Pomysł na pracę magisterską

Prognoza produkcji/cen na podstawie danych pogodowych

# Dane

Możliwe dane do wykorzystania można podzielić na kategorie:

* Dane pogodowe
* Dane agronomiczne
* Dane ekonomiczne

## Dane pogodowe

Wybrane dane open source na temat zmian klimatycznych w różnych regionach świata, najlepiej z okresu ostatnich 30 lat. Dane mogą być naziemne oraz satelitarne. Można również zdecydować się na własny zbiór danych powstały z aproksymacji obu typów danych pogodowych. Przydatne linki:

* <https://www.chc.ucsb.edu/data/chirps> - Dane CHRIPS na temat szacunków ilości opadów oraz obserwacji satelitarnych.
* <http://iridl.ldeo.columbia.edu/SOURCES/> - ogromny zbiór wszystkich możliwych do importu danych pogodowych open source.
* <https://www.cpc.ncep.noaa.gov/products/international/africa/africa.shtml> - strona centrum prognoz klimatu zorientowane na obserwację regionów Afryki.

Dane dotyczące NOAA - National Oceanic and Atmospheric Administration znajdują się już w bazie danych firmy. Dane dotyczą dziennych średnich parametrów dla różnych stacji na Ziemii.

**Liczebność danych NOAA**: 51816218

**Zakres dat**: 2000-2023

|  |  |  |
| --- | --- | --- |
| **Nazwa** | **Typ** | **Opis** |
| USAF | object | Identyfikator stacji pogodowej |
| WBAN | object | Identyfikator regionu |
| YEAR | int64 | - |
| MONTH | int64 | - |
| DAY | int64 | - |
| TEMP | float64 | Średnia temperatura (.1 Fahrenheit) |
| TEMP\_Count | int64 | Ilość obserwacji wykorzystanych do obliczenia średniej temperatury |
| DEWP | float64 | Średni punkt rosy (.1 Fahrenheit) |
| DEWP\_Count | int64 | Liczba obserwacji wykorzystanych przy obliczaniu średniego punktu rosy |
| SLP | float64 | Średnie ciśnienie na poziomie morza (.1 milibar) |
| SLP\_Count | Int64 | Liczba obserwacji wykorzystanych przy obliczaniu średniego ciśnienia na poziomie morza |
| STP | float64 | Średnie ciśnienie w stacji (.1 milibar) |
| STP\_Count | Int64 | Liczba obserwacji wykorzystanych przy obliczaniu średniego ciśnienia w stacji |
| VISIB | float64 | Średnia widoczność (.1 mila) |
| VISIB\_Count | Int64 | Liczba obserwacji wykorzystanych przy obliczaniu średniej widoczności |
| WDSP | float64 | Średnia prędkość wiatru (.1 węzła) |
| WDSP\_Count | Int64 | Liczba obserwacji wykorzystanych przy obliczaniu średniej prędkości wiatru |
| MXSPD | float64 | Maksymalna utrzymująca się prędkość wiatru (.1 węzła) |
| GUST | float64 | Maksymalny poryw wiatru (.1 węzła) |
| MAX | Float64 | Maksymalna temperatura (.1 Fahrenheit) |
| MAX\_FLAG | object | Nie do końca wiadomo. Do dyskusji. |
| MIN | float64 | Minimalna temperatura (.1 Fahrenheit) |
| MIN\_FLAG | object | Nie do końca wiadomo. Do dyskusji. |
| PRCP | float64 | Całkowity opad (deszcz i/lub stopiony śnieg) (.01 cal) "0" oznacza brak mierzalnych opadów (w tym śladowe ilości). |
| PRCP\_Flag | object | Nie do końca wiadomo. Do dyskusji. |
| SNDP |  | Głębokość śniegu (.1 cal). Jest to ostatni raport dla danego dnia, jeśli został zgłoszony więcej niż raz. |
| Fog | bool | Wskaźnik zdarzenia: Mgła |
| RainOrDrizzle | bool | Wskaźnik zdarzenia: Deszcz lub mżawka |
| SnowOrIcePellets | bool | Wskaźnik zdarzenia: Śnieg lub ziarna lodowe |
| Hail | bool | Wskaźnik zdarzenia: Grad |
| Thunder | bool | Wskaźnik zdarzenia: Burza |
| TornadoOrFunnelCloud | bool | Wskaźnik zdarzenia: Tornado lub chmura lejkowata |
| ID | Int64 | Unikalny identyfikator instancji |
| EntryDate | datetime64[ns] | Data wprowadzenia do systemu |
| ts | object | Nie do końca wiadomo. Do dyskusji. |

## Dane agronomiczne

Dane dotyczące policzalnych metryk dotyczących poszczególnych plantacji. Źródłem są ręczne kalkulacje. W przypadku tych danych problemem może okazać się liczebność zbiorów danych w zależności od regiony/kraju. Niektóre regiony dostarczyły na przestrzeni lat o wiele mniej danych, połowa rekordów to dane dotyczące Brazylii. Około 22000 rekordów dotyczących Côte d'Ivoire.

**Liczebność danych**: 111544

**Zakres dat**: 1997-2023

|  |  |  |
| --- | --- | --- |
| **Nazwa** | **Typ** | **Opis** |
| IsTempRecord | bool | W domyśle wartość logiczna określająca czy dany rekord jest tymczasowy |
| Reporter | object | Skrót państwa z którego pochodzi raport |
| Country | object | Państwo z którego pochodzi raport |
| State | object | Region z którego pochodzi raport |
| Week | int64 | Nie do końca wiadomo, do dyskusji. Być może wiek drzewka w tygodniach |
| Year | int64 | - |
| Month | int64 | - |
| SiteLetter | object | Podział obszaru na literę |
| SiteNumber | Int64 | Podział obszaru na liczbę |
| TreeNumber | Int64 | Numer drzewka |
| PCSample | object | Nie wiadomo, do dyskusji. |
| Variety | object | Nie wiadomo, do dyskusji. |
| Moisture | float64 | Z nazwy wynika, że jest to poziom zawilgocenia |
| Texture | float64 | Nie do końca wiadomo, poziom/rodzaj tekstury? |
| Canopy | float64 | Nie do końca wiadomo, stan korony? |
| Maintenance | float64 | Nie do końca wiadomo, poziom utrzymania? |
| Flowers | float64 | Liczba kwiatów |
| Leaf flushing | float64 | Nowe liście |
| WB | float64 | Nie wiadomo, do dyskusji |
| Monilia T/S | float64 | Rodzaj grzyba atakującego owoce kakaowca (prawdopodobnie tiny/small) |
| Monilia M/L | float64 | Rodzaj grzyba atakującego owoce kakaowca (prawdopodobnie medium/large) |
| Monilia M | Float64 | Rodzaj grzyba atakującego owoce kakaowca (prawdopodobnie medium) |
| Black Pod | Float64 | Zebsuty owoc kakaowca |
| Other D/D | Float64 | Nie wiadomo, do dyskusji |
| Harvest | bool | Czy owoce kakaowca zostały już zebrane? |
| Drying | bool | Czy owoce kakaowca suszą się? |
| Fermentation | bool | Czy zaczął się proces fermentacji owoców kakaowca? |
| Fertilizer | bool | Czy drzewko zostało nawiezione? |
| Fungicide | bool | Czy drzewko otrzymało środek grzybobójczy? |
| Irrigation | bool | Czy przeprowadzone zostało nawadnianie/irygacja drzewka? |
| Tiny | Float64 | Ilość najmniejszych owoców |
| Small | Float64 | Ilość małych owoców |
| Medium | Float64 | Ilość średnich owoców |
| Large | Float64 | Ilość dużych owoców |
| Mature | Float64 | Ilość dojrzałych owoców |
| Ripe | Float64 | Ilość dojrzałych owoców, nie dokońca zrozumiała różnica między Mature a Ripe, do dyskusji |
| Wilt | Float64 | Liczba zwiędłych owoców |
| Damaged | Float64 | Liczba uszkodzonych owoców |
| Healthy | Float64 | Liczba zdrowych owoców |
| Total | Float64 | Suma owoców |
| Cherelles | Float64 | Nie wiadomo, do dyskusji |
| Adults | Float64 | Nie wiadomo do czego się odnosi, do dyskusji |
| ImportDate | datetime64[ns] | Data wprowadzenia do bazy |
| ID | int64 | Unikalny identyfikator rekordu |
| ts | object | Nie do końca wiadomo. Do dyskusji. |

## Dane ekonomiczne

Dane dotyczące cen rynkowych różnych produktów, w tym tych pochodzących od owoców kakaowca (proszki naturalne, alkalizowane, itd.).

**Liczebność danych**: 1380446

**Zakres dat:** 2012-2023

Najważniejszymi kolumnami są głównie: **data wprowadzenia, data aktualizacji, identyfikator osoby wprowadzającej/aktualizującej, data wymiany, identyfikator miejsca wymiany, kod produktu, CENA, jednostka ilości/wagi produktu.**

# Pomysły

* Wykorzystanie zmieszanych/interpolowanych danych pochodzących z zarówno źródła satelitarnego jak i naziemnego. Porównanie działania modelu na różnych danych pogodowych.
* Porównanie podejścia klasycznego analizy regresji vs. Prosty model uczenia maszynowego vs. Rekurencyjna sieć neuronowa
* Prognozowanie danych agronomicznych tzn. Np. Ilość zebranych owoców kakaowca, chorych, prawdopodobieństwo posiadania choroby, ilość z rozróżnieniem w wielkościach owoców itd.
* Prognozowanie danych agronomicznych dla krajów/regionów o mniejszej liczbie rekordów z wykorzystaniem metod nadpróbkowania, np. SMOTE (o ile da się zastosować do szeregów czasowych)
* Prognozowanie danych ekonomicznych tzn. Cena produktów na podstawie danych pogodowych
* Połączenie danych pogodowych oraz danych agronomicznych w celu prognozy danych ekonomicznych.
* Porównanie na które produkty (z perspektywy ceny) bardziej wpływa pogoda a na które mniej.