Lab 10 - Projekt podsumowujący zajęć dotyczących testowania hipotez

Cel projektu

Celem projektu jest analiza datasetu Gender, Mental Illness, and Crime in the United States, 2004 (ICPSR 27521). Dane można pobrać z serwisu lub z lokalnej kopii.

Próba: NATIONAL HOUSEHOLD SURVEY ON DRUG USE AND HEALTH (NSDUH), 2004 [ICPSR 4373] obejmuje 67 760 osób. Plik do użytku publicznego zawiera 55 602 rekordów ze względu na resampling stosowany w procesie anonimizacji. Próba jest stratyfikowana na wielu poziomach, zaczynając od stanów. Osiem stanów jest uważanych za stany o dużej próbie i wnoszą około 3600 respondentów na stan. W pozostałych stanach próba obejmuje 900 respondentów. Proces próbkowania respondentów był prowadzony w sposób systematyczny. Każdy respondent, który ukończył pełny wywiad, otrzymał 30 USD gotówki. Próbkowanie obejmowało pięć grup wiekowych: 12–17 lat, 18–25 lat, 26–34 lata, 35–49 lat oraz 50 lat i więcej. Projekt próby obejmował mniej więcej równą liczbę osób w grupach wiekowych 12–17 lat, 18–25 lat i 26 lat i starszych.

Opis wszystkich kolumn można znaleźć w dokumencie `27521-0001-Codebook.pdf` Zestaw danych zawiera łącznie 3011 zmiennych. Pierwsze 2690 zmiennych pochodzi z `NATIONAL HOUSEH SURVEY ON DRUG USE AND HEALTH (NSDUH), 2004 [ICPSR 4373]` , a pozostałe 321 zmiennych zostało stworzonych przez głównego badacza. Pytania z NSDUH z 2004 r. obejmowały wiek w momencie pierwszego użycia, a także używanie kilku klas narkotyków w ciągu całego życia, coroczne i w ciągu ostatniego miesiąca. Ankieta obejmowała historię leczenia uzależnień i postrzeganą potrzebę leczenia oraz zawierała pytania z Diagnostic and Statistical Manual (DSM) of Mental Disorders, które umożliwiają zastosowanie kryteriów diagnostycznych. Ankieta zawierała pytania dotyczące leczenia zarówno uzależnień, jak i zaburzeń związanych ze zdrowiem psychicznym. Respondenci byli również pytani o nielegalne działania i rejestr aresztowań, problemy wynikające z używania narkotyków i dzielenie się igłami. Pytania obejmowały również różne tematy, takie jak środowisko sąsiedzkie, nielegalne działania, zażywanie narkotyków przez przyjaciół, wsparcie społeczne, zajęcia pozalekcyjne, narażenie na programy zapobiegania nadużywaniu substancji i programy edukacyjne oraz postrzegane postawy dorosłych wobec zażywania narkotyków i zajęć, takich jak praca w szkole. Kilka pytań koncentrowało się również na tematach związanych z zapobieganiem. Zachowano również pytania dotyczące zdrowia psychicznego i dostępu do opieki, postrzeganego ryzyka zażywania narkotyków, postrzeganej dostępności narkotyków, prowadzenia pojazdów i zachowania osobistego oraz palenia cygar. Uwzględniono również zmienne demograficzne i informacje ogólne, takie jak płeć, rasa, wiek, pochodzenie etniczne, stan cywilny, poziom wykształcenia, status zawodowy, status weterana i obecny skład gospodarstwa domowego. Zmienne utworzone przez głównego badacza są zagregowanymi danymi z pierwotnego zestawu 2690 zmiennych. Konkretnie zmienne te obejmują wskaźniki depresji, wskaźniki uzależnienia od narkotyków, interakcje z płcią i innymi zmiennymi demograficznymi oraz zmienne odnoszące się do typów nadużywania narkotyków i zachowań przestępczych.

Informacje dotyczące opracowania i prezentacji danych

Ostateczny raport z analizy przedstaw w formie jupyter notebook, możesz jednak dołączyć dodatkowe pliki ze skryptami, z których korzysta notatnik (przy założeniu że znajdują się one w tym samym katalogu co notatnik). Jeśli wykorzystujesz dodatkowe dane, które nie zostały załączone do datasetu to należy je również załączyć (gdy dane są wczytywane w notatniku powinny być zaczytywane z tego samego katalogu w którym znajduje się notatnik). Jeśli do pozyskania/przetworzenia danych z których później korzystasz w notatniku wykorzystałeś własne skrypty to również powinny być dołączone z krótką informacją dotyczącą ich przeznaczenia. Jeśli analizujesz istotność danego czynnika, np. istotność grupy wiekowej, środowiska zycia, dochodu czy też doświadczeń w zażywaniu narkotyków pamiętaj o tym, że wnioski muszą opierać się na analizie przedziałów ufności (w przypadku czynników analizowanych przy pomocy regresji), lub wynikach testów statystycznych. Zastosowany sposób analizy powinien zostać krótko uzasadniony.

Zadania

Temat projektu określa pewne ramy, które powinien objąć, szczegółowe rozwiązania proszę zrealizować w oparciu o dotychczas pozyskaną wiedzę, własną widzę dziedzinową oraz ew. dodatkowe dane, które dostarczają informacji o innych potencjalnie ważnych czynnikach, które mogą mieć wpływ na rozwój depresji.

W analizie danych poszukujemy czynników wpływających na rozwój depresji, tak, żeby po ich określeniu można było dokonać oceny ryzyka depresji w danej grupie wiekowej. Szczegółowy opis cech można znaleźć w CookBook oraz na stronie icpsr.

Jako wskaźnik wystąpienia depresji możesz wykorzystać m.in. następujące zmienne (pamiętaj, żeby nie używać ich jako zmiennych niezależnych w modelu umożliwiającym rozpoznanie depresji):

- DEPRESSIONINDEX obejmujący połączenie indeksu dla wszystkich grup wiekowych - wskaźnik natężenia depresji w skali 0-9 dla grupy dorosłych i w wieku młodzieńczym - należy zwrócić uwagę na ankiety bez udzielonej odpowiedzi kodowane jako -9
- DEP_EPISODE doświadczenie epizodu depresji w okresie całego zycia
- MDELastYr - epizod depresji w ostatnim roku
- ANYTXRXMDE - jakiklolwiek zdarzenie zawiązane z leczeniem depresji lub receptą na leki antydepresyjne w minionym roku
- Warto tutaj dokonać rozróżnienia na depresję somatyczną, która charakteryzuje się równoczesnym występowaniem zaburzeń apetytu, problemów ze snem oraz zmęczenia oraz depresję niesomatyczną przy której te 3 objawy nie występują równocześnie.

Jako wskaźnik grupy wiekowej używaj: CATAG2, który dzieli populację na 3 równoliczne grupy 12-17, 18-25, powyżej 25 lat lub CATAG3 zawierających równoliczne 5 kategorie wiekowe lub CATAG7 wyróżniających 7 kategorii wiekowych. Grupa wiekowa (12-17 (youth)) ma w pewnych obszarach inne zestawy pytań niż grupy starsze (rozróżnienie jest kodowane w nazwach kolumn YOxxx lub ADxx).

Jako wskaźnik płci: IRSEX

Jako wskaźnik rasy: NEWRACE2

Jako wskaźnik uzależnienia od narkotyków i alkoholu:

- ANYINDEX - wskaźnik uzależnienia od dowolnego rodzaju narkotyków 0-1
- doświadczenie w zażywaniu konkretnych rodzajów narkotyków: MJANDCOKE, ILLICITDRUGUSE, LSYRILLICIT, COKECRACK, OTHERILLICIT
- zażywanie narkotyków w okresie minionego roku: MARJLTYR, MJCOKELY, COCCRKLY
- zażywanie narkotóków od których upłynęło ponad 12 miesięcy: MJGT12MO, COCGT12MO, ANYGT12MO
- alkohol: ALCFMFPB

Jako wskaźnik edukacji:

- IREDUC2, EDU_DUMMY

Jako wskaźnik ekonomiczne:

- INCOME - dochód rodziny
- INCOME_R - dochód własny
- POVERTY - dochód rodziny odniesiony do wskaźnika biedy
- IRPRVHLT - prywatne ubezpieczenie zdrowotne
- WORKFORCE - informacja czy osoba pracuje/pracowała
- EMPSTAT4 - status zatrudnienia

Jako wskaźnik warunków zamieszkania

- REVERSEPOP - charakterystyka miejsca zamieszkania (gęstość zaludnienia)
- MOVESPY2 - liczba przeprowadzek w okresie ostatnich 12 miesięcy
- CACHAR, CATYPE - typ mieszkania

Jako wskaźnik konfliktów z prawem:

- CRIMEHIST, ANYSDRUG, ANYATTACK, ANYTHEFT
- NUMARREST

Stan zdrowia:

- HEALTH2 - stan zdrowia

- SCHDSICK liczba dni opuszczonych w szkole z uwagi choroby (dla uczniów)
- SCHDSKIP liczba dni opuszczony z powodu wagarów
- TXLCAD - informacja o terapii uzależnień od narkotyków lub alkoholu
- DSTNCALM, DSTTIRE, DSTSITST, DSTDEPRS, DSTCHEER, DSTNRVOS
- można również znaleźć informację o myślach samobójczych (YOWRSATP, YOWRSPLN, ADWRDLOT, ADWRSTHK) czy też problemach ze snem (YO_MDEA4, ADWRSLEP, ADWRSMOR) lub zaburzeniach apetytu (YO_MDEA3, ADWRELES)

Inne informacje rodzinne:

- IRMARIT - stan cywilny
- NOMARR2 - liczba razy kiedy osoba wchodziła w związek małżeński
- RKIDSHH - liczba dzieci respondent
- MARRIED aktualny stan cywilny
- CHILDRENINHOME

Przed użyciem danej cechy zapoznaj się z jej specyfikacją oraz informacjami o brakujących wartościach. Podejmij decyzję w jaki sposób traktować wartości brakujące, uzasadnij swój wybór w odniesieniu do najważniejszych zmiennych a w szczególności indeksu depresji.

Faza I - analiza czynnikowa i eksploracja

1. Głównym celem jest analiza czynników wpływających na możliwość rozwoju depresji. Przeanalizuj wpływ czynników związanych z pracą, zarobkami, środowiskiem zamieszkania, rasą, płcią, informacjami rodzinnymi, uzależnieniem od narkotyków i stanem zdrowia psychofizycznego.
2. Na podstawie wstępnej analizy wybierz co najmniej 6 istotnie różne czynniki dla których sprawdzisz stosując znane Ci metody oceny istotności jak te czynniki zmieniają się w zależności od grupy wiekowej i od płci, możesz również sprawdzić czy istnieją istotne zmiany dla typu depresji (somatycznej i niesomatycznej).
3. W analizie wybranych czynników dodaj wnioski oraz opisz obserwacje, tam gdzie to potrzebne możesz zamieścić wykres lub tabelę, pamiętaj jednak, że pod każdym wykresem powinien znajdować się komentarz z obserwacjami

Uwaga

1. Pamiętaj że w przypadku analizy czynników związanych ze stanem psychofizycznym mogą one być silnie powiazane z wskaźnikiem depresji (wyznaczanym na podstawie wielu cech z ankiety), stąd staraj się spojrzeć krytycznie na możliwe relacje między czynnikami które mogą powodować przecieki informacji.
2. Często czynniki oprócz podzbioru dyskretnych lub ciągłych wartości zawierają dodatkowe wartości kodujące informacje o powodzie dla którego dana odpowiedź nie została udzielona, zwróć na to uwagę i podejmij decyzję o sposobie traktowaniu obsługi tych wartości.

Sugestie

1. Do analizy możesz wykorzystać np. korelację, regresję (w tym model ols umożliwiający ocenę istotności zmiennej oraz istnienie powiązań między zmiennymi), możesz również spróbować stworzyć własne cechy. Możesz założyć, że część czynników ma charakter ciągły a część (zdecydowana większość) może być traktowana jako zmienne nominalne lub mieszane,

2. Możesz próbować również, stosujące metody uczenia nienadzorowanego ocenić czy możemy, analizując dane znaleźć pewne grupy (clustry) osób i czy te grupy są zależne np. od kategorii wskaźników depresji.

Faza II - predykcja

W fazie tej należy wykorzystać wnioski a fazy I do budowy modelu regresyjnego OLS umożliwiającego ocenę ryzyka depresji u danej osoby. Dla porównania możesz również spróbować stworzyć model oparty np. o drzewa decyzyjne i porównać jego wyniki z wynikami modelu wykorzystującego OLS. Budowa modelu oceny ryzyka składa się z następujących etapów

1. Stwórz model regresyjny umożliwiający oszacowanie natężenia depresji u danej osoby lub model, który dokonuje klasyfikacji na osoby z depresją i bez - decyzję które podejście wybrać podejmij w oparciu o własne testy i analizy.
2. Dokonaj ocenę modelu, (użyj do tego zbioru walidacyjnego), pamiętaj żeby ten zbiór miał możliwie podobny rozkład populacji
3. Przeanalizuje wyjście z modelu, którym jest prawdopodobieństwo przynależności do danej klasy, w tym celu jeżeli wykorzystałeś model regresyjny możesz zastosować na wyjściu aktywację w postaci funkcji sigmoidalnej,

4. Posortuj otrzymane wyniki po wartości prawdopodobieństwa i podziel je na 5-8 równolicznych grup, dla każdej z tych grup wyznacz ryzyko wystąpienia depresji (bazując na ilorazie liczby osób z depresją i całkowitej liczby osób). Wyświetl wykres zmian ryzyka dla poszczególnych grup
5. Stwórz tablicę przeglądania lub model, który, będzie klasyfikowała daną obserwację do jednej z wyznaczonych grup ryzyka
6. Dla zbioru walidacyjnego oceń jak wygląda liczebność poszczególnych grup ryzyka
7. Na podstawie parametrów modelu regresyjnego przeanalizuj grupę o najniższym i najwyższym poziomie ryzyka depresji, spróbuj ją scharakteryzować.

– Autorzy: *Piotr Kaczmarek*