

Податочно рударство

**Forever alone**

22.06.2023 Орданче Недев

Скопје 203206

# Содржина

Contents

[Содржина 2](#_Toc138355304)

[1.Опис на проектот 3](#_Toc138355305)

[2. Возраст на испитаниците 5](#_Toc138355306)

[3. Резултати по модел 6](#_Toc138355307)

[3.1 Naïve Bayes Classifier 6](#_Toc138355308)

[3.2 XGBoost Classifier 7](#_Toc138355309)

[3.3 Logistic Regression 8](#_Toc138355310)

[3.4 KNN(со 4) 9](#_Toc138355311)

[3.5 KMeans 10](#_Toc138355312)

[Заклучок 11](#_Toc138355313)

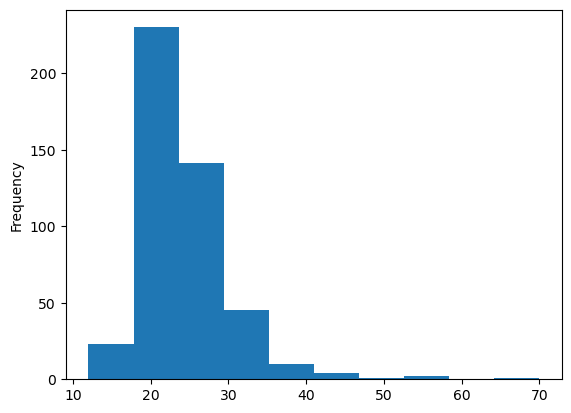
## Опис на проектот

|  |  |
| --- | --- |
| Име, презиме, индекс | Орданче Недев |
| Опис на проектот | Проектот претставува истражување на тема самоубиства по предметот податочно рударство |
| Податочно множество | <https://www.kaggle.com/code/antonaks/suicide-attempt-prediction-foreveralone-dataset/input> |
| Независни променливи на податочното множество | Време, род, сексуалност, работно титула, приход, социјален страв, депресија, дали работеле на самоподобрување, дали барале помош од други за врска, ниво на образование, девственост, дали платиле за секс и др. |
| Колоната што ја предвидуваме | Предвидуваме дали некое лице ќе изврши обид за самоубиство или не. |
| Вредности кои недостасуваат | Има вредности кои недостасуваат во колоната работна титула, но тоа е мал број на такви, па тие редици се отстрануваат. |
| Погрешни вредности, и вредности кои се двојат | Погрешни вредности има во колоната број на пријатели. Броевите се децимални. Со отфрлање на вредности над 15, спојување на 0.2 и 0.5 во 0, спојување на 15 во 13, и претварање на вредностите во целобројни се средува колоната. |
| Модели кои се користат за предвидување | Logistic Regression, Naïve Bayes Classifier, KNN, XGBoost Classifier за предвидување и Kmeans за детекција на ризични групи кај кои би се појавиле самоубиствени идеи. |
| Употребени научни трудови | 1. Број на самоубиства и методи кај различни возрасни групи: Австралиски податоци и перцепции, Џон Сноудон, 1997 2. Предвидувања со Машинско учење на самоубиствени идеи, планирање и обиди помеѓу млади Корејци: Популациско-базирана студија, Џеонгјун Ли, Тијунг Пак, 2022 |

|  |  |
| --- | --- |
| Предикатори | Род, сексуалност, возраст, пријатели, социјален страв, депресија, ниво на образование. |
| Најдобар модел | Naïve Bayes Classifier |
| Мерки за оценка на модел | Accuracy score, jaccard score, precision од classification report и roc curve |
|  |  |

## 2. Возраст на испитаниците

Просечната возраст на испитаниците е 24.03, медијаната е 23 и модата е 23.



## 3. Резултати по модел

### 3.1 Naïve Bayes Classifier

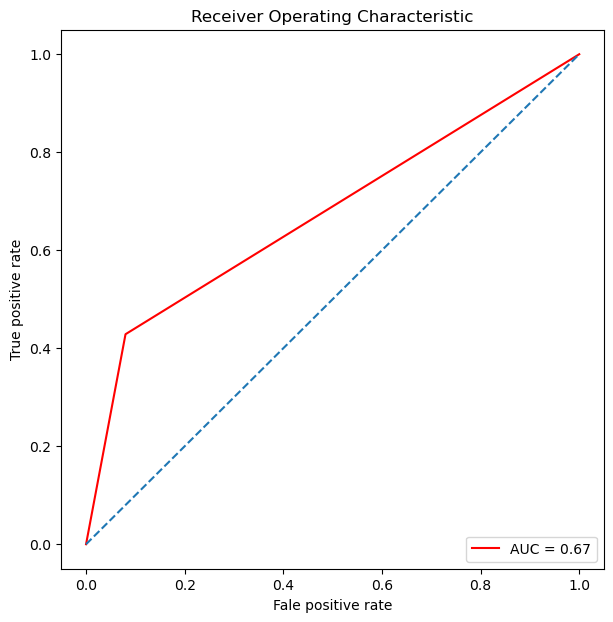
precision(Classification report)

0 0.88

1 0.53

Accuracy score: 83.47

Jaccard score: 31.03



### 3.2 XGBoost Classifier

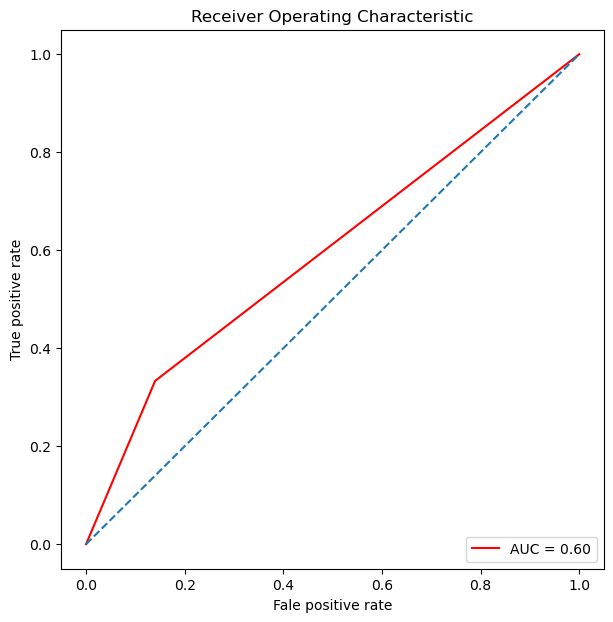
precision (Classification report)

0 0.86

1 0.33

Accuracy score: 76.85

Jaccard score: 20.0



### 3.3 Logistic Regression

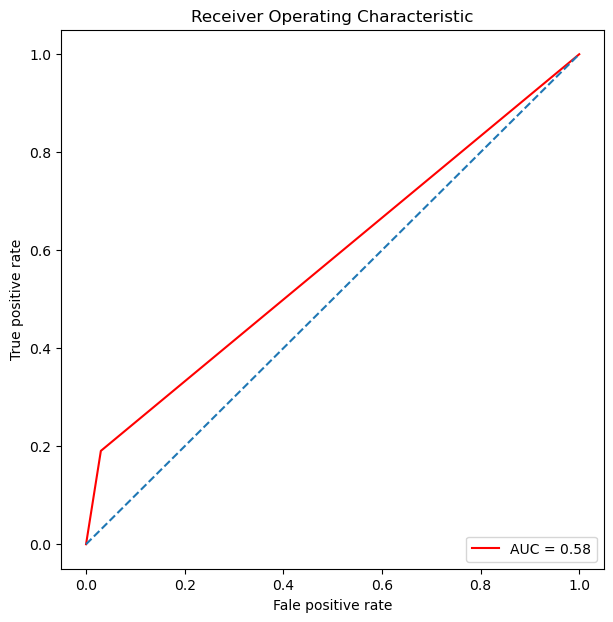
Precision(Classification report)

0 0.85

1 0.57

Accuracy score: 83.47

Jaccard index: 16.67



### 3.4 KNN(со 4)

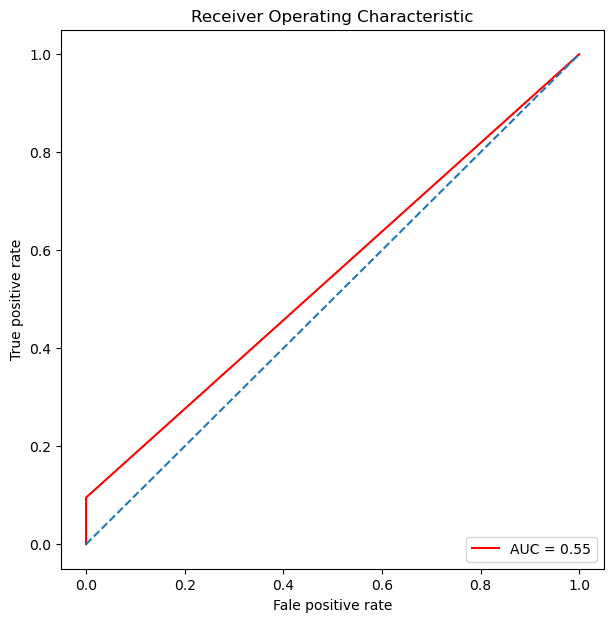
precision

0 0.84

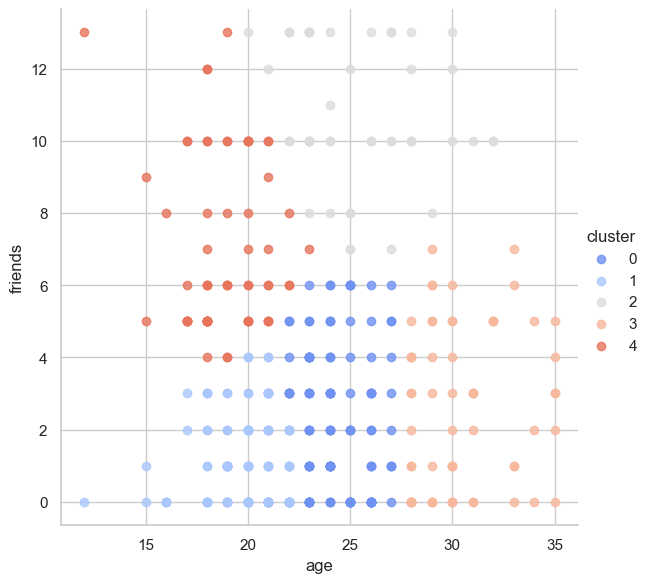
1 1.00

Accuracy score: 84.30

Jaccard score: 9.52



### 3.5 KMeans



## Заклучок

Самоубиството и самоубиствените идеи се работи кои стануваат се поактуелни во државите. Тоа е горчлив проблем на повеќе земји во развој, но и на развиените земји. Само со навремено откривање на лицата склони кон такви идеи и кон самоубиство, може исто да се превенира.

Мое лично мислење е дека алгоритмите на машинско учење можат да помогнат во решавање на ваквите проблеми и подобрување на едно општество.