
- *EnabledDate*: Fecha de disponibilidad.
- *DeadlineDate*: Fecha final de la competición.
- *ProhibitNewEntrantsDeadlineDate*: Fecha límite para nuevas entradas.
- *TeamMergerDeadlineDate*: Fecha límite para incorporación de equipos.
- *TeamModelDeadlineDate*: Fecha límite para los modelos de los equipos.
- *ModelSubmissionDeadlineDate*: Fecha límite de aportación de los modelos.
- *FinalLeaderboardHasBeenVerified*: Booleano que verifica si el tablero de líderes está verificado.
- *HasKernels*: Booleano que define la existencia de kernels.
- *OnlyAllowKernelSubmissions*: Booleano que define si la competición sólo acepta kernels.
- *HasLeaderboard*: Booleano que define la existencia del tablero de líderes.
- *LeaderboardPercentage*: Porcentaje del tablero de líderes.
- *LeaderboardDisplayFormat*: Formato del tablero de líderes.
- *EvaluationAlgorithmAbbreviation*: Abreviación del nombre del algoritmo.
- *EvaluationAlgorithmName*: Nombre del algoritmo.
- *EvaluationAlgorithmDescription*: Descripción del algoritmo empleado.
- *EvaluationAlgorithmIsMax*: Booleano que define si la evaluación del algoritmo es máxima.
- *ValidationSetName*: Nombre del conjunto de validación.
- *ValidationSetValue*: Valor del conjunto de validación.
- *MaxDailySubmissions*: Número máximo de aplicaciones por día.
- *NumScoredSubmissions*: Número de puntuación de las aportaciones realizadas.
- *MaxTeamSize*: Tamaño del equipo.
- *BanTeamMergers*: Prohibición de fusiones en equipo.
- *EnableTeamModels*: Disponibilidad de modelos de equipo.
- *EnableSubmissionModelHashes*: Disponibilidad de aportación de modelos hashes.
- *EnableSubmissionModelAttachments*: Habilitados archivos adjuntos en el modelo de envío.
- *RewardType*: Tipo de premio.
- *RewardQuantity*: Cantidad del premio.
- *NumPrizes*: Número de premios.
- *UserRankMultiplier*: Multiplicador de rango de usuario.
- *CanQualifyTiers*: Disponibilidad de calificación por niveles.
- *TotalTeams*: Número total de equipos.
- *TotalCompetitors*: Número total de competidores.
- *TotalSubmissions*: Número total de aportaciones.

DatasetTags.csv:

- *Id*: Entero identificador del registro.

- *CompetitionId*: Identificador del usuario.
- *TagId*: Identificador del tag.

Dataset.csv:

- *Id*: Identificador del conjunto de datos
- *CreatorUserId*: Identificador del creador.
- *OwnerUserId*: Identificador del propietario de los datos.
- *OwnerOrganizationId*: Identificador de la organización propietaria.
- *CurrentDatasetVersionId*: Versión actual del conjunto de datos.
- *CurrentDatasourceVersionId*: Identificador de la fuente de datos de la versión.
- *ForumId*: Identificador del foro asociado.
- *Type*: Tipo del conjunto de datos.
- *CreationDate*: Fecha de creación.
- *LastActivityDate*: Última fecha de actualización.
- *TotalViews*: Número total de visitas.
- *TotalDownloads*: Número total de descargas.
- *TotalVotes*: Número total de votos.
- *TotalKernels*: Número total de kernels con este conjunto de datos.

User.csv

- *Id*: Entero identificador del registro.
- *UserName*: String con el nombre de usuario.
- *DisplayName*: String con el nombre del usuario visible a otros.
- *RegisterDate*: Fecha de registro.
- *PerformanceTier*: Valor numérico con el rendimiento del usuario.

UserFollowers.csv:

- *Id*: Entero identificador del registro.
- *UserId*: Identificador del usuario.
- *FollowingUserId*: Identificador del usuario al que sigue.
- *CreationDate*: Fecha de creación de la relación.

Tags.csv:

- *Id*: Identificador de la etiqueta.
- *ParentTagId*: Identificador de la etiqueta padre.
- *Name*: Nombre
- *Slug*: Nombre del identificador
- *FullPath*: Ruta de herencia.
- *Description*: Descripción de la etiqueta.
- *DatasetCount*: Número de conjuntos con esta etiqueta.
- *CompetitionCount*: Número de competiciones con esta etiqueta.
- *KernelCount*: Número de kernels con esta etiqueta.

DatasetVotes:

- *Id*: Identificador del registro
- *UserId*: Identificador del usuario del voto.

- *DatasetVersionId*: Id de la versión.
- *VoteDate*: Fecha de la votación.

DatasetVersions.csv

- *Id*: Identificador de la versión del dataset.
- *DatasetId*: Identificador del conjunto de datos.
- *DatasourceVersionId*: Identificador de la versión de la fuente de datos.
- *CreatorUserId*: Identificador del usuario creador.
- *LicenseName*: Nombre de la licencia.
- *CreationDate*: Fecha de creación.
- *VersionNumber*: Número de versionado.
- *Title*: Título.
- *Slug*: Nombre.
- *Subtitle*: subtítulo.
- *Description*: Descripción.
- *VersionNotes*: Notas sobre la versión.
- *TotalCompressedBytes*: Bytes totales comprimidos.
- *TotalUncompressed*: Total descomprimido.

Como se puede intuir, la primera de las dificultades que presentan estos datos es la gran cantidad de registros que lo forman, lo cual a nivel de procesamiento requiere una alta demanda de recursos. Por otro lado, y relacionado con este primer punto, la dimensión no es sólo grande verticalmente si no también horizontalmente, existiendo un alto número de variables a analizar (con un mínimo de 3 variables por fichero).

Dentro de los datos existentes existe una alta variabilidad de estos, encontrándonos valores tanto cuantitativos como cualitativos. Los primeros podemos verlos en diversas variables como: los identificadores presentados en todos los ficheros, la valoración de rendimiento de un usuario, los votos que un usuario le ha dado a un dataset, las medallas que un usuario tiene, etc. Un claro ejemplo de variables categóricas lo encontramos en los tags, que clasifican tanto competiciones como conjuntos de datos, en estos se pueden apreciar como los diversos tags representan una categoría y además esta está jerarquizada, pudiendo un tag pertenecer a una categoría tag padre.

También nos encontramos con otro tipo de datos, como por ejemplo fechas, estas representan atributos como la creación de una relación de seguimiento entre dos usuarios, la fecha de pertenencia a una organización, la publicación de un conjunto de datos/competición. O booleanos, para representar, por ejemplo, y en el contexto de competiciones, si esta ha sido verificada o no, si existen kernels asociados, si tienen un tablón de líderes, etc. O en otro ejemplo, en un contexto de kernels, una variable que representa si este tiene un notebook asociado o no.

4. Originalidad

Podemos encontrar diversos kernels o códigos en la misma plataforma que trabajan con este conjunto de datos (los ficheros aquí señalados y otros pertenecientes al mismo conjunto).

Por ejemplo tenemos el notebook *Winning Team Submission Traces* del usuario James Trotman que muestra la evolución de los equipos en diversas competiciones a lo largo del tiempo. Otro ejemplo del mismo usuario es Leaderboard Score Landscapes, que muestra rastros de puntajes de presentación públicos/privados para toda la tabla de clasificación, para cada competencia de Kaggle.

Otra de las visualizaciones existentes (la cual se trata de una actualización de versiones anteriores pero aún así puede ser interesante) es la del usuario Anshuman Mishra titulada Kaggle Grand/Masters Map, en esta se puede apreciar un mapa con los mejores participantes de Kaggle, según su competiciones, comentarios, participación en los conjuntos de datos, medallas, etc. Como se puede apreciar es un mapa interactivo en el cual nos podemos acercar a los distintos continentes/países y ver de forma más detallada quienes son estos maestros del dato.

También hay que destacar el Notebook compartido por Andrada Oltenau, One Notebook to Rule 'Em All. En este muestra diversas gráficas sobre los tipos de competiciones existentes, los premios a lo largo de los años, las palabras más comunes en los requisitos, la especialización de los ganadores, etc. Todo con el fin de ver la mejor forma de sacarle partido a las competiciones.

A parte de estas visualizaciones, también hay que destacar que, como bien se ha descrito anteriormente, este conjunto de datos se actualiza diariamente, por lo que cualquier visualización que se haya realizado puede quedar desactualizada en poco tiempo (especialmente, si destacamos que Kaggle es una comunidad bastante activa). Es por esto, que simplemente realizar visualizaciones iguales a las presentadas pueden darnos distintos resultados. A pesar de esto, se plantea realizar un enfoque diferente de los ya planteados.

5. Cuestiones a responder

En este trabajo, se busca explorar un enfoque diferente para la obtención de datos en Kaggle. En primer lugar, se pretende crear una red de conexiones entre los usuarios de la plataforma mostrando información de estos, como podría ser el nombre de usuario, su nombre, el nivel de rendimiento que tiene o la organización a la que pertenece. Como esta red puede ser inmensa, para simplificar se propone utilizar métricas de grafos como el grado, la centralidad de cercanía o la centralidad de influencia. Esto permitirá identificar a las personas más influyentes en la comunidad de Kaggle en términos de sus relaciones con otros usuarios, en lugar de basarse en méritos como se hace en otras visualizaciones.

Una vez identificados los usuarios más influyentes, se pretende analizar en qué conjuntos de datos interactúan y la información presente en estos, como por ejemplo, las descargas que han tenido, las visitas, la organización a la que pertenece, etc. Además, con este análisis se pretende resaltar los temas más populares en el campo de la ciencia de datos según las categorías que estos conjuntos de datos tengan. Hay que destacar que ambos estudios se pueden analizar a nivel de evolución temporal.

En resumen, se busca obtener una visión más profunda de la comunidad de Kaggle a través de la red de conexiones entre sus usuarios y de los conjuntos de datos que manejan.

6. Github

Actualmente, el desarrollo del proyecto se encuentra en el siguiente repositorio de Github, KaggleMetaData. Como se puede deducir se ha llevado a cabo el control de versiones para la gestión de los documentos, los datos originales utilizados, los generados y sus correspondientes códigos.

En el código se puede ver todo el proceso realizado hasta el momento, destacando la unión de los diferentes archivos para reunir los datos de interés, la limpieza de ciertos índices de registros (que estaban en un formato incorrecto), la asociación de usuarios con su última organización vinculada y la creación del grafo a partir del cual se han obtenido nuevas métricas para generar rankings de usuarios y destacar los conjuntos de datos a los que estos les han dado un voto positivo.