Laboratorium 1. Elementy Excela

1. Wczytaj zawarte w pliku credit_g.xls. Plik ten zawiera informacje o danych będących historią dot. osób zaciągających kredyty w jednym z niemieckich banków. Uwaga. Ponieważ są to dane historyczne więc jednostką jest DM – marka niemiecka
2. Opis poszczególnych kolumn zawarty jest w pliku Credit_g_Opis_Zmiennych.txt
3. Skopiuj kolumny „age", „*credit_amount*" oraz „*duration*". Pierwsza ze zmiennych opisuje wielkość kredytu, a druga jego czas trwania.
4. Wyznacz wartości średniej i mediany dla tych dwóch kolumn. Co na tej podstawie możemy powiedzieć. Czy rozkład wartości jest skośny?,Jeśli tak to w którą stronę? Co to oznacza?
5. Wyznacz wartośc kwartyli 0.25, 0.5 0.75 Co z nich wynika?
6. Policz wartość rozstępu międzykwartylowego i odchylenia standardowego.
7. Wyznacz wartości dla histogramu dla przedziałów liczonych co 10 dla zmiennych *age* oraz *duration*. Zastanów się jak wyznaczyć wartości histogramu korzystając z funkcji suma.jeżeli(zakres;warunek;suma_zakres)
8. Narysuj histogramy
9. Zdefiniuj przedziały histogramu o wielkości 6 i stwórz histogram dla zmiennej *duration* za pomocą funkcji *częstość* (uwaga jest to funkcja blokowa wywoływana za pomocą kombinacji ctrl+shift+enter)
10. Porównaj obydwa histogramy, co można zaobserwować na obydwu wykresach. Skąd wynikają różnice?
11. Otwórz zakładkę *credit-g* i stwórz tabelę przestawną. Wstaw-> Tabela przestawna
12. Jako etykiety wierszy ustaw *employment*, a jako etykiety kolumn *purpose* (cel kredytu), a jako wartości średnią wartość z *credit_amount*.
13. Przeanalizuj wygenerowaną tabelę i znajdź różne zastanawiające wyniki. Przykładowo zaobserwuj, że dla osób kupujących nowe auta (kolumna *new car*) największy kredyt zaciągają osoby niezatrudnione (*unemploymed*) oraz osoby o stażu zatrudnienia powyżej 7 lat. Innym przykładem są wydatki na edukację. Najwyższe dla osób o okresie zatrudnienia od 1 do 4 lat. Oraz od 4 do 7. Zwróć uwagę, że osoby niezatrudnione nie wydają pieniędzy na edukację!!!
14. Dodaj do opisu wierszy lub kolumn w tablicy przestawnej pozycję *housing* która opisuje czy dana osoba posiada/nie posiada własny dom/mieszkanie. Przeanalizuj uzyskane wyniki i znajdź (wypisz) te które uważasz za ciekawe. Spróbuj nadać im interpretację.
15. Spróbuj dokonać analizy ręcznie, oceniając które czynniki wpływają na to czy kotś będzie dobrym lub złym kredytobiorcą. W tym celu wykorzystaj tabelę przestawną ustawiając jako kolumny *class* a jako wiersze *różne inne atrybuty*. Np. *employment* Które czynniki i dla jakich wartości wpływają na większe ryzyko udzielania kredytu
16. Uruchom oprogramowanie RapidMiner i spróbuj zbudować drzewo decyzji obrazujące jakie czynniki komputer wyznaczy jako mające największy wpływ na udzielenie kredytu. W tym wykorzystaj operator *Read Excel* oraz wczytaj zbiór danych credit_g.xls, następnie oznacz kolumnę class jako etykietę wierszy (operator *Set Role* gdzie kolumnę *class* ustaw jako *label*) Na wyjście podłącz drzewo decyzji (spróbuj z domyślnym drzewem oraz drzewem W-J48 dostępnym w Classification and Regression->Weka->Trees)
17. Przeanalizuj uzyskane wyniki z punktu widzenia tabeli przestawnej.

18. Spróbuj zbudować prosty model predykcyjny w oparciu o drzewo decyzji i oceń jego dokładność za pomocą operatora XValidation.