Metody planowania i analizy eksperymentów
Zadanie domowe nr 1: Analiza opisowa wybranych danych

Kemal Erdem - 183705

26 maja 2020

Streszczenie

Celem zadania była analiza zbioru danych **'Absenteeism at work'** (1). Zbiór zawiera dane dotyczące ilości godzin pracy opuszczonych przez pracowników w konkretnych dniach tygodnia w danym miesiącu. Zbiór składa się z 21 cech i 740 obiektów (przykładów). Cechy możemy podzielić na 9 zm. kategorycznych i 12 zm. numerycznych. Zbiór wykorzystany do analizy został stworzony z użyciem indeksu ustawionego na 5 różnych cechy ('Reason for absence', 'Month of absence', 'Day of the week', 'ID', 'Year') z czego piąta została usunięta ze zbioru przed publikacją. Wiąże się to z koniecznością dodatkowego opisu znajdujących się tam wartości.

Spis treści

1 Opis Danych

Zbiór danych posiada dane dotyczące pracowników firmy kurierskiej pomiędzy 07.2007 a 07.2010. Dane nie są przechowywane w najbardziej zrozumiały dla człowieka sposób ponieważ były wykorzystywane jako dane wejściowe do systemu informatycznego. Każdy wiersz w pliku .csv reprezentuje dzień tygodnia w miesiącu kalendarzowym w danym roku dla danej przyczyny nieobecności w pracy. Oznacza to iż jeżeli w danym miesiącu pracownik był niedostępny z powodu óddawania krwióraz "konsultacji dentystycznych"będziemy posiadać dwa różne rekordy (po jednym dla każdej przyczyny). Jeżeli przyczyna zajmowała więcej niż jeden dzień tygodnia, zostanie stworzony kolejny rekord dla tego dnia tygodnia. Jeżeli dana przyczyna w miesiącu kalendarzowym wystąpiła w ten sam dzień tygodnia dwu lub więcej krotnie (dentysta w kolejne poniedziałki), wartości nieobecności zostają zsumowane dla tych samych dni tygodnia. Poniżej znajdują się przykłady cech wraz z opisami. Nie jest to pełna lista cech, z każdej kategorii zostało wybrane po 4 przykłady.

1.1 Jakościowe - Nominalne

- **'Reason for absence'** - składa się z 28 różnych wartości (cała lista wartości w załączniku).
- **'Disciplinary failure'** - zmienna dychotomiczna gdzie 1 oznacza problem dyscyplinarny
- **'Education'** - zmienna dychotomiczna gdzie 1 oznacza problem dyscyplinarny
- **'Social drinker'** - zmienna dychotomiczna gdzie 1 oznacza osobę pijącą

1.2 Jakościowe - Porządkowe

- **'Month of absence'** - od 1 (styczeń) do 12 (grudzień)
- **'Day of the week'** - od 1 (poniedziałek) do 5 (piątek)
- **'Seasons'** - od 1 (wiosna) do 4 (zima)
- **'Education'** - od 1 (liceum) do 4 (doktor)

1.3 Ilościowe - Dyskretne

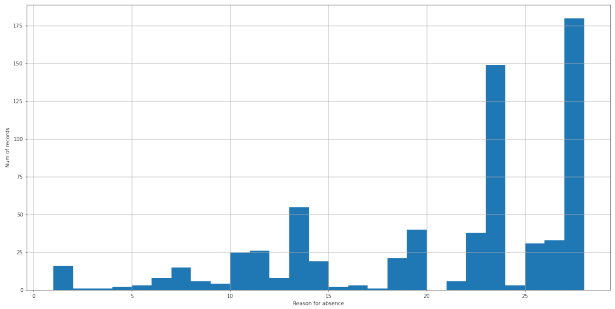
- **'Pet'** - ilość posiadanych zwierząt
- **'Son'** - ilość posiadanych dzieci
- **'Age'** - wiek (w latach)
- **'Service time'** - staż pracy (w latach)

1.4 Ilościowe - Ciągłe

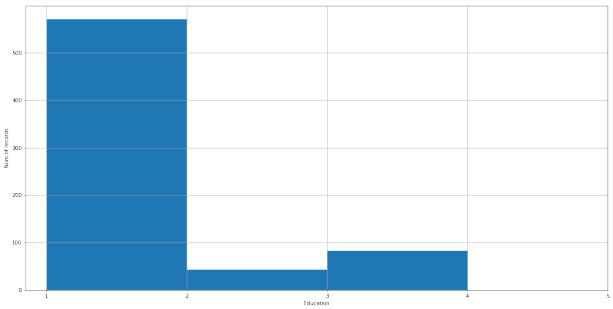
- **'Work load Average'** - średnia ilość godzin pracy
- **'Distance from Residence to Work'** - odległość od miejsca pracy
- **'Body mass index'** - wartość BMI
- **'Absenteeism time in hours'** - ilość godzin nieobecności

2 Analiza Danych

Do analizy danych zostały wykorzystane następujące narzędzia: **python**, **pandas**, **matplolib**, **seaborn**. Część cech użytych do analizy musiała zostać poddana przetworzeniu ponieważ była zapisana w formacie USA (tydzień zaczyna się od niedzieli). Wszystkie operacje dokonane na danych są dostępne w dołączonym kodzie źródłowym.
Wykorzystując analizę graficzną można znaleźć dominujące wartości wśród niektórych cech kategorycznych jak np. **'Reason for absence'** (wartość dominująca to "physiotherapy", ang. fizykoterapia) czy **'Education'** (wartość dominująca to "wykształcenie średnie")



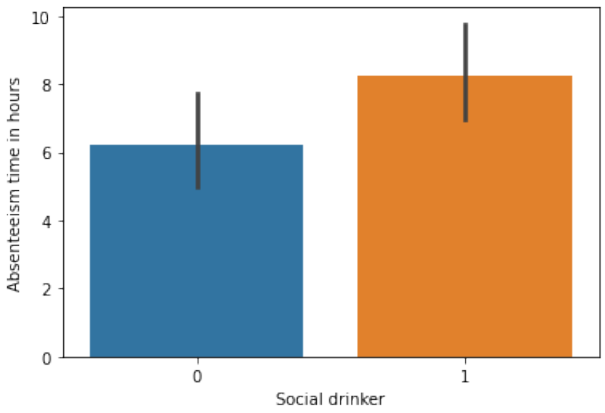
Rysunek 1: Reason of absence



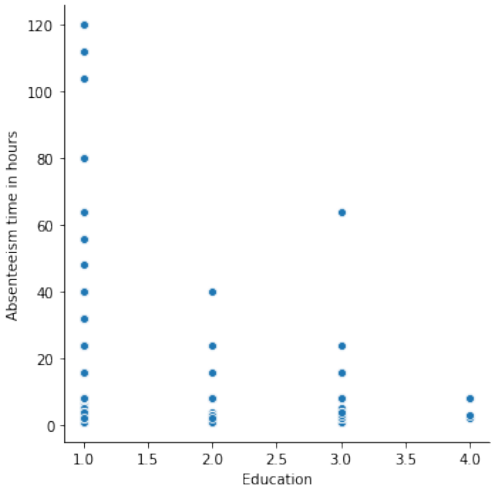
Rysunek 2: Education

2.1 Korelacje

Analiza korelacji danych znalazła dość oczywiste korelacje pomiędzy danymi (Waga -> BMI), nie była jednak w stanie znaleźć znaczącej korelacji z czasem nieobecności (największa dodatnia wartość to 0.147, największa ujemna korelacja to 0.131). Wykresy korelacji znajdują się w załączonym pliku (odpowiednio "Data Correlations"i Ćorrelation with Ąbsenteeism time in hours").Mimo niskich wartości można wyciągnąć wnioski z korelacji pomiędzy piciem alkoholu a średnią ilością opuszczonych godzin (Rys. 3) oraz Edukacją a ilością opuszczanych godzin (Rys. 4). Osobom ze skłonnością do alkoholu częściej zdarza się być nieobecnym w pracy (chociaż różnica jest niewielka). Większą zależność widać pomiędzy edukacją a nieobecnością, im wyższa edukacja tym średnia rozkład ilości opuszczonych godzin maleje.



Rysunek 3: Alkohol a niedostępność



Rysunek 4: Edukacja a niedostępność

Literatura

[1] Ferreira R. P. Sassi R. J. Martiniano, A. Absenteeism at work data set, 2018. URL: http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work.