José Ortín López (1667573)
Arnau Muñoz Barrera (1665982)
Albert Vacas Martínez (1665473)

Grup 15

Informe Pràctica 2 GABD

Exercici 1

Per aquest primer exercici l'objectiu és implementar la funció *insertVectorDataset* per importar conjunts de dades des de l'UCI a una base de dades Oracle. Aquesta base de dades està formada per dues taules:

1. **Dataset**: Taula que emmagatzema el nom, el nombre de característiques (feat_size), el nombre de classes (numclasses) i metadates en format JSON (info)
2. **Samples**: Taula que emmagatzema les mostres individuals associades a un data set, incloent un vector de característiques (features) en format de text compatible amb Oracle (CLOB) i una etiqueta de classe (label).

La funció *insertVectorDataset* està formada principalment per les següents parts:

- Primer de tot, verifiquem si el dataset que es passa per paràmetre ja existeix a la base de dades mitjançant la funció *exists*. Aquesta funció assegura que la dataset ja està introduïda amb una consulta SQL que busca el nom del dataset a la taula dataset.
- Seguidament, fem l'extracció de dades del dataset. El dataset es carrega des de la llibreria *ucimlrepo* i s'extreu la informació següent: les dades de la mostra (matriu data), noms de les característiques (features_names), etiquetes i noms de les classes (target i target_names) i detalls del dataset (convertits a JSON per al camp info). Aquesta extracció de dades està implementada sota un mecanisme de tractament d'excepcions que ens permet identificar errors en l'extracció de dades i facilitar la feina en el moment de buscar errors dintre de la funció.
- Per tal d'assegurar la consistència de les dades fem una neteja d'aquestes amb un doble bucle en el qual validem que cada valor de tota la matriu de dades són caràcters i guardem aquestes dades filtrades en un nou array (*cleaned_data*).

- Amb el filtre de dades ja fet, la funció insereix el dataset a la taula *dataset*. S'obté un identificador únic mitjançant una seqüència creada anteriorment (seq). Després, les mostres s'insereixen una per una a la taula *samples*. De nou fem ús del mecanisme de tractament d'excepcions que ens permet identificar errors en la inserció de dades en les dues taules (*dataset* i *samples*) i facilitar la feina en el moment de buscar errors dintre de la funció.

Amb aquesta funció ens permet importar conjunts de dades de l'UCI a Oracle en la que garantim una validació prèvia de duplicats abans de la inserció i la conversió a dades a formats compatibles.

Exercici 2

Per aquest segon exercici l'objectiu és implementar la base de dades proposada aplicant certes millores a l'estructura, segons el nostre criteri i conveniència a l'hora de procedir en els següents exercicis.

Els canvis significatius que hem hagut d'aplicar han estat:

- Creació d'una clau composta a la taula REPETICIO, formada per ID_DATASET, que fa referència a la taula DATASET i el NUM_REPETICIO que fa referència al NUM de la taula REPETICIO.
  Aquesta clau ha estat creada ja que considerem que el més òptim per associar e identificar una repetició és identificar-la segons el seu número, però també, associar-la amb el dataset corresponent que es faci servir.
- També hem hagut d'afegir com a claus primàries a la taula EXPERIMENTS, l'ID_DATASET, NUM_REPETICIO, NOMCURT_CLASSIFICADOR i VALORS, a més dels 3 ja proveïts a l'esquema de l'enunciat.
  Aquestes claus han estat establertes ja que la nostra base de dades funciona identificant un experiment, segons un id d'un dataset corresponent, la repetició en la que es troba, el nom del classificador i els seus possibles valors.
- Finalment, també hem creat una clau primària composta a la taula PARAMETERS, que està formada pel NOMCURT_CLASSIFICADOR i VALORS, referenciant a les taules corresponents.
  Aquesta clau composta ens permet identificar a la taula parameters, segons la combinació del nom del classificador i dels valors, únicament.

José Ortín López (1667573)
Arnau Muñoz Barrera (1665982)
Albert Vacas Martínez (1665473)

Exercici 3

Per resoldre aquest exercici, hem modificat el script original **testUCI.py** per tal que es connecti a la base de dades Oracle i pugui recuperar els datasets de la UCI. Primer, hem implementat una funció en Python que carrega les dades des de la base de dades utilitzant consultes SQL amb un usuari Oracle amb privilegis restringits, **TestUCI**, qui només pot inserir i actualitzar dades als experiments però no modificar les taules **Dataset** i **Samples**. Després, hem desenvolupat una funció en Python que envia els resultats dels experiments a una funció en **PL/SQL**. Aquesta funció és responsable d'inserir o actualitzar els resultats a les taules corresponents, retornant un valor que indica si l'operació s'ha completat correctament.

A més, hem configurat l'script per realitzar fins a 50 repeticions per cada experiment, assegurant així la fiabilitat dels resultats. Finalment, hem verificat que l'script manté la interfície original, tal com s'especifica en l'enunciat, i hem implementat una gestió adequada dels permisos a nivell d'usuari i base de dades.

****** Degut als imprevistos que hem trobat i la manca de recursos que creiem que té la base de dades per inserir un volum tan gran de dades, és possible que no haguem pogut inserir completament totes les dades. Llavors, els resultats del exercici 3 s'han vist influènciats ja que, encara que els 3 primers datasets han estat completament inserits correctament, possiblement ens ha faltat la major càrrega de dades del dataset amb id 60. ******

Exercici 4

Per resoldre l'exercici, hem creat una vista materialitzada que mostra, per cada experiment, el dataset, el classificador, els paràmetres emprats, la data, i les mitjanes de les mètriques de rendiment (accuracy i f-score). Aquesta vista s'ha implementat amb l'usuari GestorUCI, que té el rol adequat per gestionar aquesta informació.

Un cop creada la vista materialitzada, hem desenvolupat una segona vista que calcula, per cada conjunt de dades i classificador, el millor conjunt de paràmetres segons la màxima accuracy, i que també inclou la desviació estàndard d'accuracy i f-score.

A l'hora de valorar l'eficiència i rendiment de les consultes amb i sense índexs, hem hagut de crear un pla d'execució concret a tractar i l'hem analitzat. L'estudi del rendiment ens permet observar:

```
Index IDX_CLASSIFICADOR borrado.


Index IDX_PARAMETERS borrado.


Explicado.


PLAN_TABLE_OUTPUT
------------------------------------------------------------------------------------------
Plan hash value: 22193053

------------------------------------------------------------------------------------------
| Id  | Operation            | Name                | Rows  | Bytes | Cost (%CPU)| Time     |
------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT     |                     |    71 |  5538 |     3   (0)| 00:00:01 |
|   1 |  MAT_VIEW ACCESS FULL| VISTAMATERIALITZADA |    71 |  5538 |     3   (0)| 00:00:01 |
------------------------------------------------------------------------------------------
```

```
Index IDX_CLASSIFICADOR creado.


Index IDX_PARAMETERS creado.


Explicado.


PLAN_TABLE_OUTPUT
------------------------------------------------------------------------------------------
Plan hash value: 22193053

------------------------------------------------------------------------------------------
| Id  | Operation            | Name                | Rows  | Bytes | Cost (%CPU)| Time     |
------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT     |                     |    71 |  5538 |     3   (0)| 00:00:01 |
|   1 |  MAT_VIEW ACCESS FULL| VISTAMATERIALITZADA |    71 |  5538 |     3   (0)| 00:00:01 |
------------------------------------------------------------------------------------------
```

Que en les dos casos el temps i cost de l'execució és el mateix, permeten entendre que en el nostre cas no suposa un canvi o millora significativa.

Entenem que en casos de volums de dades encara més grans i significatius, aquest adició si que suposaria la diferència entre un funcionament òptim o funcionament lent, però pel nostre context concret, no varia.

Finalment, hem assegurat que les vistes segueixen el format requerit per l'enunciat, retornant els valors mitjans i la desviació estàndard de les mètriques de forma clara i organitzada de manera concatenada amb els valors i a una mateixa columna de les vistes.

José Ortín López (1667573)
Arnau Muñoz Barrera (1665982)
Albert Vacas Martínez (1665473)

Autoavaluació

	Exercici 1	Exercici 2	Exercici 3	Exercici 4
Nota	3/3	1/1	4/4	2/2

Autors: Grup 15

Arnau Muñoz Barrera - 1665982

Pep Ortín López - 1667573

Albert Vacas Martínez - 1665473