Универзитет Св. Кирил и Методиј – Скопје  
  
Факултет за информатички науки и компјутерско инженерство

**Предмет: Вовед во науката за податоци  
  
Тема: Избор на податочно множество за Алцхајмерова болест и негова анализа преку соодветна методологија.**

Github линк до проектот : <https://github.com/majkltancin/proekt-vnp.git>

Ментор: Изработил:

Ана Тодоровска Мајкл Танчин

Скопје, 2025

Содржина

[**Вовед** 3](#_Toc200915948)

[**Користени алатки и библиотеки** 4](#_Toc200915949)

[**Опис на избраното податочно множество** 4](#_Toc200915950)

[**Анализа на податоците** 5](#_Toc200915951)

[**Подготовка на податоците** 6](#_Toc200915952)

[**Креирање на модели** 6](#_Toc200915953)

[**Споредба на модели** 7](#_Toc200915954)

[**Заклучок** 8](#_Toc200915955)

[**Користена литератира** 8](#_Toc200915956)

# **Вовед**

Целта на овој проект е да се обработат податоци за алцхаемерова болест и да се проба со различни модели на машинско учење да се проучи податочното множество и да се предвиди дали некој пациент ја има или ја нема таа болест.

Зашто е битно да се создаваат тренираат и користат вакви модели на машинско учење за предвидување на оваа болест поред специјалисти во оваа област би се запрашале некои.

Ова е болест која го напaѓа мозокот кај поголем процент од повозрасната генерација и се работи за болест која за краток временски период од нејзиното добивање доведува до губење на конгнитивните функции, памтење, промена на личноста и способноста за функционирање. Со тоа наjчесто кога пациентите ќе се здобијат со симптоми веќе е доцна односно болеста е развиена до некој степен, затоа овој начин на предвидување на оваа болест е доста корисен односно им помага на специјалистите да ја потврдат или порано да ја увидат оваа дијагноза и да им помогнат на луѓето.

Со тоа овој проект ќе има за цел:

* Да го визуализираме и анализиреме податочното множество
* Да провериме дали има некои вредности кои фалат и да се дополнат доколку има таков случај
* Да се доведат потребните податоци во ист формат пред креирањето на моделите
* Креирање на повеќе модели
* Финални споредба на моделите и увидување на битни карактеристики за истите

# **Користени алатки и библиотеки**

За изработка на проект од овој тип и за споредба на неколкуте модели користени се следните алатки и библиотеки.

* Програмски јазик : Python
* CSV за вчитување на податоците
* Pandas за креирање на дата фрејм предобработка и манипулација со истиот
* Numpy за справување со нумерички вредности
* Seaborn за визуализација на податоците

# **Опис на избраното податочно множество**

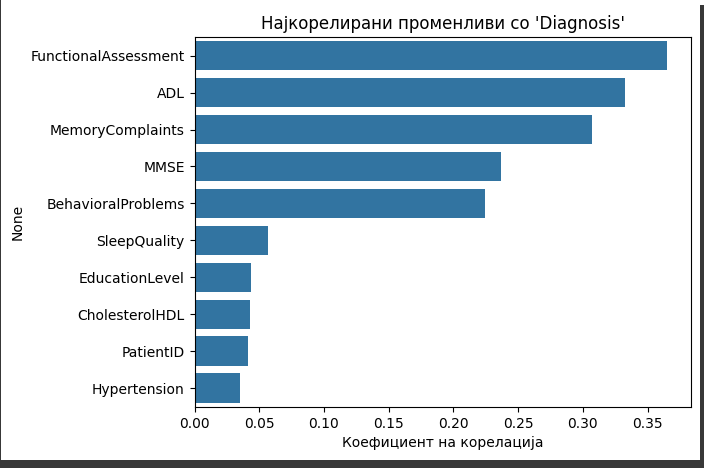
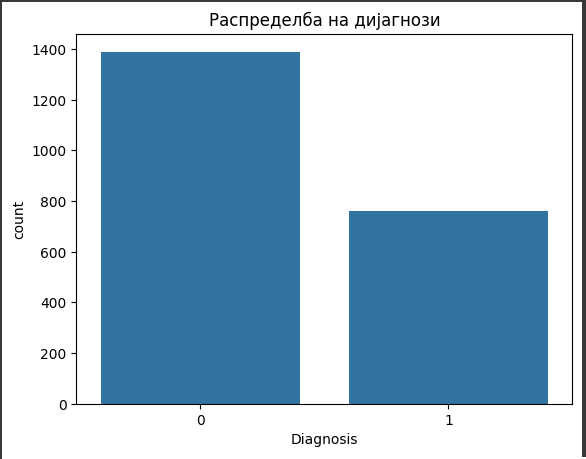
Користениот датасет е именуван како alzheimers\_disease\_data.csv и претставува збир на демографски карактеристики , навики ,фамилијарни поврзаности , когнитивни и медицински мерења кај пациенти кои се предмет на истражување поврзано со Алцхајмеровата болест. Податоците се добиени од kaggle кој се наменети за истражувачки дејности и креирање на модели за предвидување .

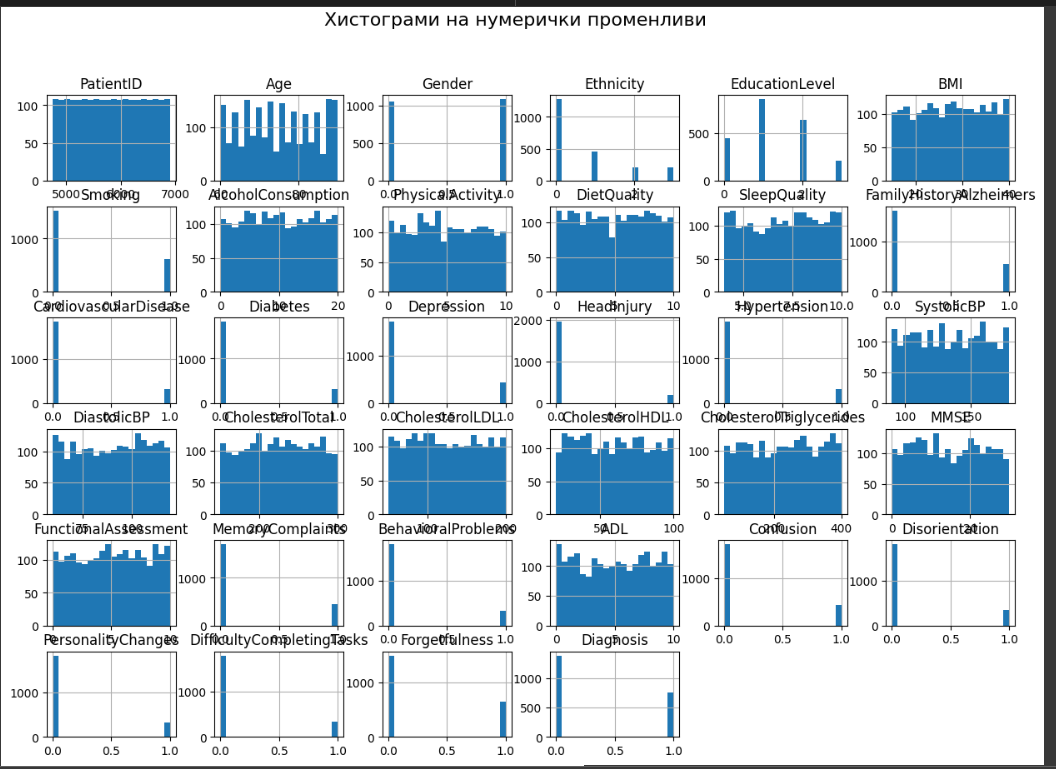
Датасетот содржи приближно 2,149 записа (редови), каде секој запис претставува податок за еден пациент. Секој запис содржи повеќе од 30 променливи, меѓу кои: возраст, пол, степен на образование, навики, повреди, болести, и најбитно дијагноза.

Променливите се од различен тип: нумерички (на пример, возраст) и категориски (на пример, пол дијагноза. Сите податоци се собрани и спремни за анализа.

# **Анализа на податоците**

Се креираат хистограми кои се генерирани врз база на податоците од кои се добиваат значајни информации како на пример како е распределено множеството на податоци кој податоци најдобро корелираат. Од тука па натаму се користат овие податоци за да се одреди правецот на понатамошната обработка на податоците и кои модели би се користеле.





# **Подготовка на податоците**

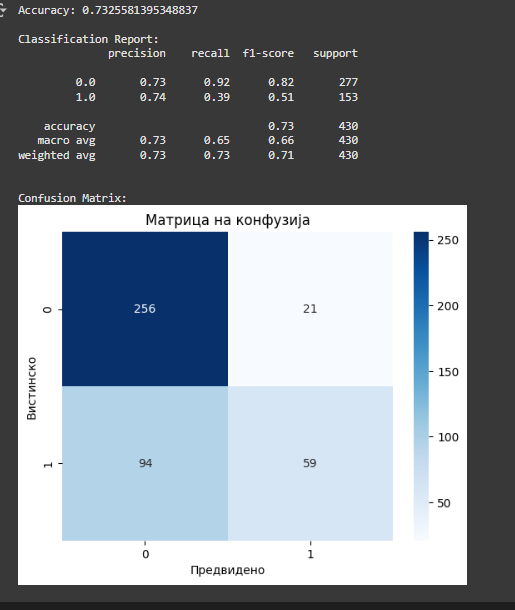
Секако е потребно да се направи проверка за нул вредности и импутација на истите за секој случај доколку се најдат нул вредности во податочното множество. Затоа што како што знаеме овие вредности се најчесто проблем за моделите на учење и голема е веројатноста тие и да не се извршат.

Понатаму се острануваат непотребни колони и продолжуваме со проверката дали има вредности кои се категориски а текстуални и промена на истите во соодветна бројка со помош на LabelEncoder.

# **Креирање на модели**

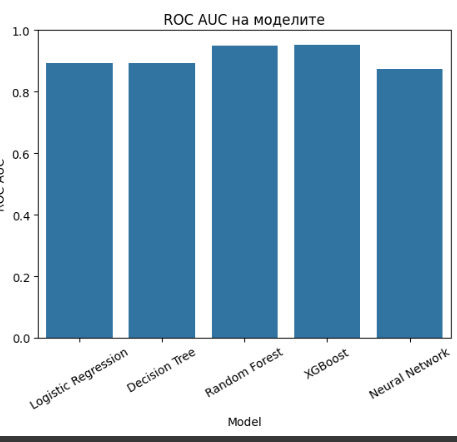
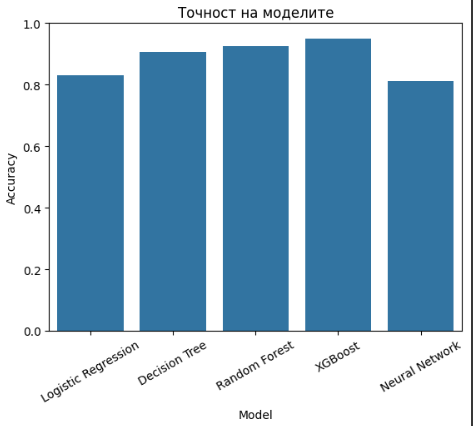
Се креираат најразлични модели почнувајки од не толку сложени како кнн, логистичка регресија, дрва на одлучување и RandomForest, кон посложени како xgboost и невронски мрежи.

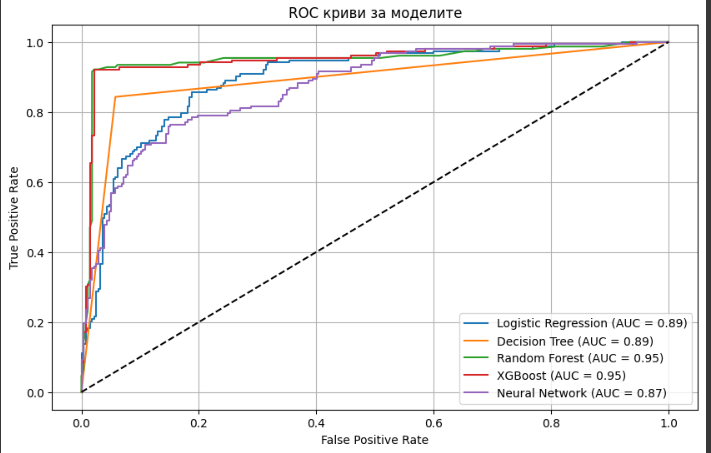
Се креираат сите модели засебно на крај се проверува precision, recall, f1-score и со колкава точност погодува тој модел.



# **Споредба на модели**

Се собираат вредностите со која точност погодуваат моделите и се прикажуваат како хистограмио и роц крива. Според овие хистограми може добро да се визуализира дека овие модели кој подобро кој полошо покажуваат дека перформираат така што добро ги предвидуваат вредностите.





# **Заклучок**

Со овој проект покажуваме дека машинското учење може да понуди прецизни алатки за поддршка на раната дијагноза на алцхамеровата болест. Највлијателни променливи кој доведоа до вакви резултати беа проблеми со меморијата и финкционирањето, проблеми со однесувањето и мрежата за секојдневни активности.

Најдобри резултати беа постигнати со XGBoost, што укажува на силниот потенцијал на напредните модели. Овие модели можат да помогнат во рана интервенција и подобрување на квалитетот на животот на пациентите.

# **Користена литератира**

<https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data>

<https://www.wikipedia.org/>

<https://scikit-learn.org/stable/>

<https://courses.finki.ukim.mk/>