***Кисляк Марини*. Прикладна та теоретична статистика, 2 курс маг.**

**Лабораторна робота №1. *Варіант 6.***

1. Для аналізу даних перш за все потрібно було їх перевірити на наявність пропусків та визначити по яких змінних буде проводитися аналіз. Дані містили по 20 номінальних та 20 числових змінних. Для аналізу використовувалися лише числові змінні. Спостереження у яких були пропущені значення було видалено.

Після проведення аналізу за допомогою методів KNN, SVM та Decision.Trees було отримано такі результати:

Confusion Matrix KNN:

[[13 14]

[14 32]]

Accuracy KNN: 0.6164383561643836

Accuracy SWM: 0.6438356164383562

Confusion Matrix SWM:

[[ 9 8]

[18 38]]

Accuracy DT: 0.6986301369863014

Confusion Matrix DT:

[[16 11]

[11 35]]

2. для наступного експерименту було додано ще одну факторну змінну. Кожному фактору було присвоєно число і після цього проведено повторну класифікацію.

#added factor data

Confusion Matrix KNN:

[[16 11]

[17 29]]

Accuracy KNN: 0.6164383561643836

Accuracy SWM: 0.684931506849315

Confusion Matrix SWM:

[[13 9]

[14 37]]

Accuracy DT: 0.5753424657534246

Confusion Matrix DT:

[[13 17]

[14 29]]

3. Висновок. Отже, на першому наборі змінних класифікатори СВМ та КНН спрацювали приблизно однаково, правильно класифіковано було майже 64% тестових даних. Найкраще спрацював метод ДТ- правильно класифіковано майже 70% даних. На другому наборі, при додаванні факторної змінної, найкраще спрацював метод СВМ. Досить невисоку якість класифікації у всіх прийомах можна пояснити невеликою кількістю даних. Можливо потрібно не видаляти рядки з пропущеними значеннями, а замінити пропущені значення на середні. Після цього якість класифікації може покращитися.