Praca Domowa Analiza Wielowymiarowa 2022/2023

Maciej Nasiński, Pawel Strawiński oraz Bartosz Baranowski

11/28/22

Opracowanie powinno mieć formę raportu. Jesteś zobowiązana/zobowiązany dostarczyć raport w formie drukowanej (wydruk dwustronny) oraz elektronicznej (akceptowane formaty .pdf, oraz kody w pliku tekstowym). Raport końcowy powinien zawierać reprodukowalne kody wykorzystane do obliczania wyników.

Nieprzekraczalny termin dostarczenia raportu to 14 stycznia 2023 godzina 18.00. Prace należy przesłać na adres elektroniczny pstrawinski@wne.uw.edu.pl oraz mnasinski@wne.uw.edu.pl, a wersję drukowaną pozostawić na portierni budynku 00-241 Warszawa, Długa 44/50 wejście od ulicy Długiej. Opracowanie należy opatrzyć imieniem, nazwiskiem i numerem indeksu autora. Opracowania anonimowe nie będą brane pod uwagę.

Wskazówka: Oceniając wartość merytoryczną opracowania będzie przede wszystkim brana pod uwagę argumentacja uzasadniającą. Ważne decyzje oraz wyniki powinny zostać zilustrowane odpowiednimi wykresami i/lub wartościami stosownych statystyk. Nie ma błędnych odpowiedzi są mniej lub bardziej trafne. Limit długości tekstu: 15000 znaków

Faza odkrycia zamiennika dla Viagry - Współpraca z Novartis

Na bazie prawdziwej historii, gdzie brak pewnych informacji w patencie spowodował opracowanie podobnego leku przez konkurencję

Pracujesz jako cheminformatyk w dziale Innowacji dla jednej z dużych firm farmaceutycznych ze Stanów Zjednoczonych. Konkurencyjna firma właśnie wydała oświadczenie, że znalazła i dostała pozwolenie na zarejestrowanie przełomowego leku na problemy natury męskiej, Viagrę.

Twój szef przyszedł do Ciebie z pytaniem, czy na podstawie struktury opatentowanej cząsteczki, jesteście w stanie znaleźć podobną cząsteczkę, która znacznie przyspieszyłaby czas wprowadzeniu waszego leku na rynek. Jak na pewno dobrze wiesz, przeprowadzenie

wszystkich badań i prac koniecznych do wprowadzenia nowego leku na rynek zajmuje ponad 12 lat i kosztuje średnio ponad 1 mld euro.

Mając dostęp do bazy danych Twojej firmy (molecules.csv), możesz znaleźć informacje dotyczące cząsteczek w Twojej firmie. Dane zostały pozyskane z bazy PubChem na mocy licencji CC-BY-NC 4.0.

Każda cząsteczka posiada jej unikalny identyfikator (*cpd_id*), strukturę zapisaną w formacie tekstowym (*smiles*) oraz wektor bitowy opisujący jej strukturę w numeryczny sposób (*512 wymiarów*). Po załadowaniu danych ustaw ziarno generatora liczb losowych (*ang. seed*) na wartość odpowiadającą numerowi Twojego albumu.

Twoim celem jest odnalezienie podobnej cząsteczki wykorzystując algorytmy analizy skupień (poznane na wykładzie - **uwaga, nie wszystkie algorytmy muszą zadziałać!!!**). Zbadaj charakterystyki zmiennych oraz dokladnie opisz proces odnaleznienia zamiennika.

Dla ułatwienia, Viagra (lub Sildenafil) jest pierwszą cząsteczką w bazie danych (index 0). Po procedurze klastrowania otrzymasz parę cząsteczek (**w zależności od liczby klastrów!**). Cząsteczki możesz porównać używając strony Pubchem, czyli aplikacji webowej do przeglądania cząsteczek chemicznych dostępnych w otwartym dostępie:

`https://pubchem.ncbi.nlm.nih.gov/compound/<tutaj podaj cpd_id>#section=2D-Structure.`

Przykład dla Viagry: https://pubchem.ncbi.nlm.nih.gov/compound/135398744#section=2D-Structure

Posiadając obliczone skupienia, znajdź molekułę, która wygląda *prawie* tak samo i znajdź strukturalną różnicę pomiędzy nimi (wykres na pubchem).

Faza Badan Klinicznych - Inspirowane rozmowami rekrutacyjnymi w FAAMG

Lek o podobnym dzialaniu do Viagry (zamiennik Viagry) został już wynaleziony i wstępnie przebadany. Twoim celem jest omówienie badania oraz wyników, odpowiedz z argumentacją na zadane pytania. Zorganizowano badanie kliniczne zamiennika viagry z dwoma grupami po 100 losowo wybranymi uczestnikami w każdej. Uczestnicy nie wiedzieli czy otrzymali prawdziwy lek. Pierwsza grupa otrzymała placebo i 10 + EPS uczestników było zadowolonych z efektu. W drugiej grupie 50 uczestników było zadowolonych z efektów, to ta grupa otrzymała lek. EPS jest równy nr_albumu moda 40, np. 338914 moda 40 == 34 oraz 351300 moda 40 == 20. W analizie zakładamy błąd pierwszego rodzaju na poziomie 5%, ale możesz spróbować uargumentować i użyć inny. Przy pytaniach natury otwartej liczy się też wasza kreatywność - wskazówka odpowiedzi na pytania otwarte powinny być w miarę zwięzłe (kilkaset znaków).

Pytania:

1. Badanie już sie odbyło. Proszę wyjaśnij w jaki sposób oszacować minimalną niezbędną liczebność próby. Jakie czynniki, ograniczenia i założenia są istotne w tym procesie?

2. Co daje nam losowa próba i czy uchroniła nas przed wszystkimi problemami/ryzykami, proszę dokładnie wyjaśnić? Warto zwrocić uwagę, że w tym przypadku mamy grupę kontrolną.
3. Proszę policzyc przedziały ufności proporcji dla każdej z grup. Co wpływa na wielkość przedziałów i w jakim kierunku? Proszę zwizualizować przedziały.
4. Czy zamienisz wyniki badania z postaci otrzymanych proporcji/liczebności zadowolenia, na wektor binarny z wynikami dla wszystkich uczestników (0101110…)? Opowiedz tak lub nie i uargumentuj.
5. Proszę policzyć czy zadowolenie (proporcja osób zadowolonych w każdej grupie) na przestrzeni grup jest statystycznie istotnie różne, oraz czy ilość osób zadowolonych w grupie z zamiennikiem Viagry jest statystycznie istotnie większa niż w grupie placebo.
6. Jaka jest wielkość efektu (Cohen h dla proporcji - arcsin), jak go oceniasz? Jaką wartość decyzyjną ma ta wielkość?
7. Wytłumacz czym jest moc testu. Jaka jest moc testu dla różnicy proporcji i hipotezy, że proporcja w grupie zamiennika Viagry jest większa. Jak sie zmieni moc testu gdyby w grupie placebo byłoby tylko 50 uczestników (wykorzystaj ogólnodostępne kalkulatory aby zwalidować wyniki).
8. Znając otrzymane wyniki czy rekomendujesz powtórzenie badania?
9. Gdyby w każdej grupie było by po 1000 uczestników, jakie są plusy oraz minusy większej próby? Czy zmiania coś fakt, że w przypadku zwiększenia próby każda grupa musiała by być zbadana z miesięcznym interwałem?

W badaniu zebrano wiele dodatkowych informacji takich jak wiek uczestników. Na podstawie tabeli z częstościami zbadaj wpływ wieku na zadowolenie, niezależność.

	18-30 lat	30-50 lat	50-70 lat
Niezadowolony	EPS	10	40 - EPS
Zadowolony	25	15	10

10. Przyjrzyj sie tabeli jak myslisz czy wiek może być powiązany z wynikami badania zadowolenia, uargumentuj?
11. Podeprzyj swoje wnioski testem statystycznym i przeanalizuj wyniki. Czy twoje założenia się potwierdziły?
12. Załóżmy że w badaniu brała udział dwukrotnie większa liczba uczestników i proporcje w tabeli zostały zachowane, jak zmienią sie wyniki testu z poprzedniego pytania.
13. Czy z obecną wiedzą rekomendujesz próbę dopuszczenia leku na rynek i masz jakieś przeciwskazania?