BIRMINGHAM CITY
University

CMP6230 Draft Pipeline

Lewis Higgins - Student ID 22133848

Contents

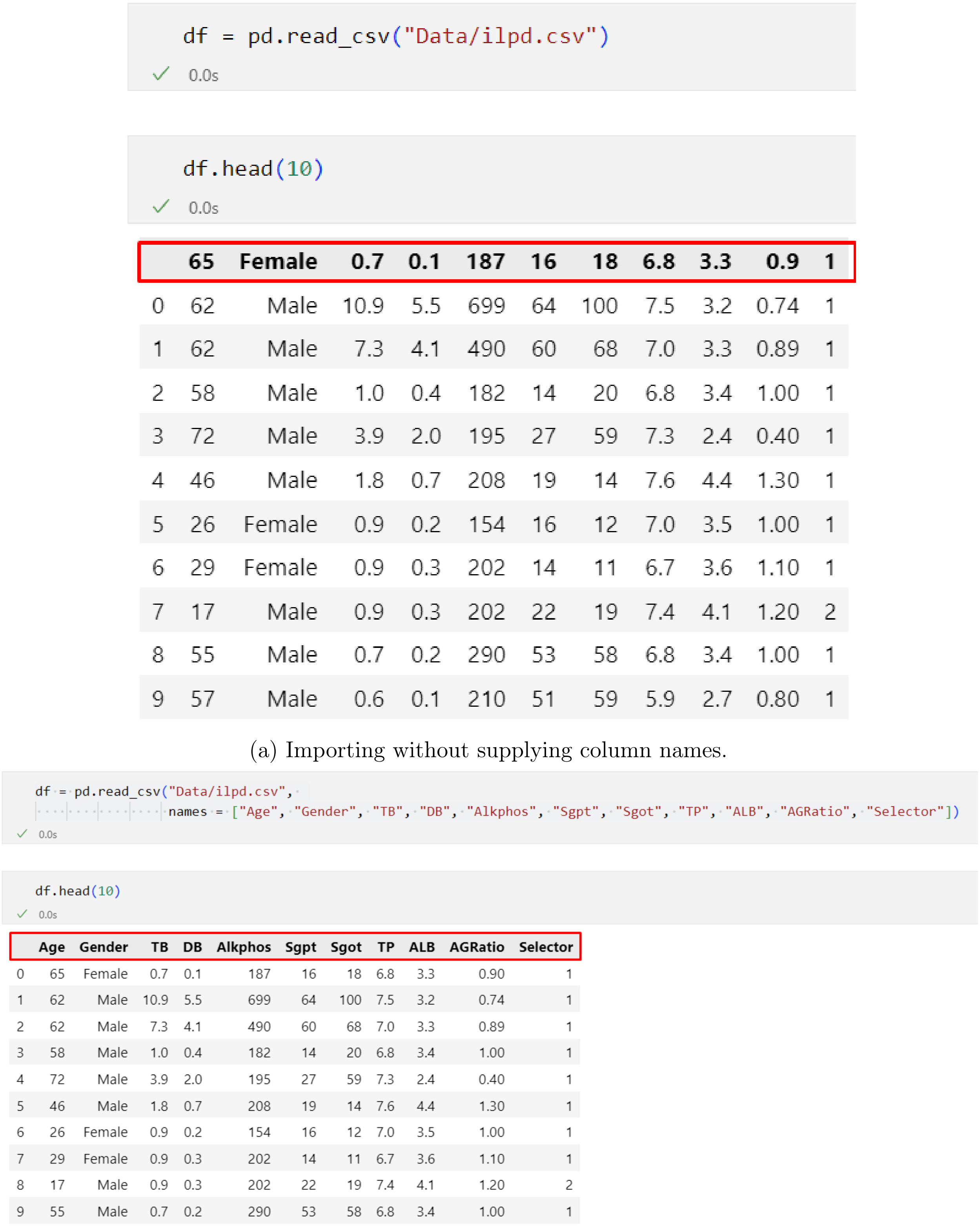
Candidate Data Sources

For the first stage of the pipeline, data ingestion, three data sources will be identified in order to find the one that would be most optimal for the production and deployment of a machine learning model to complete a supervised learning task.

1.1 Candidate 1 - Indian Liver Patient Dataset

This dataset consists of real data sourced from hospitals northeast of Andhra Pradesh in India. It was obtained from the UCI Machine Learning Repository (Bendi Ramana and N. Venkateswarlu, 2022), and has been previously used by Straw and Wu (2022) in their analysis of sex-related bias in supervised learning models. The UCI ML Repository is a popular host of datasets used by students, educators and researchers worldwide for machine learning (UCI Machine Learning Repository, 2024), and hosts these datasets on the cloud for public download and usage, as long as credit is given. This dataset in particular aims to assist in the diagnosis of liver disease due to increasing mortality rates from conditions like liver cirrhosis, and contains 584 records with 10 features as well as the "Selector" classification column, where those wihout liver disease are classed as 1, and those with liver disease are classed as 2. For the purposes of the ML model, these can be changed to 0 and 1 respectively. The dataset is a single flat-file Comma-Seperated Values (CSV) file, which stores data by seperating each column with commas and each row with line breaks. This CSV file uses a One Big Table (OBT) schema, as seen in the entity relationship diagram in Figure **??**, wherein all of the data within this dataset is stored in a single table. Descriptions of the columns in the dataset, as well as the associated data types, can be found in Table 1.1.

A minor issue with this file is that it has no headers in its CSV file, meaning that when imported, Pandas will interpret the first row of data as the names of the columns, though this can be combated by adding the "names" argument when calling Pandas' "read_csv" function, seen below in Figure 1.1a.



```
df = pd.read_csv("Data/ilpd.csv")
✓ 0.0s
```

```
df.head(10)
✓ 0.0s
```

	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
0	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
1	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
2	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
3	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
4	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.30	1
5	26	Female	0.9	0.2	154	16	12	7.0	3.5	1.00	1
6	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.10	1
7	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.20	2
8	55	Male	0.7	0.2	290	53	58	6.8	3.4	1.00	1
9	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.80	1

(a) Importing without supplying column names.

```
df = pd.read_csv("Data/ilpd.csv",
                 names = ["Age", "Gender", "TB", "DB", "Alkphos", "Sgpt", "Sgot", "TP", "ALB", "AGRatio", "Selector"])
✓ 0.0s
```

```
df.head(10)
✓ 0.0s
```

	Age	Gender	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	AGRatio	Selector
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
5	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.30	1
6	26	Female	0.9	0.2	154	16	12	7.0	3.5	1.00	1
7	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.10	1
8	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.20	2
9	55	Male	0.7	0.2	290	53	58	6.8	3.4	1.00	1

(b) Importing with the column names.

Figure 1.1: Importing the erroneous CSV using Pandas. The column headers are highlighted in a red box.

Column	Format	Description
Age	Integer	The patient's age. **Ages of 90 or over were listed as 90 before this dataset was published.**
Gender	String (Binary)	The patient's gender, either "Male" or "Female".
TB	Float	Total bilirubin. Bilirubin is a substance produced by the liver, and a high presence of it may be indicative of liver problems (Mayo Clinic, 2024).
DB	Float	Direct bilirubin. This is a slightly different form of bilirubin that is formed after the liver has processed it.
Alkphos	Integer	Levels of alkaline phosphate - an enzyme in the body produced by the liver. Too much may indicate liver disease. (Cleveland Clinic, 2024)
Sgpt	Integer	Another enzyme found in the liver, where too much can indicate liver problems.
Sgot	Integer	Levels of AST in the blood, where too much indicates liver problems.
TP	Float	Total proteins.
ALB	Float	Albumin - a protein in blood plasma. Too little of this may indicate liver problems.
A/G Ratio	Float	The ratio of albumin to globulin, which is another blood protein.
Selector	Integer	The classifier, indicating if the person has liver disease. The target column for the ML model.

Table 1.1: The data types of each column in the Indian Liver Patient Dataset.

This dataset can be used to develop a supervised machine learning model for binary classification using the ten predictor variables and the ground truth Selector column, which will be used in measuring the accuracy of the model. There is a clear positive purpose for developing such a model; as previously mentioned, mortality rates from liver disease are high, and an early diagnosis that could leverage the power of machine learning can greatly enhance the odds of successful treatment.

1.2 Candidate 2 - Loan Approval Classification Dataset

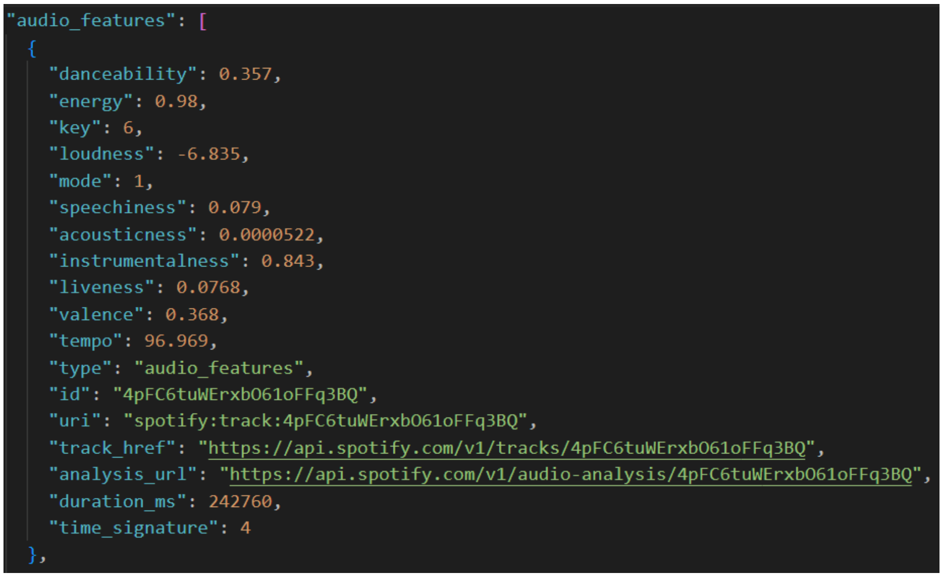
This dataset was sourced from Kaggle's cloud servers under an Apache 2.0 license, which states that the dataset can be used as long as credit is given to the original author, and takes the form of a flat-file CSV using a One Big Table schema. Unlike Candidate 1, this dataset does not consist of real data, and instead consists of synthetic data. This is likely due to the fact that this dataset, if it used real data, would contain extremely personal information that could not be shared online due to legislation such as GDPR. This particular dataset is an enhanced version of a different credit risk dataset, which also did not provide an original source and is presumably synthetic data. The dataset consists of 45,000 records and 14 features, with one of these being the ground truth target variable "loan_status", which is whether the person should be given a loan or not. As such, it is well suited for a binary classification model, using the first 13 features as predictor variables. This can also be observed from the 28 notebooks on Kaggle that utilise this dataset. The data types for each column can be seen in the entity relationship diagram in Figure ?? and descriptions of each column can be seen in Table 1.2.

Column	Description
person_age	The age of the person.
person_gender	The person's gender.
person_education	The person's highest level of education.
person_emp_exp	The person's years of employment experience.
person_home_ownership	Home ownership status (for example rent, own, mortgage)
loan_amnt	The amount of money requested.
loan_intent	The purpose of the loan.
loan_int_rate	The interest rate of the loan.
loan_percent_income	Loan amount as a percentage of the person's yearly income.
cb_person_cred_hist_length	Length of credit history in years.
credit_score	Credit score of the person.
previous_loan_defaults_on_file	If the person has defaulted on a loan before.
loan_status	Whether the loan should be approved. 1 if yes, 0 if no.

Table 1.2: The descriptions of each column in the dataset.

1.3 Candidate 3 - Spotify Likes Dataset

This dataset was sourced from Kaggle, a platform similar to the UCI ML repository in its purpose for students and researchers that acts as a search engine for datasets, but also allows its users to host competitions, upload their machine learning models, and also upload their own Python notebooks. This dataset is stored on their servers on the cloud, and is free to download and use. The data itself is split over a CSV file and two JavaScript Object Notation (JSON) files. JSON files store data in **key-value pairs**, such as in the example snippet of this dataset depicted in Figure 1.2.



```
"audio_features": [
  {
    "danceability": 0.357,
    "energy": 0.98,
    "key": 6,
    "loudness": -6.835,
    "mode": 1,
    "speechiness": 0.079,
    "acousticness": 0.0000522,
    "instrumentalness": 0.843,
    "liveness": 0.0768,
    "valence": 0.368,
    "tempo": 96.969,
    "type": "audio_features",
    "id": "4pFC6tuWErxbO61oFFq3BQ",
    "uri": "spotify:track:4pFC6tuWErxbO61oFFq3BQ",
    "track_href": "https://api.spotify.com/v1/tracks/4pFC6tuWErxbO61oFFq3BQ",
    "analysis_url": "https://api.spotify.com/v1/audio-analysis/4pFC6tuWErxbO61oFFq3BQ",
    "duration_ms": 242760,
    "time_signature": 4
  },
```

Figure 1.2: A snippet of the JSON data, viewed in Visual Studio Code.

Every row in the JSON files is part of "audio_features", and is seperated by curly braces {}. Each column is then given as a key-value pair, such as the first row in the image, where "danceability" is the key, and 0.352 is the associated value.
This dataset does consist of real data, sourced from the author's personal liked songs directly via the Spotify API. There are 195 rows of data, with 100 liked songs, and 95 disliked songs. Liked and disliked songs are seperated into two JSON files, named "dislike" and "good". The CSV has 14 features, though the JSON files have 18, depicted in Figure **??**.

While a machine learning classification problem can definitely be performed on this dataset to identify if the author would like a song, it has significantly less of a positive impact than Candidates 1 and 2, as this dataset is the author's subjective belief rather than objective fact that can be applied to other people. Nevertheless, the data types and descriptions of each column can be found in Table 1.3.

Column	Format	Description
Danceability	Float	How suitable a song is for dancing, calculated from the tempo, rhythm stability, beat strength and overall regularity. 1.0 means it is very danceable.
Energy	Float	The intensity and activity of a song. For example, death metal is high energy, whereas classical music is low intensity. 1.0 is the most energetic.
Key	Integer	The musical key the song is in, converted to an integer using standard pitch class notation.
Loudness	Float	The averaged decibel volume of a song, typically between -60 and 0 dB.
Mode	Integer (Binary)	Whether a song is in major or minor scale. 1 is major and 0 is minor.
Speechiness	Float	The calculated presence of spoken words in a song.
Acousticness	Float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Instrumentalness	Float	Whether a song has no vocals.
Liveness	Float	Whether a live audience can be heard as part of a song.
Valence	Float	The musical positiveness of a song.
Tempo	Float	The beats per minute of a song.
Duration_MS	Integer	The duration of a song in milliseconds.
Time signature	Integer	The estimated time signature of the song.
Liked	Integer (Binary)	The target variable, indicative of whether the author liked the song or not.
Type	String	Always "audio_features". Not a relevant predictor.
ID	String	Spotify's own unique ID for a song. Not a relevant predictor.
URI	String	Spotify's URI for the song. Not a relevant predictor.
Track HREF	String	A link to the song on Spotify's API. Not a relevant predictor.
Analysis URL	String	A link to the song's audio analysis data. Not a relevant predictor.

Table 1.3: The descriptions of each column in the Spotify songs dataset (Spotify, 2024). Red columns are only present in the CSV, whereas green columns are only present in the JSONs.

These measurements and the descriptions are part of Spotify's API, and are automatically calculated when songs are uploaded to the service. The ground truth of the dataset is present in the CSV file as the "liked" classifier column, and a train/test split can be implemented for predictions, which is aided by the fact that this dataset is well balanced (100 liked to 95 disliked).

Datasets

Bendi Ramana and N. Venkateswarlu (2022). *ILPD (Indian Liver Patient Dataset)*. DOI: 10.24432/C5D02C.

Bibliography

Bendi Ramana and N. Venkateswarlu (2022). *ILPD (Indian Liver Patient Dataset)*. DOI: 10.24432/C5D02C.

Cleveland Clinic (2024). *Alkaline Phosphatase (ALP): What It Is, Causes & Treatment*. Cleveland Clinic. URL: https://my.clevelandclinic.org/health/diagnostics/22029-alkaline-phosphatase-alp (visited on 11/12/2024).

Mayo Clinic (2024). *Bilirubin test - Mayo Clinic*. URL: https://www.mayoclinic.org/tests-procedures/bilirubin/about/pac-20393041 (visited on 11/12/2024).

Spotify (2024). *Web API Reference | Spotify for Developers*. URL: https://developer.spotify.com/documentation/web-api/reference/get-audio-features (visited on 11/12/2024).

Straw, Isabel and Honghan Wu (Apr. 25, 2022). "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction". In: *BMJ Health & Care Informatics* 29 (1). Publisher: BMJ Publishing Group Ltd. ISSN: 2632-1009. DOI: 10.1136/bmjhci-2021-100457.

UCI Machine Learning Repository (2024). *About - UCI Machine Learning Repository*. URL: https://archive.ics.uci.edu/about (visited on 11/12/2024).