**Az adatbázis szerkezete**

Az információk a programomhoz használt filmekről a „movies.db” adatbázisban kerültek letárolásra. Az adatbázishoz SQLite formátumot használtam annak egyszerűsége, hordozhatósága miatt, illetve egy nagy előnye, hogy megbízhatóan tud működni gyakorlatilag szinte minden programozási környezetben.

A felhasznált adatok egy kaggle.com-ról származó adatbázisból származnak, melynek forrása a következő: <https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata>

A felhasznált adatbázis a Tmdb-ről származó adatokkal lett feltöltve, így az általam készített adatbázis adatainak az eredeti forrása a Tmdb-n szereplő adatok.

Az adatbázis 11 darab táblát tartalmaz, amelyek a következőek: Movies, Countries, Directors, Genres, Keywords, Languages, Movies\_Countries, Movies\_Directors, Movies\_Genres, Movies\_Keywords, Movies\_Languages.

A **Movies** **tábla** a következő mezőkből áll:

* ID:
  + Típus: INTEGER
  + Leírás: A film egyedi azonosítója, ez a primary key (elsődleges kulcs).
* Title:
  + Típus: TEXT
  + Leírás: A film címe angolul, amely az esetek nagy részében megegyezik a filmek eredeti címével.
* Genre:
  + Típus: TEXT
  + Leírás: A film műfaja(i) felsorolás szerűen, vesszővel elválasztva. Egyes filmeknél csak 1 darab van belőle, más, több műfajba is beleillő filmeknél több is lehet.
* Released:
  + Típus: INTEGER
  + Leírás: A film megjelenésének éve.
* Runtime:
  + Típus: INTEGER
  + Leírás: A film játékideje percben megadva.
* Gender\_of\_the\_protagonist:
  + Típus: INTEGER
  + Leírás: A főszereplő nevét jelzi. Az értéke 1, ha a főszereplő nő, ha férfi akkor pedig 2 az értéke.
* Main\_actor:
  + Típus: TEXT
  + Leírás: A főszereplő színész neve.
* Keywords:
  + Típus: TEXT
  + Leírás: Az adott filmhez kapcsolódó, a Tmdb-ről származó kulcsszavak vesszővel ellátott felsorolása, melyek használatával könnyebben találhatunk hasonlóságokat a filmek között.
* Director:
  + Típus: TEXT
  + Leírás: Az adott film rendezőjének/rendezőinek neve.
* Language:
  + Típus: TEXT
  + Leírás: A film eredeti nyelve, illetve a benne elhangzó nyelvek felsorolása
* Pruduction\_countries:
  + Típus: TEXT
  + Leírás: A készítő országok neveinek felsorolása.
* Tmdb\_score:
  + Típus: REAL
  + Leírás: Az adott film Tmdb-n szereplő értékelése, amely a felhasználói értékelések átlaga. Az értékelés egy 0-tól 10-ig terjedő skálán történik, így a pontszám is a két érték közé eshet csak.
* Number\_of\_ratings:
  + Típus: INTEGER
  + Leírás: Megmutatja, hogy a Tmdb oldalán eddig hányan értékelték az adott filmet, tehát a „Tmdb\_score” nevű mező hány értékelésnek az átlaga.
* Popularity:
  + Típus: REAL
  + Leírás: Az adott film Tmdb szerinti népszerűségi indexe, melyet a felhasználók tevékenységeiből számolnak.
* Budget:
  + Típus: INTEGER
  + Leírás: Az adott film költségvetésének összege amerikai dollárban megadva.
* Revenue:
  + Típus: INTEGER
  + Leírás: Az adott film bevételének összege amerikai dollárban megadva.

A **Countries** **tábla** a következő mezőkből áll:

* ID:
  + Típus: INTEGER
  + Leírás: Az adott ország egyedi azonosítója, ez a primary key.
* Country\_Name:
  + Típus: TEXT
  + Leírás: Az adott ország neve.

A **Directors** **tábla** a következő mezőkből áll:

* ID:
  + Típus: INTEGER
  + Leírás: Az adott rendező egyedi azonosítója, ez a primary key.
* Director\_Name:
  + Típus: TEXT
  + Leírás: Az adott rendező neve.

A **Genres** **tábla** a következő mezőkből áll:

* ID:
  + Típus: INTEGER
  + Leírás: Az adott műfaj egyedi azonosítója, ez a primary key.
* Genre\_Name:
  + Típus: TEXT
  + Leírás: Az adott műfaj neve.

A **Keywords** **tábla** a következő mezőkből áll:

* ID:
  + Típus: INTEGER
  + Leírás: Az adott kulcsszó egyedi azonosítója, ez a primary key.
* Keyword\_Name:
  + Típus: TEXT
  + Leírás: Az adott kulcsszó neve

A **Languages** **tábla** a következő mezőkből áll:

* ID:
  + Típus: INTEGER
  + Leírás: Az adott nyelv egyedi azonosítója, ez a primary key.
* Language\_Name:
  + Típus: TEXT
  + Leírás: Az adott nyelv elnevezése

A **Movies\_Countries** **kapcsolótábla** a következő mezőkből áll:

* Movie\_ID:
  + Típus: INTEGER
  + Leírás: Az adott film egyedi azonosítója, foreign key, amely a Movies tábla ID mezőjére mutat.
* Countries\_ID:
  + Típus: INTEGER
  + Leírás: Az adott ország egyedi azonosítója, foreign key, amely a Country tábla ID mezőjére mutat.

A **Movies\_Directors** **kapcsolótábla** a következő mezőkből áll:

* Movie\_ID:
  + Típus: INTEGER
  + Leírás: Az adott film egyedi azonosítója, foreign key, amely a Movies tábla ID mezőjére mutat.
* Directors\_ID:
  + Típus: INTEGER
  + Leírás: Az adott rendező egyedi azonosítója, foreign key, amely a Director tábla ID mezőjére mutat.

A **Movies\_Genres** **kapcsolótábla** a következő mezőkből áll:

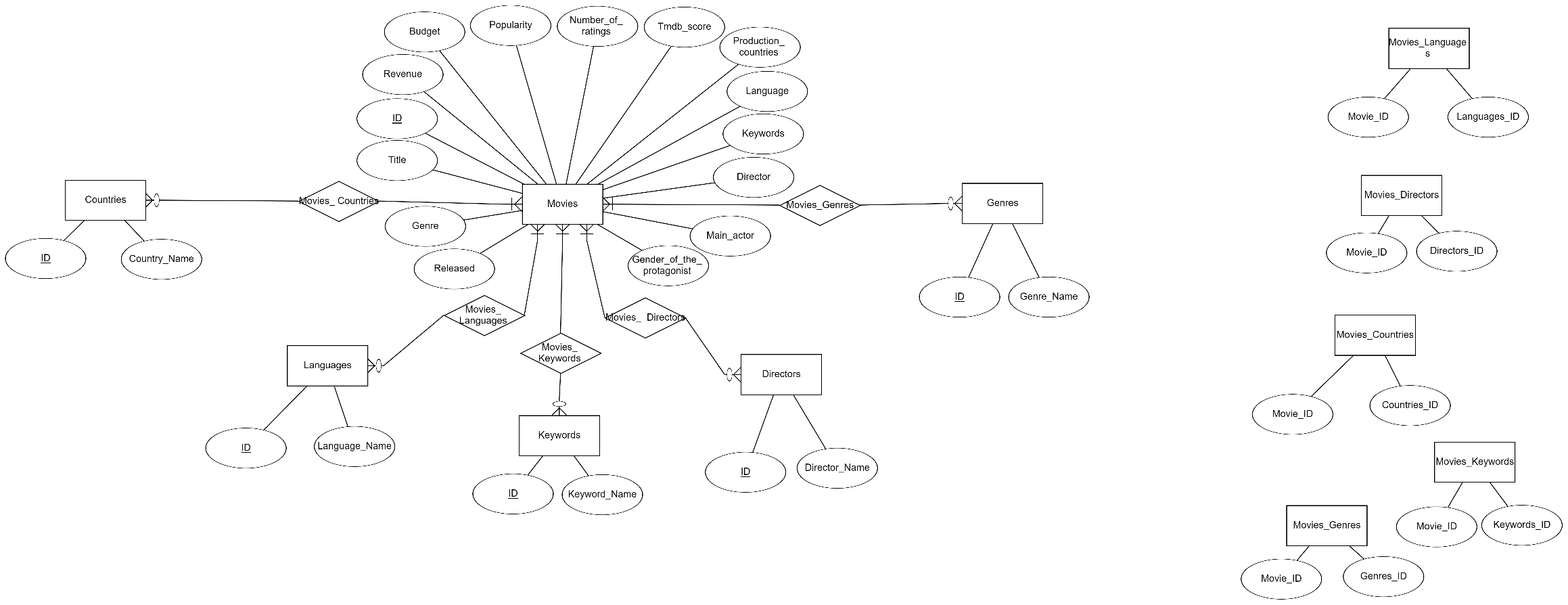
* Movie\_ID:
  + Típus: INTEGER
  + Leírás: Az adott film egyedi azonosítója, foreign key, amely a Movies tábla ID mezőjére mutat.
* Genres\_ID:
  + Típus: INTEGER
  + Leírás: Az adott műfaj egyedi azonosítója, foreign key, amely a Genre tábla ID mezőjére mutat.

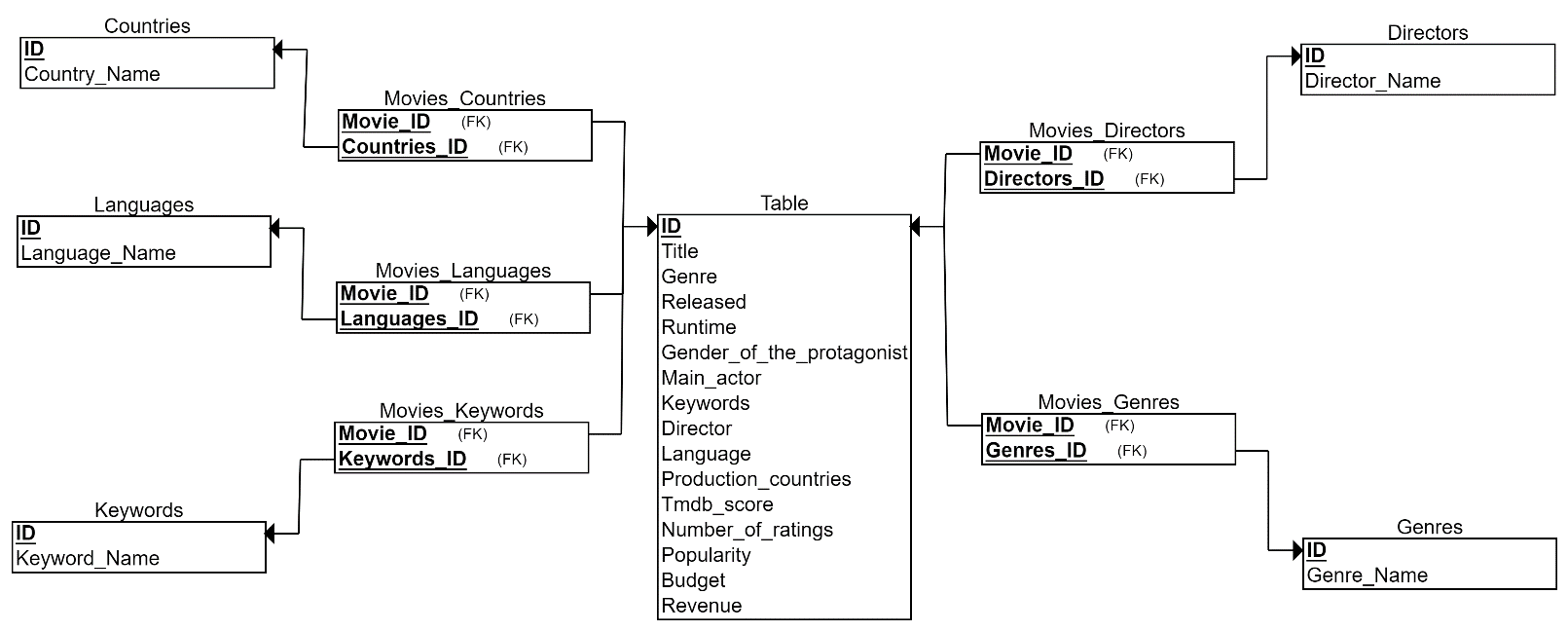
A **Movies\_Keywords** **kapcsolótábla** a következő mezőkből áll:

* Movie\_ID:
  + Típus: INTEGER
  + Leírás: Az adott film egyedi azonosítója, foreign key, amely a Movies tábla ID mezőjére mutat.
* Keywords\_ID:
  + Típus: INTEGER
  + Leírás: Az adott kulcsszó egyedi azonosítója, foreign key, amely a Keyword tábla ID mezőjére mutat.

A **Movies\_Languages** **kapcsolótábla** a következő mezőkből áll:

* Movie\_ID:
  + Típus: INTEGER
  + Leírás: Az adott film egyedi azonosítója, foreign key, amely a Movies tábla ID mezőjére mutat.
* Languages\_ID:
  + Típus: INTEGER
  + Leírás: Az adott nyelv egyedi azonosítója, foreign key, amely a Language tábla ID mezőjére mutat.

**Az adatbázis ER modellje**

**Az adatbázis relációs modellje**

**Döntési fa a film nyelvére vonatkozóan**

A fa kézzel került összerakásra, a leírásához JSON formátumot választottam. Tanító halmaznak egy 20 elemű halmazt vettem, ami az adatbázisom utolsó 20 db filmjéből áll, azok alapján alakítottam ki az elágazásokat.

Az elkészült fa működésének hatékonyságát megpróbáltam lemérni. A többnyelvű filmeknél, pl. a "Magyar, Angol"-t egy nyelvnek vettem, hiszen a döntési fa kialakításánál is ahol több nyelv volt jelen egy filmben, azokat együttesen vettem egy nyelvként. Hatékonyságra az első 20 db filmet vizsgálva 85% jött ki eredményként, mivel az adatbázis ezen szakaszán nagyrészt angol nyelvű filmek voltak csak, így itt jobban működött a fa. A 61-80. filmre is megnéztem a hatékonyságát, itt láthatóan kevésbé működött jól a fa, csak 20%-os hatékonyságot adott, ami azzal magyarázható, hogy ebben az intervallumban sokkal vegyesebb nyelvű filmek szerepelnek, illetve sok olyan nyelvkombináció (pl. "Afrikai, Angol", "Angol, Japán") jelent meg, amelyek a döntési fában nem szerepeltek, hiszen az utolsó 20 filmnél nem voltak ezek jelen. A két szeletet együttvéve, a 40 filmet összességében vizsgálva 52,5%-os hatékonyság jön ki eredményként.