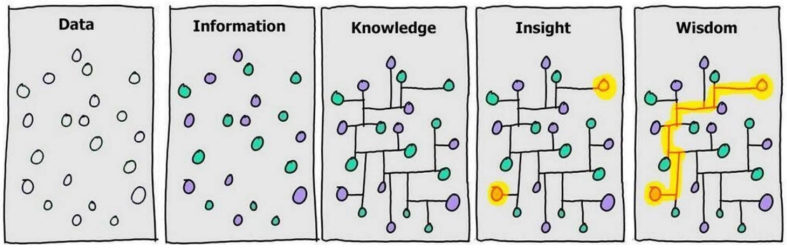
Pràctica 2
Anàlisi de Dades en un Sistema d'Informació Cinematogràfic

- El termini d'entrega de la pràctica finalitza el **5 de maig de 2.024 a les 23:55**.
- Heu de penjar un **informe PDF** al Campus Virtual explicant com heu plantejat i resolt la pràctica. No és necessari adjuntar cap codi.
- La pràctica es realitza en grup de **dues o tres persones**.
- La nota d'aquesta pràctica equival a un **14%** de la nota global de l'assignatura.

Objectiu

Una vegada heu aconseguit estructurar les dades correctament a la vostra base de dades, l'empresa FilmLib vol obtenir coneixement a partir de totes les dades que disposa per tal de posicionar-se al mercat. No obstant, tindre dades no és suficient, cal analitzar-les utilitzant diverses tècniques per tal de descobrir tendències o patrons, predir comportaments, anticipar-se a esdeveniments, detectar els problemes precoçment, i prendre decisions de manera informada. Com bé sabeu, "la informació és poder". Tots aquests recursos s'utilitzen a dia d'avui dins l'àmbit del *data science*, una disciplina cada vegada més important[1].



En aquesta segona pràctica, ens centrarem en **l'anàlisi de les dades**, és a dir, aplicar diverses tècniques i mètodes analítics per tal d'interpretar les dades i obtenir coneixement.

Fase 1: Anàlisi exploratòria de dades

La primera tasca consisteix en realitzar una **anàlisi exploratòria de dades** (en anglès, *exploratory data analysis* –EDA–), un primer pas cap a analitzar les dades. Més concretament, una anàlisi exploratòria és un tractament estadístic utilitzat per explorar, descriure, resumir i entendre les dades amb les que es treballen. Aquest tipus d'anàlisi és

[1] T. H. Davenport and D. J. Patil, "Is Data Scientist Still the Sexiest Job of the 21st Century?", in Harvard Business Review, July 2022, URL: `https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century`

extremadament útil abans de realitzar anàlisis més avançades com, per exemple, mineria de dades, mineria de procés o filtratge col·laboratiu, entre d'altres.

En primer lloc, haureu d'identificar el tipus de cadascuna de les variables que teniu: algunes seran categòriques, d'altres numèriques discretes, d'altres numèriques contínues, etc. En funció del tipus, hi podreu aplicar els estudis que corresponguin per tal d'interpretar aquella variable. Per exemple, podeu realitzar taules de freqüències, estudiar la distribució de probabilitat d'una variable, calcular mesures de posició central (com la moda, la mediana i la mitjana) o de posició no central (com els quartils i els percentils), calcular el grau de dispersió (com el rang inter-quartil, la variància i la desviació típica), o estudiar les mesures de forma (com el coeficient d'asimetria o la curtosi) de la distribució de la variable. Amb tot això, podeu fer-vos preguntes del següent estil:

- Quina plataforma de *streaming* té més contingut d'acció? I la que menys?
- Quina és la distribució entre sèries i pel·lícules a cada plataforma de *streaming*?
- Quins són els deu actors que han protagonitzat més pel·lícules?
- Quina proporció de contingut és troba en més d'una plataforma de *streaming*?
- Hi ha alguna zona del món que hagi produït molt més contingut que la resta?
- Hi ha persones que hagin actuat i dirigit una mateixa sèrie o pel·lícula?

Els estudis anteriors són relativament simples. Però, a partir d'aquí, podeu realitzar estudis més elaborats i complexes que us permetin **comparar, correlacionar i desenvolupar models matemàtics entre les característiques de les sèries i pel·lícules**. Per exemple, podeu realitzar taules de contingència, estudiar la distribució d'una variable condicionada a una altra variable, calcular la covariància o el coeficient de correlació de Pearson entre dues variables, aproximar models de regressió lineals i calcular-ne el seu coeficient de determinació, aproximar models no lineals, etc. A continuació, us proposem una sèrie d'estudis a tall d'exemple:

- Quina és la distribució de duració de les sèries a cada plataforma de *streaming*? I de les pel·lícules?
- Quina plataforma de *streaming* té millor contingut?
- En quina dècada es van fer les millor pel·lícules?
- La distribució dels gèneres de contingut és igual entre les diverses plataformes de *streaming*?
- Existeix alguna relació entre la valoració i la popularitat de les sèries i pel·lícules?
- La popularitat d'una sèrie depèn de la seva longevitat?
- Existeix una relació entre l'any de les sèries i pel·lícules amb la seva valoració? És a dir, podem dir que el contingut més nou té millor valoració que el contingut més antic?
- Es correlaciona la valoració del contingut a IMDB i a TMDB?
- Les sèries i pel·lícules més ben valorades solen ser d'un gènere en concret?

En una anàlisi exploratòria, **la presentació i visualització dels resultats és molt important**. Penseu que moltes vegades, els resultats d'una anàlisi exploratòria es presenten a usuaris no tècnics (directius, personal d'altres departaments...), de manera que

el missatge s'ha de transmetre de manera entenedora visualment. Per aquest motiu, no només haureu d'explicar el procediment analític que heu realitzat en cadascun dels estudis, sinó que també haureu de presentar els resultats a través de les visualitzacions i gràfics que creieu més oportuns. Els gràfics de barres, gràfics de sectors, histogrames, gràfics de caixes i gràfics de dispersió són essencials per aquesta tasca. A partir d'aquí, haureu d'**explicar què observeu i quines conclusions extraieu dels estudis realitzats**.

> Teniu llibertat per realitzar els estudis que vulgueu. Es valorarà positivament la vostra capacitat per realitzar altres estudis, a més de la dificultat d'aquests. Tanmateix, podeu utilitzar els llenguatges de programació (Python, R, Java...) i les llibreries (NumPy, SciPy, Matplotlib, ggplot2...) que preferiu.

Fase 2: Sistema de recomanació de contingut

Una de les funcionalitats demanades per l'empresa FilmLib és la de ser capaços de recomanar contingut cinematogràfic (ja siguin sèries o pel·lícules) a usuaris. Netflix[2] i Amazon[3] són dues empreses altament reconegudes per tindre excel·lents sistemes de recomanació sobre els quals basen els seus serveis. A diferència d'aquestes, FilmLib no té informació de les preferències dels usuaris, de manera que no poden aplicar algorismes de filtratge col·laboratiu. Per això, us demanen **dissenyar, implementar i testejar un algorisme propi de recomanació de contingut**.

Fase 2.1: Disseny de l'algorisme

A l'hora de dissenyar un algorisme de recomanació, cal conèixer les preferències dels usuaris. En el vostre cas, haureu de **demanar aquesta informació a l'usuari** a través de la consola o d'una interfície gràfica. Podeu demanar qualsevol informació que creieu d'interès i que sigui rellevant pel vostre algorisme. Per exemple, podeu preguntar-li si prefereix veure sèries o pel·lícules (o si és indiferent), si vol que siguin (o no siguin) d'algun gènere en concret, si han de tindre alguna restricció d'edat (per si és per menors d'edat), si la duració del contingut és un inconvenient, si vol que hi actuï algú en concret, si la popularitat o la valoració d'aquesta és important, etc. També, podeu proporcionar la recomanació distingint entre cada plataforma de *streaming* o per una plataforma en concret (la que us indiqui l'usuari). El vostre sistema no té perquè proporcionar únicament una recomanació, sinó que pot recomanar k sèries o pel·lícules.

A més a més, el vostre sistema ha de ser capaç de fer recomanacions en **casos particulars**. Per exemple, què passa si l'usuari ja ha vist la sèrie/pel·lícula que li heu recomanat? Podeu anar fent recomanacions fins que l'usuari accepti la vostra recomanació. Un altre exemple: si un usuari té unes preferències extremadament rares (és a dir, no hi ha cap sèrie/pel·lícula que coincideix el 100% amb el que vol), com reacciona el vostre sistema? Igual que Netflix, podeu dissenyar un indicador de coincidència de la vostra recomanació

[2] `https://research.netflix.com/research-area/recommendations`
[3] `https://www.argoid.ai/blog/decoding-amazons-recommendation-system`

amb les preferències introduïdes per l'usuari.

> El disseny de l'algorisme és totalment lliure. Podeu implementar heurístiques pròpies i integrar els mètodes de mineria de dades que creieu oportuns. Tanmateix, podeu utilitzar els llenguatges de programació (Python, R, Java...) i les llibreries (SciPy, scikit-learn, caret, mlr, Java-ML...) que preferiu.

Fase 2.2: Joc de proves

Abans de començar a fer les proves, convé pensar-les. Heu de demostrar que el vostre sistema fa recomanacions amb criteri. Per això, convé fer una **quantitat important de proves** per verificar els diversos casos i perfils d'usuaris que podeu tenir. També, heu de demostrar la robustesa del vostre sistema davant casos particulars o molt específics. És important **extreure conclusions i raonar** el perquè dels resultats obtinguts durant les recomanacions.

> El professor es reserva el dret de realitzar les proves que cregui convenients per validar el funcionament del vostre algorisme, si ho creu oportú.

Lliurament

El lliurament de la pràctica es farà mitjançant la tasca habilitada al Campus Virtual fins el dia **5 de maig de 2.024 a les 23:55**. Haureu de lliurar un **informe PDF**[4], demostrant que heu assolit les diverses fases de la pràctica, on hi consti la següent informació:

- Portada amb el nom de la pràctica, el nom dels membres del grup, el nom de l'assignatura, el curs i la data.
- Índex.
- Breu explicació del problema i objectius principals a assolir.
- Fase 1: explicació detallada dels diversos estudis de l'anàlisi exploratòria de dades. Per a cada estudi, indiqueu clarament la pregunta que us heu plantejat, l'explicació de com l'heu resolt, els seus corresponents resultats analítics, les visualitzacions i les conclusions obtingudes.
- Fase 2.1: explicació detallada del disseny i implementació de l'algorisme de recomanació de contingut, incloent l'heurística i mètodes emprats, les variables considerades, la tecnologia utilitzada, la interacció amb l'usuari, parts del codi que vulgueu ressaltar...

[4] Recomanem l'ús de LaTeX utilitzant la plataforma Overleaf: `www.overleaf.com`

- Fase 2.2: resultats detallats del joc de proves incloent, per cada joc, el seu objectiu, el perfil de l'usuari a testejar, el(s) resultat(s) obtingut(s), la justificació de la recomanació..., així com qualsevol altra heurística que vulgueu destacar.
- Conclusions.
- Bibliografia (si cal).

En aquesta pràctica no cal adjuntar el codi: ja haureu posat a la documentació les parts que voleu destacar d'ell.

Avaluació

Per valorar la pràctica es tindran en compte les ponderacions següents:

- Qualitat tècnica de la fase 1 (quantitat i dificultat d'estudis realitzats, ús d'estadístics, resolució analítica, visualitzacions i extracció de conclusions): 40%
- Qualitat tècnica de la fase 2.1 (algorisme dissenyat, heurístiques utilitzades, variables considerades, interacció amb l'usuari...): 20%
- Qualitat del joc de proves de la fase 2.2 (quantitat d'experiments, varietat d'experiments, recomanacions obtingudes...): 20%
- Qualitat tècnica de l'informe i comunicació escrita: 20%