АВГУСТ, 2024

Модел за предсказване на риска от инсулт (Stroke Risk Prediction Model)

Финален проект за Strypeslab Advanced Python Academy

Мария Симеонова

1. Въведение

Цел на проекта

Целта на този проект е да се разработи модел, който може да предсказва риска от инсулт, базиран на различни лични и здравословни фактори, като възраст, пол, здравословно състояние и начин на живот. Този модел може да бъде използван за подпомагане на медицински специалисти при оценката на риска за пациенти, което би могло да допринесе за по-ранна интервенция и превенция.

Кратко описание на модела и какво прави

Проектът използва машинно обучение, за да създаде модел, който анализира различни фактори, свързани с пациента, и предсказва дали този пациент е изложен на висок или нисък риск от инсулт. Моделът е базиран на Random Forest алгоритъм и е обучен с помощта на балансиран набор от данни, за да може по-добре да различава между пациенти с висок и нисък риск от инсулт. След обработка и анализ на данните, моделът постига точност от 87.81%, което показва неговата ефективност в предсказването на риска за пациенти без инсулт, но същевременно отбелязва области за бъдещо подобрение при предсказанията за висок риск.

2. Изисквания

Софтуерни изисквания

- **Python**: Версия 3.6 или по-висока
- **Библиотеки**:
    - wxPython: За графичен потребителски интерфейс
    - pandas: За обработка на данни
    - imblearn: За техники за баланс на данни (SMOTE и RandomUnderSampler)
    - scikit-learn: За модели на машинно обучение и метрики за оценка на модела
- **Допълнителни инструменти**:
    - numpy: За числови операции (често изисквана от други библиотеки)
    - matplotlib (по избор): За визуализация на данни и резултати

Хардуерни изисквания

- **Процесор**: Минимум 1 GHz (препоръчително 2 GHz или по-висок)

- **RAM**: Минимум 2 GB (препоръчително 4 GB или повече за по-добра производителност при работа с големи данни)
- **Дисково пространство**: Минимум 100 MB за инсталация на софтуера и библиотеки, плюс допълнително пространство за съхранение на данни и резултати

Тези изисквания осигуряват необходимата среда за разработка, изпълнение и тестване на проекта за предсказване на риска от инсулт.

3. Данни

Описание на набора от данни

Наборът от данни, използван в проекта, е предоставен от Kaggle и е насочен към предсказване на риска от инсулт. Той съдържа информация за пациенти, която включва различни здравословни и демографски характеристики, използвани за изграждане на модел за предсказване на вероятността от инсулт. Данните са събрани от различни източници и предоставят ценна информация за анализ и обучение на машинно обучение модели.

Линк към използваните данни: Stroke Prediction Dataset (kaggle.com)

Преглед на атрибутите и техните значения

- **id**: Уникален идентификатор на пациента. (Изключен от анализа, тъй като не носи информация относно риска от инсулт).
- **gender**: Пол на пациента. Възможни стойности:
    - "Male" (Мъж)
    - "Female" (Жена)
    - "Other" (Друг) (Изключен от анализа, тъй като не се среща в реалния набор от данни).
- **age**: Възраст на пациента в години.
- **hypertension**: Хипертония. Възможни стойности:
    - 0: Пациентът не страда от хипертония
    - 1: Пациентът страда от хипертония
- **heart_disease**: Сърдечни заболявания. Възможни стойности:
    - 0: Пациентът не страда от сърдечни заболявания

    - 1: Пациентът страда от сърдечни заболявания
- **ever_married**: Състояние на брака. Възможни стойности:
    - "No" (Не)
    - "Yes" (Да)
- **work_type**: Вид на работата. Възможни стойности:
    - "children" (Деца)
    - "Govt_job" (Държавна работа)
    - "Never_worked" (Не е работил)
    - "Private" (Частен сектор)
    - "Self-employed" (Самостоятелно зает)
- **Residence_type**: Тип на мястото на живеене. Възможни стойности:
    - "Rural" (Селски)
    - "Urban" (Градски)
- **avg_glucose_level**: Средно ниво на глюкоза в кръвта.
- **bmi**: Индекс на телесна маса (Body Mass Index).
- **smoking_status**: Статус на тютюнопушенето. Възможни стойности:
    - "formerly smoked" (Пушил в миналото)
    - "never smoked" (Никога не е пушил)
    - "smokes" (Пуши)
    - "Unknown" (Неизвестно) (Информацията не е налична за пациента)
- **stroke**: Инсулт. Възможни стойности:
    - 0: Пациентът не е имал инсулт
    - 1: Пациентът е имал инсулт

Тези атрибути предоставят основната информация, необходима за анализа и изграждането на модела, като се използват за обучение и тест на предсказателната способност на модела.

4. Процес на разработка

4.1 Описание на correlations.py

Файлът **correlations.py** е отговорен за обработката на данните и извършването на корелационен анализ. В този файл се изпълняват следните стъпки:

Зареждане на данните: Данните се зареждат от CSV файл (healthcare-dataset-stroke-data.csv) с помощта на pandas.

Обработка на липсващи стойности:

- Първоначално се премахват редовете с липсващи стойности в колоната bmi. След това липсващите стойности се попълват с медианата на bmi, изчислена от наличните данни.

Почистване на данните:

- Премахват се редове с "Other" в колоната gender и "Unknown" в колоната smoking_status.
- Колоната id, която служи като уникален идентификатор, се изключва от анализа.

Кодиране на категориалните променливи:

- Колоните gender, ever_married, Residence_type, work_type, и smoking_status се преобразуват в числови стойности за улесняване на анализа.

Подготовка на данните за моделиране:

- Данните се разделят на обучаваща и тестова извадка с помощта на train_test_split.

Обучение на модела:

- Използва се Random Forest класификатор за обучение върху обучаващия набор от данни.

Оценка на модела:

- Оценяват се резултатите на модела чрез метрики като точност, матрица на объркването и отчет за класификация.

Анализ на важността на признаците:

- Изчисляват се важността на признаците за модела и се създава DataFrame за визуализиране на важността на признаците.

Анализ на корелации:

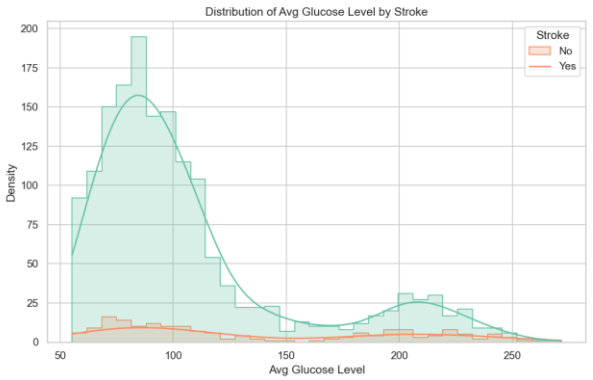
- Извършват се корелационни анализи между непрекъснати променливи и целевата променлива (stroke).

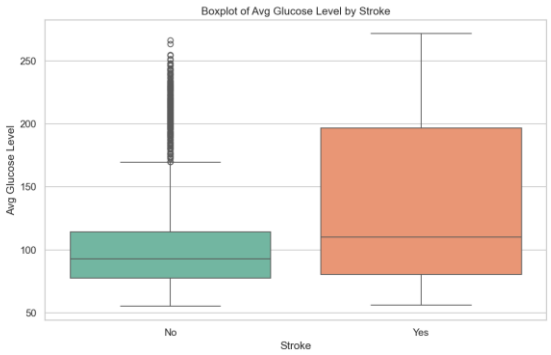
- Провеждат се Chi-Square тестове за категориални променливи, за да се определи тяхната връзка с целевата променлива.
- Основната цел на корелационния анализ е да отговори на няколко ключови въпроса, свързани с факторите, които могат да повлияят на вероятността от получаване на инсулт. Тези въпроси включват:
    - *Играе ли възрастта роля при получаването на инсулт?*
    - *По-податливи ли са пушачите към инсулт?*
    - *По-голям ли е рискът от инсулт при хора със сърдечни заболявания?*
    - *По-податливи ли са мъжете на инсулти?*
- Целта на анализа е да се изследват връзките между тези фактори и вероятността от инсулт, като се използват корелации и статистически тестове.

Визуализации и изводи:

Създават се различни визуализации, включително хистограми, кутии и разпръснати диаграми, за да се изследват разпределенията на променливите и техните връзки с инсулт.

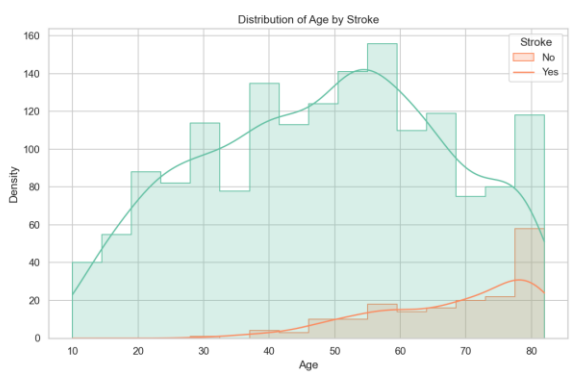
Разпределение на средното ниво на глюкоза според инсулт

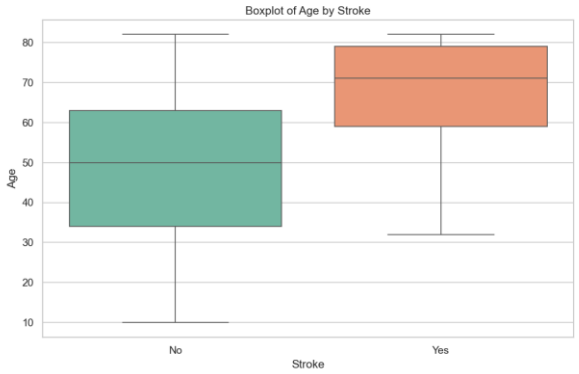


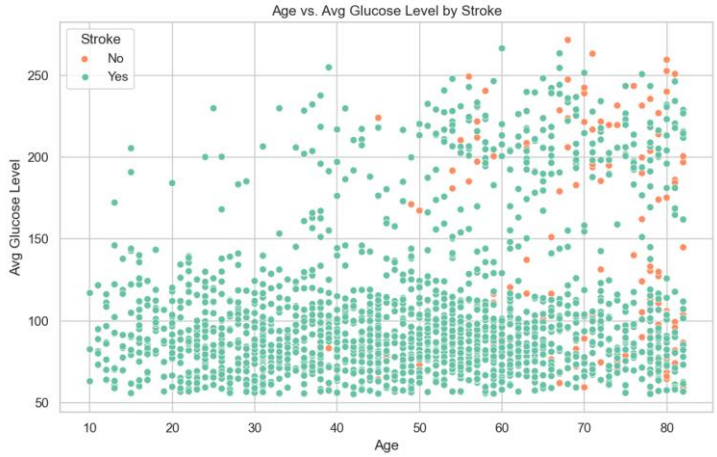


Заключение: Средното ниво на глюкоза (avg_glucose_level) има *най-висока важност* (Importance: 0.278987) сред разгледаните фактори. Корелацията между средното ниво на глюкоза и инсулт е 0.157, което показва умерена връзка. Лицата с по-високо средно ниво на глюкоза са по-податливи на инсулт.

Разпределение на възрастта според инсулт

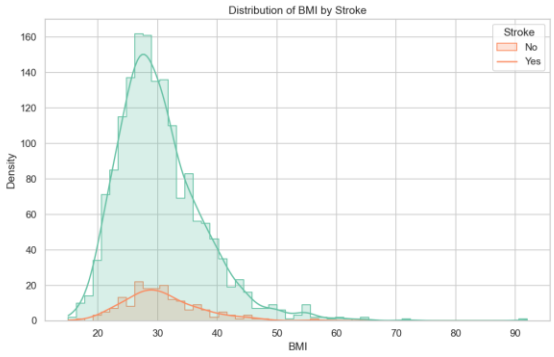


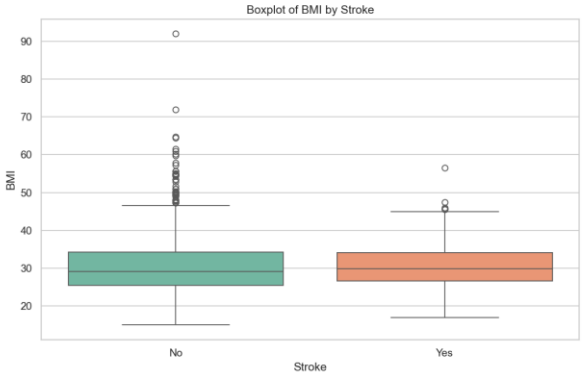


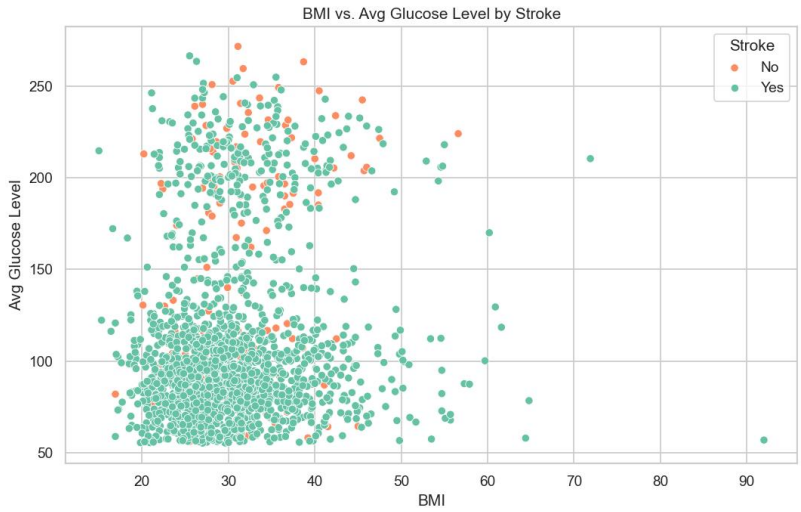


Заключение: Възрастта (age) е *вторият по важност фактор* (Importance: 0.268266) с корелация 0.314. Този резултат показва значителна *положителна връзка между възрастта и риска от инсулт*, като по-възрастните индивиди са по-изложени на риск.

Разпределение на BMI според инсулт

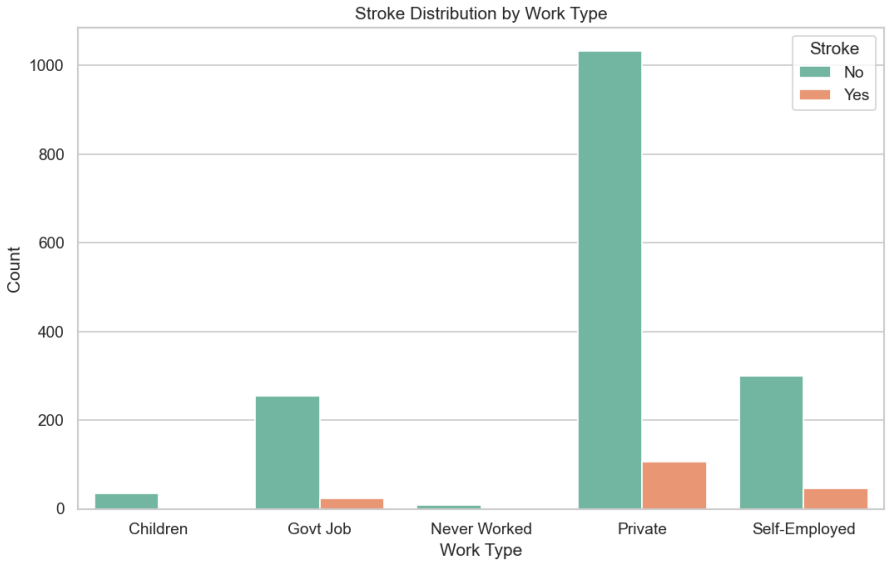






Заключение: BMI (body mass index) е *третият по важност фактор* (Importance: 0.224400) с много ниска корелация от 9.1496e-05, което показва, че този фактор има *по-слаба връзка с инсулта*, въпреки важността му в модела.

Разпределение на типа работа според инсулт

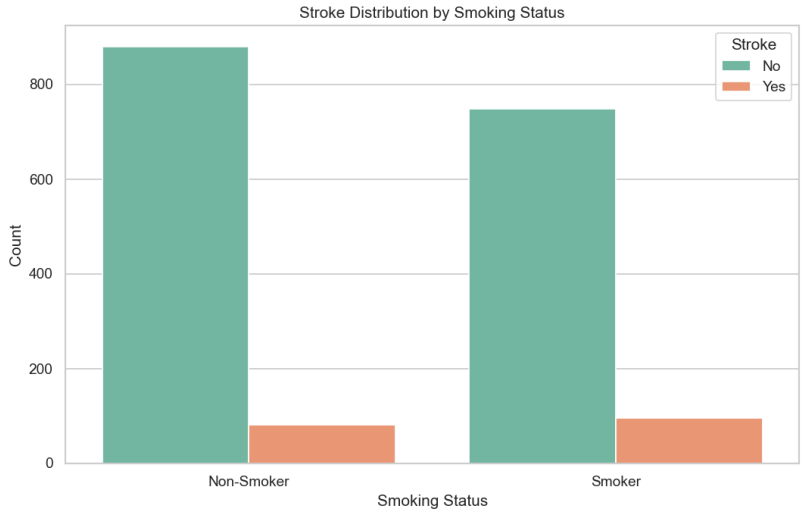


Заключение: Типът работа (work_type) има *средна важност* (Importance: 0.052506) с Chi-Square стойност 13.94 и P-стойност 0.007, което предполага, че типът работа *може да влияе върху вероятността за инсулт*.

Разпределение на местожителството според инсулт

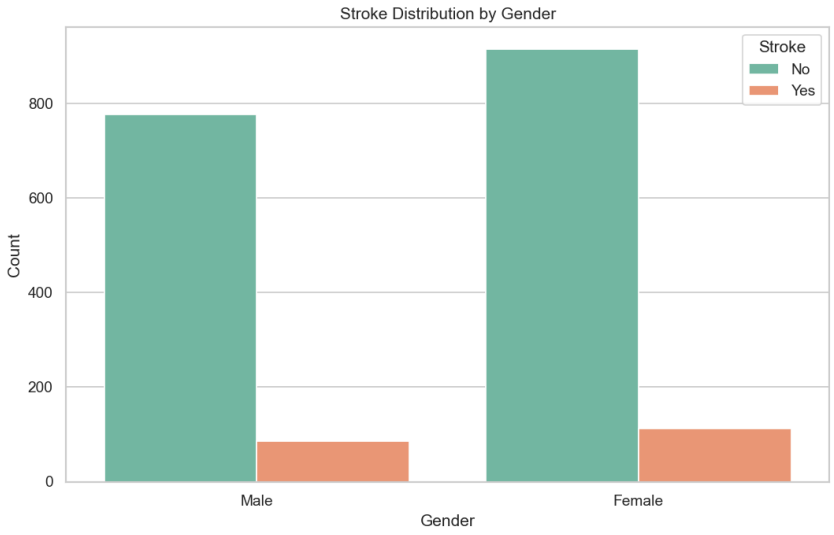
Заключение: Типът на местожителство (Residence_type) има *сравнително ниска важност* (Importance: 0.034281) и незначима Chi-Square стойност, което предполага, че *няма значителна връзка между местожителството и риска от инсулт*.

Разпределение на пушенето според инсулт



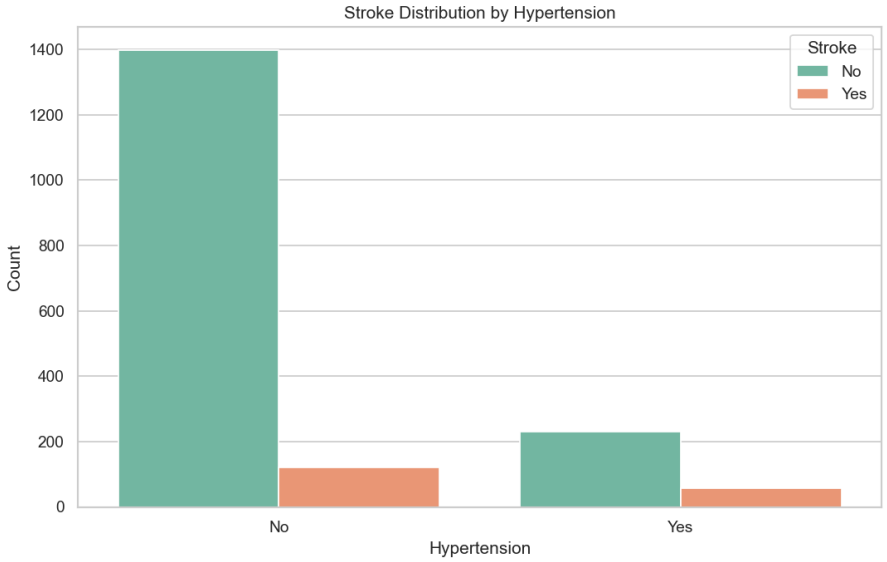
Заключение: Статусът на пушене (smoking_status) показва *умерена важност* (Importance: 0.032423) с P-стойност 0.0187, което предполага, че пушачите са *по-податливи на инсулт*.

Разпределение на пола според инсулт



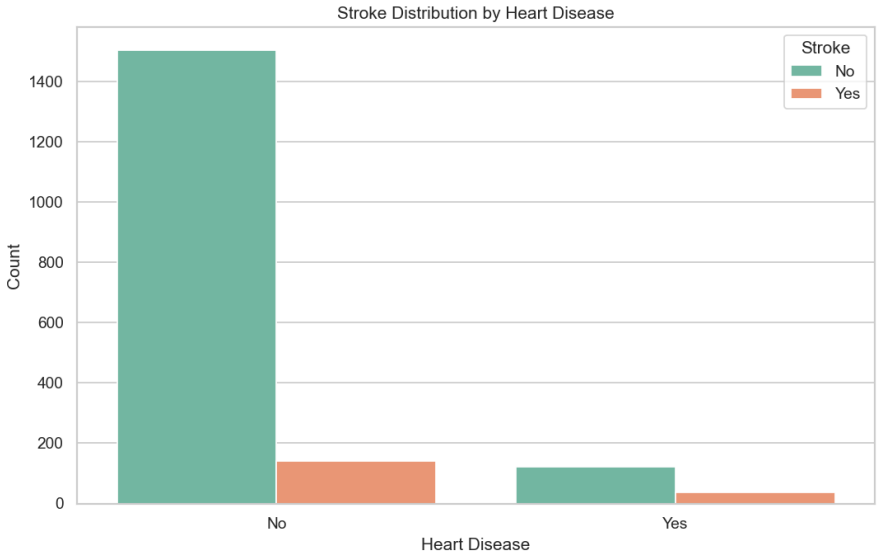
Заключение: Полът (gender) показва *ниска важност* (Importance: 0.032110) и незначима Chi-Square стойност, което означава, че полът сам по себе си *не е определящ фактор* за риска от инсулт.

Разпределение на хипертонията според инсулт



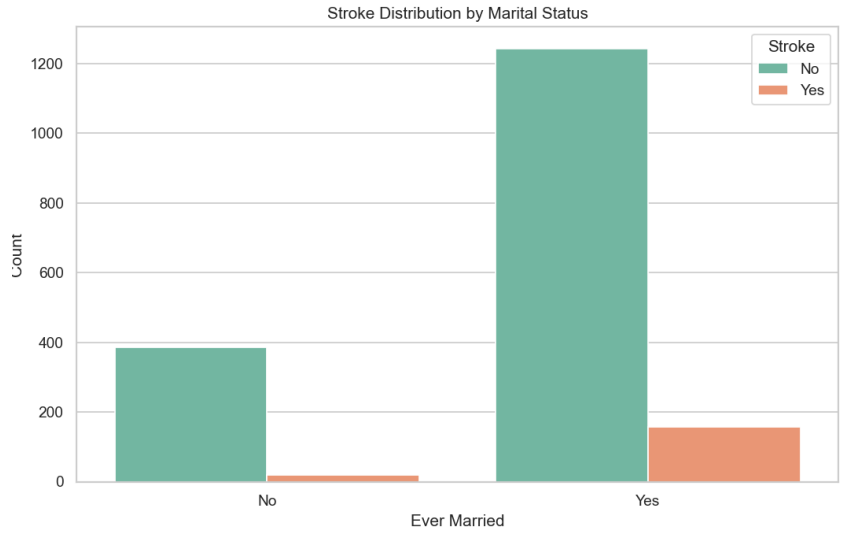
Заключение: Хипертонията (hypertension) показва *умерена важност* (Importance: 0.031945) с много значима P-стойност (1.3486e-08), което показва *силна връзка между наличието на хипертония и риска от инсулт*.

Разпределение на сърдечните заболявания според инсулт



Заключение: Сърдечните заболявания (heart_disease) имат *сравнително ниска важност* (Importance: 0.025300), но с много значима P-стойност (1.4190e-07), което подсказва, че хората със сърдечни заболявания имат *по-голям риск от инсулт*.

Разпределение на семейното положение според инсулт



Заключение: Семейното положение (ever_married) показва *ниска важност* (Importance: 0.019783) и значима P-стойност, което подсказва, че *съществува връзка между семейното положение и риска от инсулт*.

4.2 Описание на choosebestclassifier.py

Този скрипт има за цел да избере най-добрия класификатор за прогнозиране на риска от инсулт от предоставения набор от данни. Използват се различни машинни обучителни алгоритми и се извършва оценка на техните представяния, за да се определи най-подходящият модел. Основни стъпки в скрипта:

Зареждане и Предварителна Обработка на Данните по същият начин както в 4.1

Разделяне на Данните:

- Данните са разделени на обучаващи и тестови набори.
- Приложен е SMOTE (Synthetic Minority Over-sampling Technique) за справяне с дисбаланса между класовете.

Оценка на Класификаторите:

- Инциализиране на няколко класификатора: Логистична регресия, Решаващо дърво, Случайна гора, Градиентен бустинг, Поддържаща векторна машина, К-най-близките съседи и Наивен Байес.
- Трениране на класификаторите и оценяване на техните представяния връз основа на метрики като точност, прецизност, извличане, F1 оценка и AUC (площ под ROC кривата).
- Визуализиране на матриците на объркванията и ROC кривите за сравнение на представянето на класификаторите.

Резултати:

- Показва сравнителен анализ на представянето на различните класификатори.
- Визуализира резултатите чрез графики и диаграми за по-добро разбиране на ефективността на моделите.

Цел на скрипта:

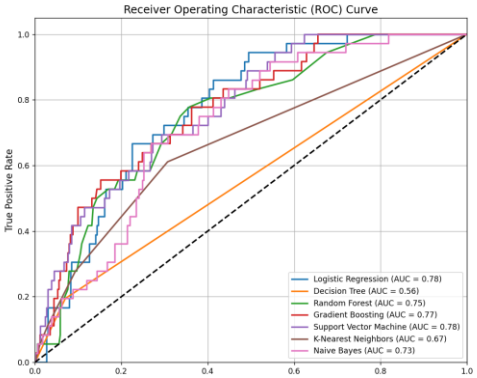
- Целта на **choosebestclassifier.py** е да идентифицира най-ефективния класификатор за прогнозиране на риска от инсулт връз основа на наличните данни. Резултатите от анализа помагат да се избере най-добрият модел за по-нататъшна употреба и приложение.

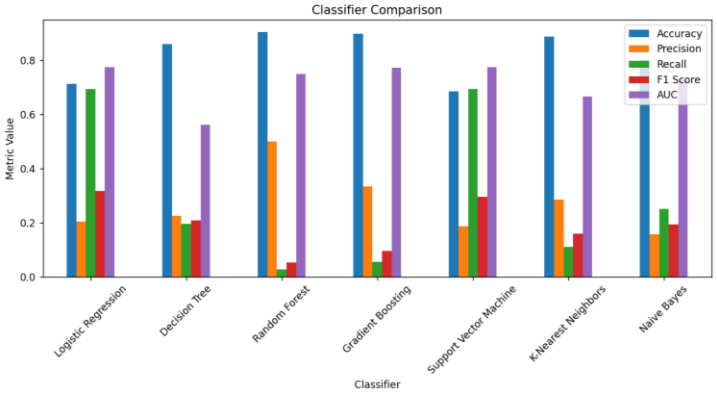
Визуализации:

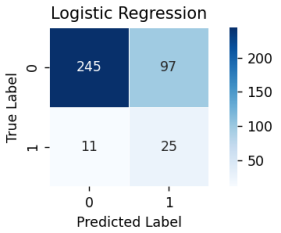
- **Матрици на объркванията:** Показват точността на предсказанията на класификаторите.
- **ROC криви:** Сравняват способността на класификаторите да различават между положителни и отрицателни примери.

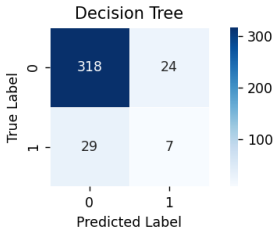
- **Бар графики:** Показват сравнението на метриките (точност, прецизност, извличане, F1 оценка, AUC) за различните класификатори.

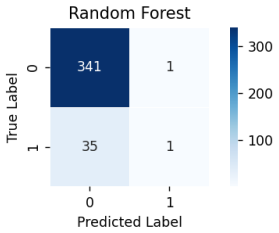
Преди прилагане на SMOTE (Synthetic Minority Over-sampling Technique) за справяне с дисбаланса между класовете:

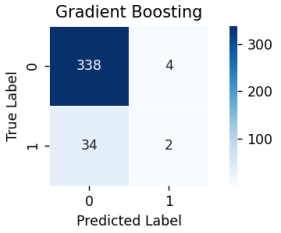


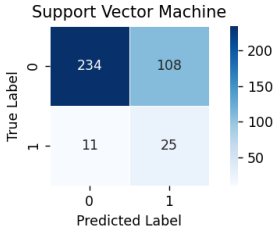


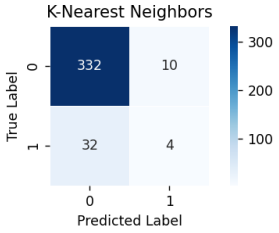


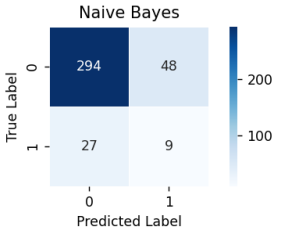




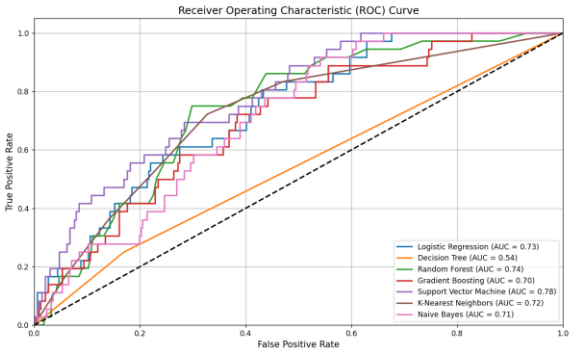


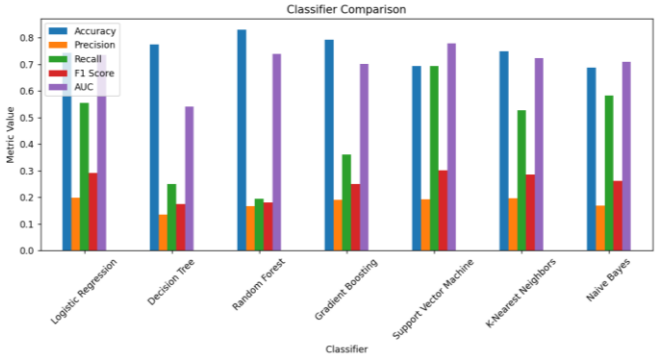


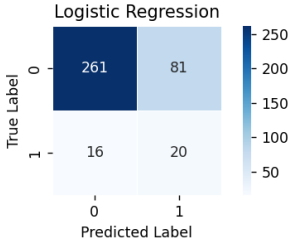


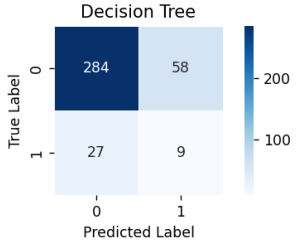


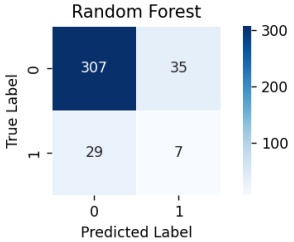
След прилагане на SMOTE:

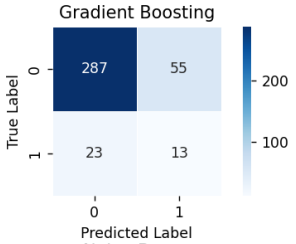


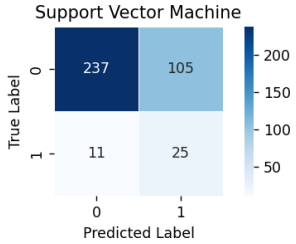


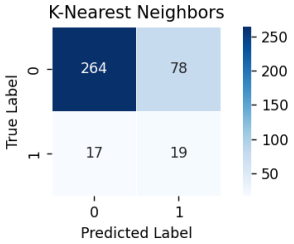


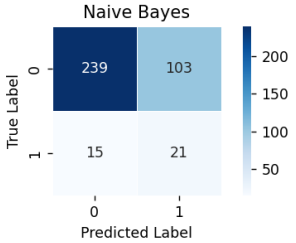












Основни Заключения:

- **Най-добър Класификатор:** Случайната гора (Random Forest) показва най-висока точност (0.828) и добра F1 оценка (0.217), но с по-ниска прецизност и извличане. Това я прави подходящ избор за моделиране на риска от инсулт, въпреки че резултатите могат да се подобрят с допълнителна настройка на параметрите.
- **Най-добро Извличане:** Поддържаща векторна машина (SVM) има най-високо извличане (0.694), но с по-ниска точност. Това може да е полезно в контексти, когато е важно да се идентифицират всички положителни случаи на инсулт, дори ако това води до по-висок процент на фалшиви положителни резултати.

- **Най-ниска Прецизност:** Логистичната регресия и Наивният Байес имат най-ниска прецизност (около 0.17-0.18), което означава, че те често дават грешни предсказания за положителни случаи.

4.3 Описание на prediction.py

Този файл съдържа основния код за обучение и предсказване на риска от инсулт, използвайки Random Forest класификатор. Кодът преминава през различни стъпки, включително предварителна обработка на данните, справяне с несъответствията в класи, балансиране на класовете чрез методите SMOTE и RandomUnderSampler, и финално предсказване на риска връз основа на входните данни. Основни стъпки в скрипта:

1. **Зареждане на данни**: Данните се зареждат от CSV файл (healthcare-dataset-stroke-data.csv) и след това се обработват за премахване на липсващи стойности в колоната bmi, както и за филтриране на некоректни или неизвестни стойности за пол и статус на пушене.

2. **Предварителна обработка на данните**: Категориалните променливи като gender, ever_married, Residence_type, work_type и smoking_status се преобразуват в числови стойности чрез използване на мапинг.

3. **Баланс на класовете**: Използват се SMOTE и RandomUnderSampler за да се справи с дисбаланса на класовете в тренировъчния набор от данни.

4. **Обучение на модела**: Random Forest класификаторът се обучава върху балансирания тренировъчен набор.

5. **Оценка на модела**: Моделът се оценява върху тестовия набор от данни чрез изчисляване на точност, матрица на объркванията и отчет за класификация.

6. **Предсказване на риска от инсулт**: В края на скрипта има функция predict_stroke, която позволява на потребителя да въведе лични данни (пол, възраст, хипертония, сърдечни заболявания и т.н.), за да получи предсказание дали е в риск от инсулт.

Основни Метрики и Резултати:

```
Accuracy: 0.8781163434903048
Confusion Matrix:
 [[306  27]
 [ 17  11]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.92      0.93       333
           1       0.29      0.39      0.33        28

    accuracy                           0.88       361
   macro avg       0.62      0.66      0.63       361
weighted avg       0.90      0.88      0.89       361
```

- **Точност на модела**: Моделът постигна точност от 87.81% върху тестовия набор от данни. Това показва, че моделът е ефективен в предсказването на риска от инсулт, въпреки че има предизвикателства при предсказването на по-редки случаи (клас 1 - хората с висок риск).
- **Матрица на объркванията**: Показва броя на правилните и неправилните предсказания.
    - True Negatives (TN): 306 - Моделът правилно идентифицира 306 случая, при които няма риск от инсулт.
    - False Positives (FP): 27 - Моделът погрешно класифицира 27 случая като висок риск от инсулт, когато всъщност няма такъв риск.
    - False Negatives (FN): 17 - Моделът пропусна 17 случая, при които има висок риск от инсулт.
    - True Positives (TP): 11 - Моделът правилно идентифицира 11 случая с висок риск от инсулт.
- **Отчет за класификация**:
    - Прецизност (precision):
        - Клас 0 (Нисък риск): 0.95
        - Клас 1 (Висок риск): 0.29
    - Извличане (recall):
        - Клас 0 (Нисък риск): 0.92
        - Клас 1 (Висок риск): 0.39
    - F1 Оценка:
        - Клас 0 (Нисък риск): 0.93
        - Клас 1 (Висок риск): 0.33

Тези метрики показват, че моделът е по-ефективен в предсказването на случаи с нисък риск от инсулт, но има трудности при разпознаването на редките случаи с висок риск.

Предсказване на риска от инсулт:

Функцията predict_stroke позволява на потребителя да въведе определени характеристики и след това да получи предсказание, базирано на обучената Random Forest модел. Тази функция е полезна за потребители, които искат да оценят личния си риск от инсулт.

4.4 Описание на prediction_with_interface.py

Основни Компоненти:

Разположение на UI & стилове:

- Разположението се управлява с помощта на wx.GridBagSizer, което е гъвкаво за подреждане на елементи.
- Включени са потребителски цветове за фона, текста и бутоните, което подобрява визуалния изглед.
- wx.StaticText за заглавията и резултатите използва различни размери и цветове на шрифта, за да се подчертае важността.

Обучение на модела:

- Функцията за обучение чете и предварително обработва набора от данни, обработва липсващи стойности и конвертира категорийни данни в числови.
- Моделът се обучава с комбинация от SMOTE и RandomUnderSampler за справяне с дисбаланса на класовете преди използване на RandomForestClassifier.

Обработка на прогнозите:

- Входните полета се валидират за коректност (например възраст, нива на глюкоза и BMI в определени граници).
- Ако възникне грешка, се показва съобщение за грешка с червен текст и по-малък шрифт, за да уведоми потребителите.
- Резултатите от прогнозите се показват с по-голям шрифт, като цветът показва нивото на риск (зелен за нисък риск, червен за висок риск).

5. Инструкции за използване

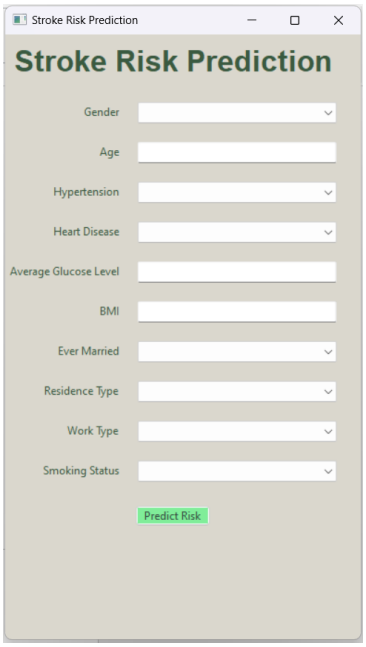
Стартиране на проекта

1. **Изисквания:** Уверете се, че всички необходими библиотеки, включени във файла, са инсталирани. Например, използвайте командата pip install -r requirements.txt в терминала, за да инсталирате всички нужни зависимости.

2. **Стартиране:** За да стартирате проекта, просто пуснете файла *prediction_with_interface.py*. Това автоматично ще зареди прозореца с потребителския интерфейс.

Използване на потребителския интерфейс

3. След като стартирате файла, ще се отвори прозорец, в който можете да въведете данните си.



4. Някои полета имат падащи менюта (dropdown менюта), от които можете да изберете подходящата стойност:

    1. **Gender:** Male, Female
    2. **Hypertension:** No, Yes
    3. **Heart Disease:** No, Yes
    4. **Ever Married:** No, Yes
    5. **Residence Type:** Rural, Urban
    6. **Work Type:** Children, Govt_job, Never_worked, Private, Self-employed
    7. **Smoking Status:** Never smoked, Formerly smoked, Smokes

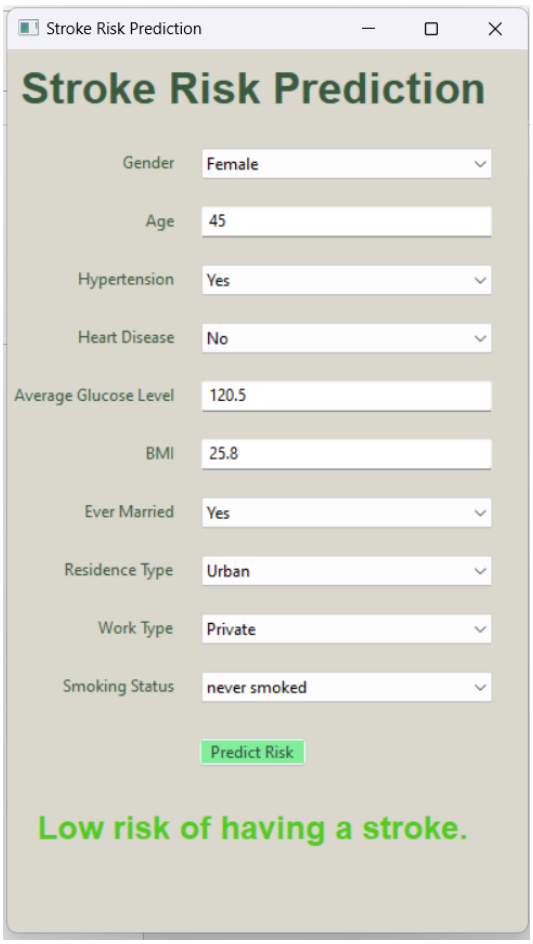
    Полетата **Age**, **Average Glucose Level** и **BMI** изискват въвеждане на числови стойности. Уверете се, че въвеждате валидни стойности за тях.

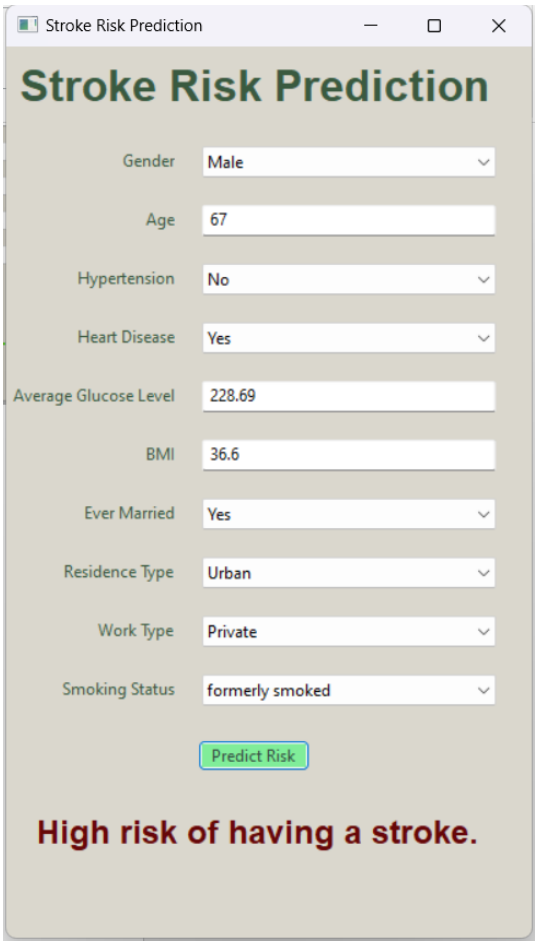
Примери за входни данни и очаквани резултати

- **Входни данни:**
    - Gender: Female
    - Age: 45
    - Hypertension: Yes
    - Heart Disease: No

    - Average Glucose Level: 120.5
    - BMI: 25.8
    - Ever Married: Yes
    - Residence Type: Urban
    - Work Type: Private
    - Smoking Status: Never smoked
- **Очакван резултат:** **Нисък риск от инсулт.**

- **Входни данни:**
    - Gender: Male
    - Age: 67
    - Hypertension: N
    - Heart Disease: Yes
    - Average Glucose Level: 228.69
    - BMI: 36.6
    - Ever Married: Yes
    - Residence Type: Urban
    - Work Type: Private
    - Smoking Status: Formerly smoked
- **Очакван резултат:** **Висок риск от инсулт.**





6. Резултати и анализ

Анализ на резултатите от модела

Моделът постигна точност от 87.81% при тестовите данни. Това показва добра способност за предсказване на риска от инсулт, особено за пациенти без инсулт, но има ниска точност при предсказване на случаи с висок риск от инсулт.

Обсъждане на точността и предсказанията

Въпреки че моделът показва висока обща точност, резултатите показват значителен дисбаланс в предсказанията. Моделът има висока точност за отрицателните случаи (без инсулт), но значително по-ниска точност при идентифициране на положителни случаи (с инсулт), което може да се дължи на дисбаланса в данните. Необходимо е да се разгледат допълнителни методи за подобряване на способността за предсказване на положителните случаи.

7. Заключение

Обобщение на постигнатото

Проектът успешно разработи модел за предсказване на риска от инсулт, базиран на различни фактори като възраст, пол, здравословно състояние и начин на живот. Моделът постигна задоволителна обща точност от 87.81%, което демонстрира неговата ефективност в предсказването на риска от инсулт за пациенти без инсулт.

Следващи стъпки и възможности за подобрение

Въпреки добрите резултати, моделът показва ниска точност при идентифициране на пациенти с висок риск от инсулт. За да се подобри тази област, могат да се изследват следните стъпки:

- Използване на различни техники за балансиране на данните, като SMOTE в комбинация с други методи за ресемплиране.
- Изследване на допълнителни модели и хиперпараметрично търсене за подобряване на точността на предсказанията.
- Разширяване на набора от данни с допълнителни релевантни характеристики или използване на различни източници на данни за повишаване на представителността на модела.