Proyectos finales

1) 1025 Pokemon Stats (all_pokemon_data.csv)

I used PokeAPI to first pull all the current Pokémon (as of March 2025). Each was checked for an extra forms (regional, mega, and any other forms that have stat/typing changes) and added a row with their stats as well. The information pulled are name, national dex (repeats for Pokémon with multiple forms), typing, generation, evolution stage, number of evolutions, color ID, catch rate, height (dm), weight (hg), BST, health, attack, defense, special attack, special defense, speed. I added a secondary typing flag, legendary status flag, what alternative form they have, alternative form flag for any other than base, height in inches, and weight in pounds. Below is a more detailed description of each column.
Huge thanks to PokeAPI for their tool!

- Name: The Pokémon's name, if they have an alternate form it is added with a dash. All lower-case text entries.
- National Dex: The ID per the national dex. Alternate forms will have MATCHING ID numbers with their original forms (i.e. charizard and mega charizard-x and mega charizard-y all share the dex ID 6)
- Primary Typing: The Pokémon's primary typing.
- Secondary Typing: The Pokémon's secondary typing. Left blank if single typing.
- Secondary Typing Flag: True if there is a secondary typing, False if solo typing.
- Generation: Marked with generation-# in roman numerals. (i=1, ii=2, iii=3, iv=4, v=5, vi=6, vii=7, viii=8, ix=9)
- Legendary Status: True if legendary, mythical, mega, ultra beast, or paradox. False if not
- Form: If there are is an alternate form with a stat change it will be listed with it (also used to change the name). (zygarde-50, zygarde-10-power-construct, zygarde-complete, zygarde-10 have the forms Base, 10-power-construct, Complete, 10)
- Alt Form Flag: True if it's not the base form.
- Evolution Stage: one, two, or three at it's current stage. (charmander = 1, charmeleon = 2, charizard (base and mega X/Y) = 3)
- Number of Evolutions: Total evolutions in the line (extra forms not included) with the options of one, two, or three. (charmander, charmeleon, charizard (base and mega X/Y) all display 3 as there are three total)
- Color ID: According to Niantic's coloring system which is mildly a mystery.
- Catch Rate: The Pokémon species modifier for the catch rate equation. The higher the number the easier it is to catch that species of Pokémon. (Learn More).
- Height (dm): Pokémon's height in decimeters. This was from PokeAPI.
- Weight (hg): Pokémon's weight in hectograms. This was from PokeAPI.

- Height (in): Pokémon's height in inches. This was converted from dm by multiplying by 3.93701.
- Weight (lb): Pokémon's weight in pounds. This was converted from hg by multiplying by 0.220462.
- Base Stat Total: Total of the in-game inherent species stats referred to as Base Statistics (Health, Attack, Defense, Special Attack, Special Defense, Speed).
- Health: Total health that a species can start with. Can be modified by EVs.
- Attack: Total attack that a species can start with. Can be modified by EVs.
- Defense: Total defense that a species can start with. Can be modified by EVs.
- Special Attack: Total special attack that a species can start with. Can be modified by EVs.
- Special Defense: Total special defense that a species can start with. Can be modified by EVs.
- Speed: Total speed that a species can start with. Can be modified by EVs.

2) Music Genre Classification (tres archivos csv: test, train y submission)

Context
Dataset is acquired from one of the MachineHack Hackathon

Content
Training dataset: 17,996 rows with 17 columns

Column details: artist name; track name; popularity; ‘danceability’; energy; key; loudness; mode; ‘speechiness’; ‘acousticness’; ‘instrumentalness’; liveness; valence; tempo; duration in milliseconds and time_signature.

Target Variable: 'Class’ such as Rock, Indie, Alt, Pop, Metal, HipHop, Alt_Music, Blues, Acoustic/Folk, Instrumental, Country, Bollywood,

Test dataset: 7,713 rows with 16 columns

3) Dry Bean Dataset Classification (Dry_Bean_Dataset.csv)

Seven different types of dry beans were used in this research, taking into account the features such as form, shape, type, and structure by the market situation. A computer vision system was developed to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification. For the classification model, images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. Bean images obtained by computer vision system were subjected to segmentation and feature extraction stages, and a total of 16 features; 12 dimensions and 4 shape forms, were obtained from the grains.

4) SPAM E-mail Database (spambase_csv.csv)

The “spam” concept is diverse: advertisements for products/websites, make money fast schemes, chain letters, pornography… Our collection of spam e-mails came from our postmaster and individuals who had filed spam. Our collection of non-spam e-mails came from filed work and personal e-mails, and hence the word ‘george’ and the area code ‘650’ are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

Attribute Information:

The last column denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

For the statistical measures of each attribute, see the end of this file. Here are the definitions of the attributes:

48 continuous real [0,100] attributes of type

word_freq_WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A “word” in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurrences) / total characters in e-mail

1 continuous real [1,…] attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,…] attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters

1 continuous integer [1,…] attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.