Міністерство освіти і науки України
Національний університет «Львівська політехніка»
Інститут комп’ютерних наук та інформаційних технологій
Кафедра «Системи штучного інтелекту»

Лабораторна Робота №2

З предмету: «Видобування великих даних»

Виконав
студент групи КН-311
Ткачук Орест
Прийняла :
Вовк О. Б.

Львів-2021

Лабораторна робота №2

Завдання: застосувати один або більше метод штучного інтелекту для аналізу даних та пояснити отриманий результат.

Хід роботи

В лабораторній роботі №2 я спробую знайти залежність між кількістю соц. працівників та кількістю самогубств в країнах, а також знайти оптимальну кількість працівників для мінімізації рівня самогубств. Також спробую перевірити гіпотезу про те, що суїциди чоловіків відбуваються частіше, ніж суїциди жінок.

Для початку необхідно нормалізувати дані. Також введемо дві додаткових колонки які будуть відображати загальну кількість суїцидів та соц. працівників по країнах.

```
normalized = feature_sel_merge

normalized['all_age'] = normalized[[' 80_above', ' 70to79', ' 60to69 ', ' 50to59 ', ' 40to49', ' 30to39', ' 20to29']].sum(axis = 1, skipna = True)

normalized['all_spec'] = normalized[['Psychiatrists', 'Nurses', 'Social_workers', 'Psychologists']].sum(axis = 1, skipna = True)

columns = ['Psychiatrists', 'Nurses', 'Social_workers', 'Psychologists'] + [' 80_above', ' 70to79', ' 60to69 ', ' 50to59 ', ' 40to49', ' 30to39', ' 20to29', ' 10to19']

for i in range(len(columns)):

    mean = np.mean(normalized[(columns)[i]])

    std = np.std(normalized[(columns)[i]], ddof=1)

    normalized[(columns)[i]] = (normalized[(columns)[i]]-mean)/std

normalized
```

Sex	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19	Psychiatrists	Nurses	Social_workers	Psychologists	all_age	all_spec
Both sexes	-0.019805	-0.601763	-0.840125	-0.681184	-0.466502	-0.146265	-0.006493	-0.303706	-0.529141	-0.559881	-0.361339	-0.339544	-0.398634	-0.509605
Male	0.610997	-0.208432	-0.538767	-0.397842	-0.124262	0.449013	0.676209	0.204131	-0.529141	-0.559881	-0.361339	-0.339544	0.168184	-0.509605
Female	-0.506234	-0.947418	-1.127468	-0.987500	-0.843844	-0.802079	-0.756345	-0.871289	-0.529141	-0.559881	-0.361339	-0.339544	-0.910775	-0.509605
Both sexes	-0.590637	-0.709036	-0.805084	-0.512711	-0.247117	-0.459038	-0.420590	0.263876	-0.355645	-0.259497	-0.313759	-0.306752	-0.672019	-0.360171
Male	-0.437378	-0.566006	-0.657908	-0.237026	0.139000	-0.186623	-0.442974	-0.303706	-0.355645	-0.259497	-0.313759	-0.306752	-0.461512	-0.360171
Female	-0.710578	-0.844119	-0.952259	-0.773079	-0.606909	-0.731453	-0.409399	0.861332	-0.355645	-0.259497	-0.313759	-0.306752	-0.863388	-0.360171
Both sexes	0.457739	0.633853	0.442402	0.023342	-0.431401	-0.529664	-0.409399	-0.453070	-0.553486	-0.534974	-0.360352	-0.343648	0.267514	-0.505232
Male	1.783756	1.670818	1.206311	0.559394	-0.106711	-0.226980	0.015891	-0.094597	-0.553486	-0.534974	-0.360352	-0.343648	1.308199	-0.505232
Female	-0.390735	-0.176648	-0.209374	-0.451447	-0.738540	-0.822258	-0.823496	-0.811544	-0.553486	-0.534974	-0.360352	-0.343648	-0.499786	-0.505232

Рис.1 – нормалізований датасет

Далі для цих даних використаємо кілька методів аналізу даних. Першим методом буде зменшення розмірності датасету. Для цього використаємо алгоритм PCA з 90% відхиленням.

```
normalized_all = normalized[normalized['Sex'] == ' Both sexes'].drop(['all_age', 'all_spec'], axis = 1)

all_columns = suicide_columns + spec_columns

pca = PCA(0.95).fit(normalized_all[(all_columns)])

components = pca.transform(normalized_all[(all_columns)])

filtered = pca.inverse_transform(components)

restored_all = copy(normalized_all)

print(pca.components_)

print(pca.singular_values_)

print('n_components: ', pca.n_components_)

for i in range(len(all_columns)):

    restored_all[(all_columns)[i]] = filtered.T[i]
```

Наші компоненти ти власні значення будуть мати наступний вигляд

```
components:  [[ 0.04146243  0.08769341  0.1749342   0.29003458  0.32521057  0.32296493
   0.30974689  0.31108227  0.39964533  0.31548873  0.36890747  0.27882408]
 [-0.33219024 -0.38984652 -0.36481918 -0.2891628  -0.22280697 -0.19641386
  -0.14543139 -0.13155269  0.27229887  0.22640913  0.36408576  0.36921161]
 [-0.42937221 -0.41698533 -0.29918533 -0.05806393  0.15281419  0.24944052
   0.3307915   0.42476831 -0.16055884 -0.00506238 -0.26697622 -0.27635937]
 [ 0.02030626  0.01453418 -0.02596081 -0.05207036 -0.02522985  0.01857901
   0.12293007  0.27957018 -0.31530416 -0.69148595  0.26763533  0.50253349]
 [ 0.24744212  0.18644117 -0.0261069  -0.2628448  -0.27804415 -0.11332522
   0.15524939  0.35234652 -0.48562823  0.56195025 -0.04584662  0.20522366]]
values:  [18.65677731 16.79702617 11.80656572  9.39564926  5.98441012]
n_components:  5
```

Рис.2 – головні компоненти та власні значення алгоритму PCA

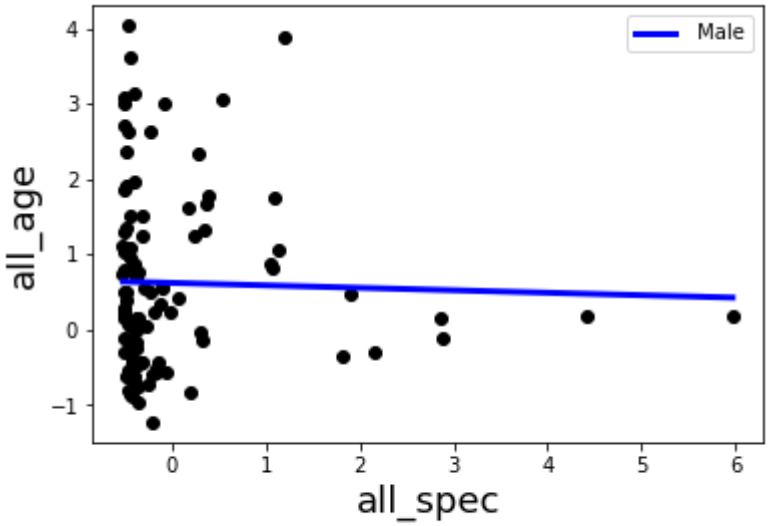
	Country	Sex	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19	Psychiatrists	Nurses	Social_workers	Psychologists
0	Afghanistan	Both sexes	-0.019805	-0.601763	-0.840125	-0.681184	-0.466502	-0.146265	-0.006493	-0.303706	-0.529141	-0.559881	-0.361339	-0.339544
3	Albania	Both sexes	-0.590637	-0.709036	-0.805084	-0.512711	-0.247117	-0.459038	-0.420590	0.263876	-0.355645	-0.259497	-0.313759	-0.306752
6	Angola	Both sexes	0.457739	0.633853	0.442402	0.023342	-0.431401	-0.529664	-0.409399	-0.453070	-0.553486	-0.534974	-0.360352	-0.343648
9	Antigua and Barbuda	Both sexes	-0.952682	-1.038798	-0.657908	-1.110026	-1.045678	-1.074494	-1.148059	-1.229762	-0.421406	-0.253780	-0.181658	-0.059124
12	Argentina	Both sexes	-0.601742	-0.577925	-0.524750	-0.382526	-0.185690	0.015167	0.564291	1.339296	2.475407	1.286426	6.306907	7.456035
15	Armenia	Both sexes	-0.332985	-0.347489	-0.517742	-0.466763	-0.255893	-0.529664	-0.644427	-0.572561	-0.024185	-0.065874	-0.349040	-0.322289
18	Azerbaijan	Both sexes	-0.781654	-0.776577	-0.910209	-0.796053	-0.738540	-0.751632	-0.901839	-0.811544	-0.078472	-0.266543	-0.356222	-0.309067

Рис.3 – оригінальні дані до PCA

	Country	Sex	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19	Psychiatrists	Nurses	Social_workers	Psychologists
0	Afghanistan	Both sexes	-0.450509	-0.499148	-0.543266	-0.539128	-0.435007	-0.350629	-0.232279	-0.137032	-0.605754	-0.496816	-0.429371	-0.248469
3	Albania	Both sexes	-0.669211	-0.720381	-0.695741	-0.568400	-0.380930	-0.270132	-0.134867	-0.035262	-0.418956	-0.248113	-0.346778	-0.227024
6	Angola	Both sexes	0.576679	0.586642	0.359367	-0.051252	-0.321336	-0.400937	-0.458108	-0.522353	-0.574000	-0.521801	-0.418218	-0.271808
9	Antigua and Barbuda	Both sexes	-0.743222	-0.908907	-1.009973	-1.104190	-1.094800	-1.102318	-1.076304	-1.153786	-0.240094	-0.326490	-0.222645	-0.114645
12	Argentina	Both sexes	-0.370901	-0.570822	-0.630236	-0.508513	-0.352218	-0.069779	0.734087	1.473042	2.700296	1.172791	6.415886	7.233538
15	Armenia	Both sexes	-0.363718	-0.405657	-0.411502	-0.424829	-0.431425	-0.469069	-0.528825	-0.641659	-0.042725	-0.071860	-0.302039	-0.355323
18	Azerbaijan	Both sexes	-0.749892	-0.854229	-0.851541	-0.809856	-0.746445	-0.764127	-0.791629	-0.891800	-0.111061	-0.264205	-0.319498	-0.325539

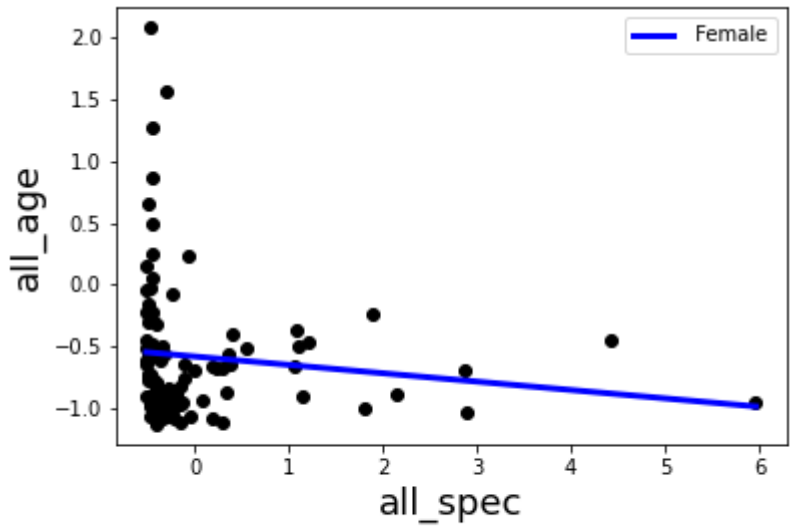
Рис.4 – дані відновлені за допомогою PCA

Наступний метод аналізу даних – передбачення за допомогою лінійної регресії. Спробуємо побудувати лінії лінійної регресії для чоловіків, жінок та загальну.



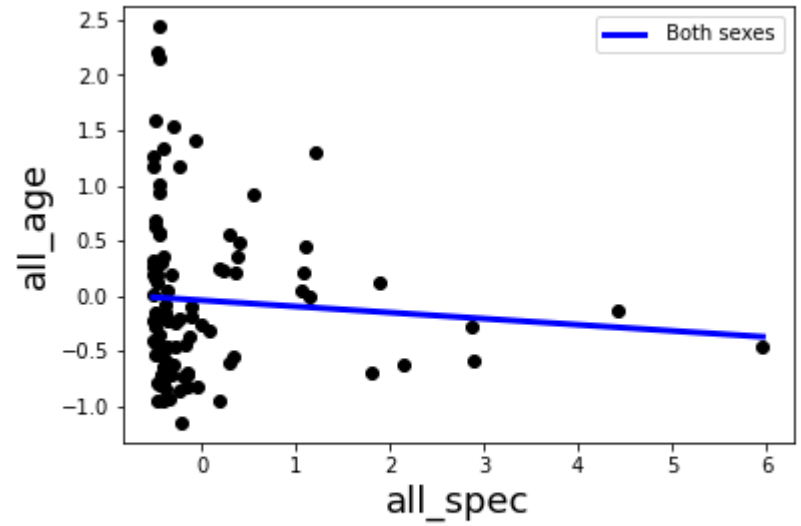

```
MAE: 0.9419638329989664
r2: 0.000760566095744819
```

Рис.5 – графік лінійної регресії для чоловіків




```
MAE: 0.3690486896843543
r2: 0.015237500725599706
```

Рис.6 – графік лінійної регресії для жінок




```
MAE: 0.5823019991636419
r2: 0.005510781045115931
```

Рис.7– графік загальної лінійної регресії

З результатів можна припустити, що при низькій кількості соц. працівників суїциди перестають лінійно залежати від них. Оскільки значення r2 знаходиться в околі 0, можна зробити висновок, що модель лінійної регресії не відрізняється від взяття середнього значення, відповідно лініа регресія дає некоректний прогноз. Гіпотезу про залежність суїцидів від кількості соц. працівників підтвердити не вдалося.

Оскільки лінійна регресія не змогла дати коректний прогноз, спробуємо кластеризувати дані за допомогою алгоритму K-means для перевірки гіпотези, що суїциди жінок відбуваються не так часто, як суїциди чоловіків. Кластеризувати будемо по загальній кількості суїцидів та соц. працівників.

```
y_pred = KMeans(n_clusters=2, random_state=47).fit_predict(X)

plt.scatter(X[:, 0], X[:, 1], c=y_pred)

plt.xlabel('suicide_rate')

plt.ylabel('spec_rate')

plt.title('Male/Female suicide clustering')

plt.show()
```

```
plt.scatter(X[:, 0], X[:, 1], c = data[:,0])

plt.xlabel('suicide_rate')

plt.ylabel('spec_rate')

plt.title('Male/Female')

plt.show()
```

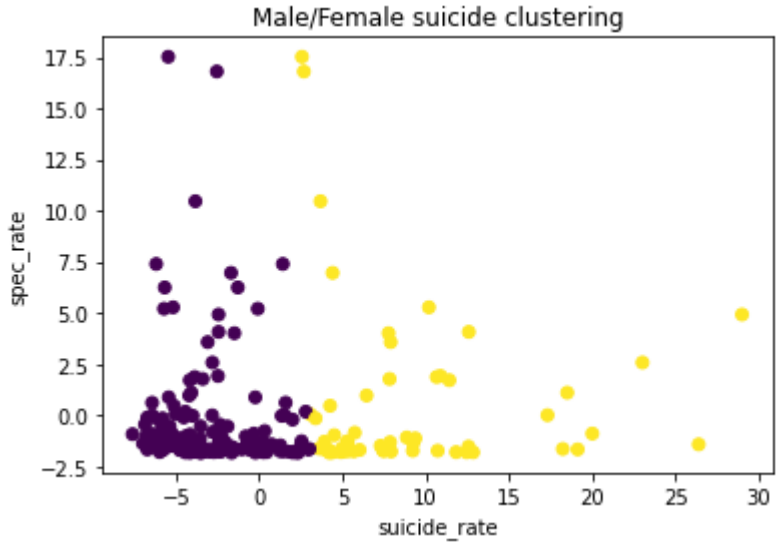


Рис.8 – кластеризація K-means

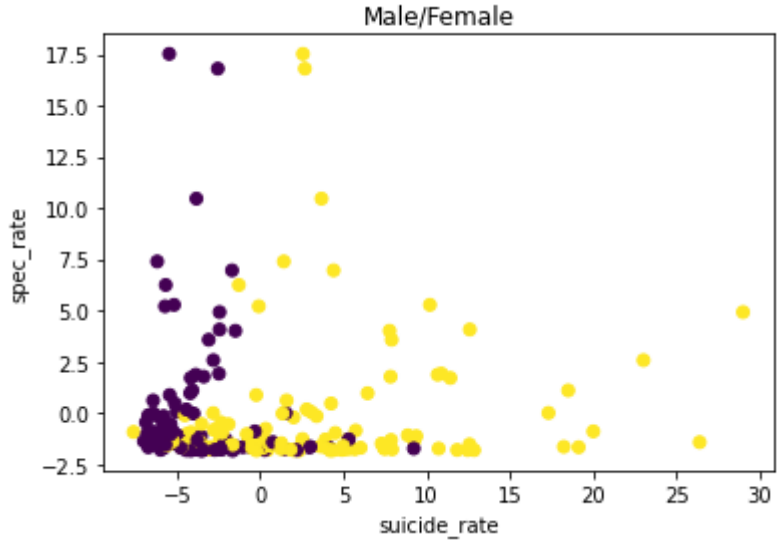


Рис.9 – поділ за статтю

Кластеризація K-means за загальною кількості суїцидів та соц. працівників дала схожий поділ, що і поділ за колонкою статі з початкового набору даних.

Далі застосуємо до цього датасету класифікацію. Спочатку розділимо дані на тренувальні та тестові.

```
svr_lin = SVR(kernel='rbf', C=1.0, gamma='auto')

data = normalized[normalized['Sex'] != ' Both sexes'][['Sex', 'all_age', 'all_spec']]

data['Sex'].replace([" Male", " Female"], [1,0], inplace=True)

data = data.to_numpy()

X = data[:, 1:3]

y = data[:,0]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=47)
```

Для класифікації використаємо алгоритм SVM.

```
C = 1.0

models = (svm.SVC(kernel='linear', C=C),

          svm.SVC(kernel='poly', degree=2, gamma='auto', C=C),

          svm.SVC(kernel='poly', degree=3, gamma='auto', C=C),

          svm.SVC(kernel='sigmoid'),

          svm.SVC(kernel='rbf', gamma=0.7, C=C),

          svm.SVC(kernel=my_poly))


titles = ('linear kernel',

          'polynomial kernel 2',

          'polynomial kernel 3',

          'sigmoid kernel',

          'RBF kernel',

          'my kernel')
```

```
fig, sub = plt.subplots(6, 1)

fig.set_size_inches(5, 30)

plt.subplots_adjust(wspace=0.4, hspace=0.4)


X0, X1 = X_train[:, 0], X_train[:, 1]

xx, yy = make_meshgrid(X0, X1)

for clf, title, ax in zip(models, titles, sub.flatten()):

        clf = clf.fit(X_train, y_train)

        ax.set_title(title)

        plot_contours(ax, clf, xx, yy, cmap=plt.cm.PiYG, alpha=0.8)

        print(clf.score(X_test, y_test))

        print(clf.predict(X_test))

        print(y_test)

        ax.scatter(X_test[:, 0], X_test[:, 1], c=y_test, cmap=plt.cm.PiYG, s=20,
edgecolors='k')

        ax.set_xlim(xx.min(), xx.max())

        ax.set_ylim(yy.min(), yy.max())

        ax.set_xlabel('suicide_rate')

        ax.set_ylabel('spec_rate')

        ax.set_xticks(())

        ax.set_yticks(())

        ax.set_title(title)


plt.show()
```

Отримаємо наступні результати:

Точність - 77%

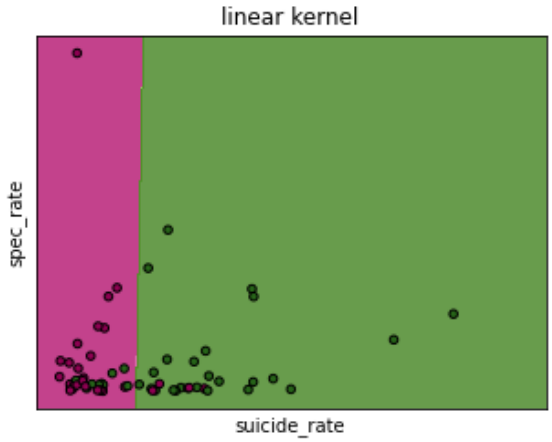


Рис.10 – класифікація з лінійним ядром

Точність - 57%

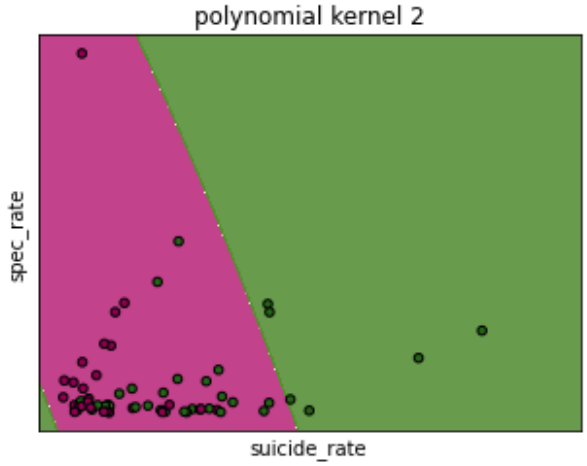


Рис.11 – класифікація з квадратичним ядром

Точність - 80%

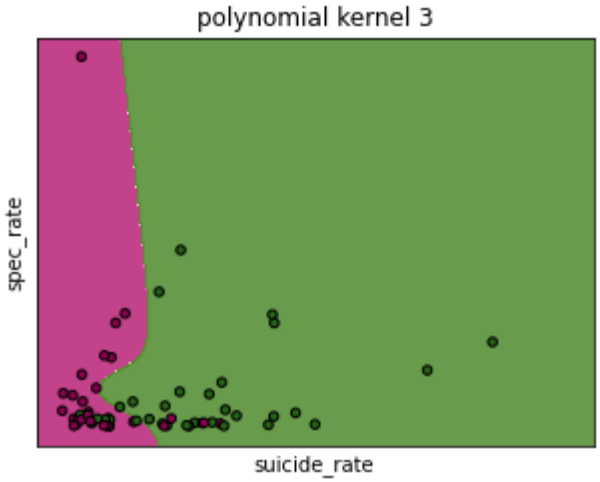


Рис.12 – класифікація з ядром 3 степеню

Точність - 74%

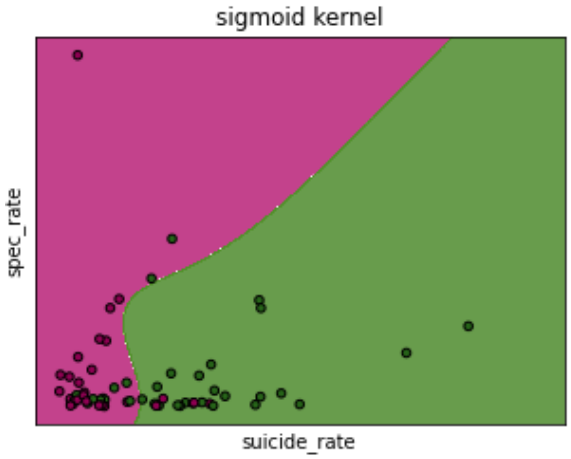


Рис.13 – класифікація з сігмоїдними ядром

Точність - 79%

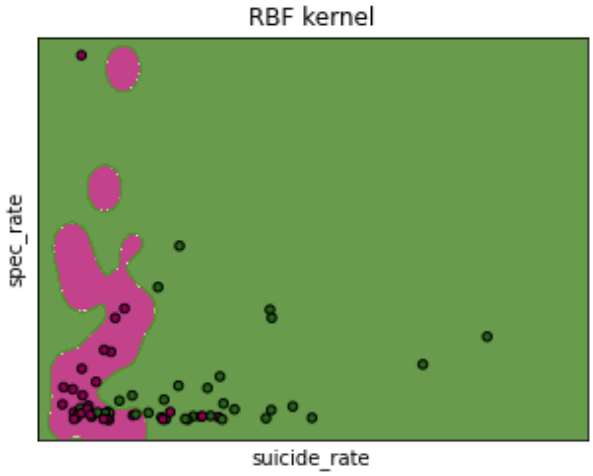


Рис.14 – класифікація з RBF ядром

Точність - 80%

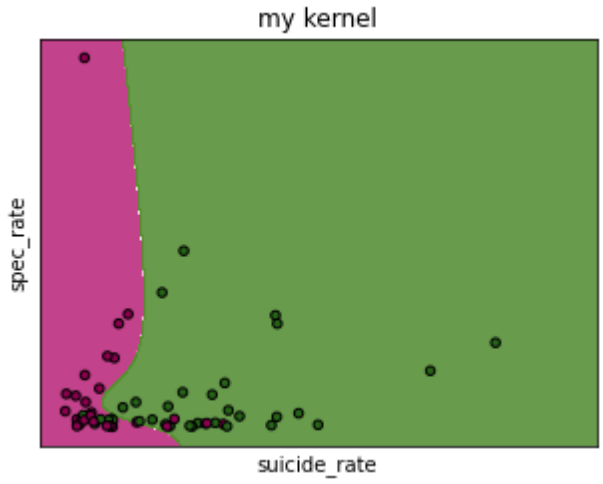


Рис.15 – класифікація з власним ядром

Висновок:

В даній лабораторній роботі я застосував 4 різні методи аналізу даних.

В першому методі я використав зменшення вимірності за допомогою алгоритму PCA а також відновив дані після зменшення вимірності.

В другому методі я спробував перевірити гіпотезу про залежність суїцидів від кількості соц. працівників і якщо вона підтвердиться – то спробувати передбачити оптимальну кількість соц. працівників для запобігання суїцидам. Лінійна регресія не змогла дати коректний прогноз і передбачення було неможливо виконати.

В третьому методі я кластеризував дані за загальною кількістю суїцидів та соц. працівників для перевірки гіпотези про те, що кількість суїцидів чоловіків менша за кількість суїцидів чоловіків. Кластеризація за допомогою K-means поділила на 2 кластери, які збігаються з поділом за колонкою статі з початкового набору даних.

В останньому методі я застосував SVM для класифікації статі групи людей від загальної кількості суїцидів та соц. працівників. Результати показали хорошу точність класифікації.