



АНАЛИЗ И ПРЕДСКАЗВАНЕ НА РИСКА ОТ ДИАБЕТ

Изготвил: Ерика Карамучева,
ФН:2101321067

ЦЕЛИ НА ПРОЕКТА

Целта на проекта е да се проследят факторите, които влияят върху риска от развитие на диабет. Ранното предсказване на диабет е ключово за предотвратяване на сериозни усложнения като сърдечно-съдови заболявания, увреждане на нервната система и бъбречна недостатъчност. Навременната диагностика позволява по-ефективно управление на състоянието чрез промени в начина на живот и медикаментозна терапия, което значително подобрява качеството на живот на пациентите. За целта са използвани данните от публичната база- <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>. С тяхна помощ, както и с помощта на софтуера Orange ще се опитаме да разберем кои са предпоставките, водещи до заболяването.

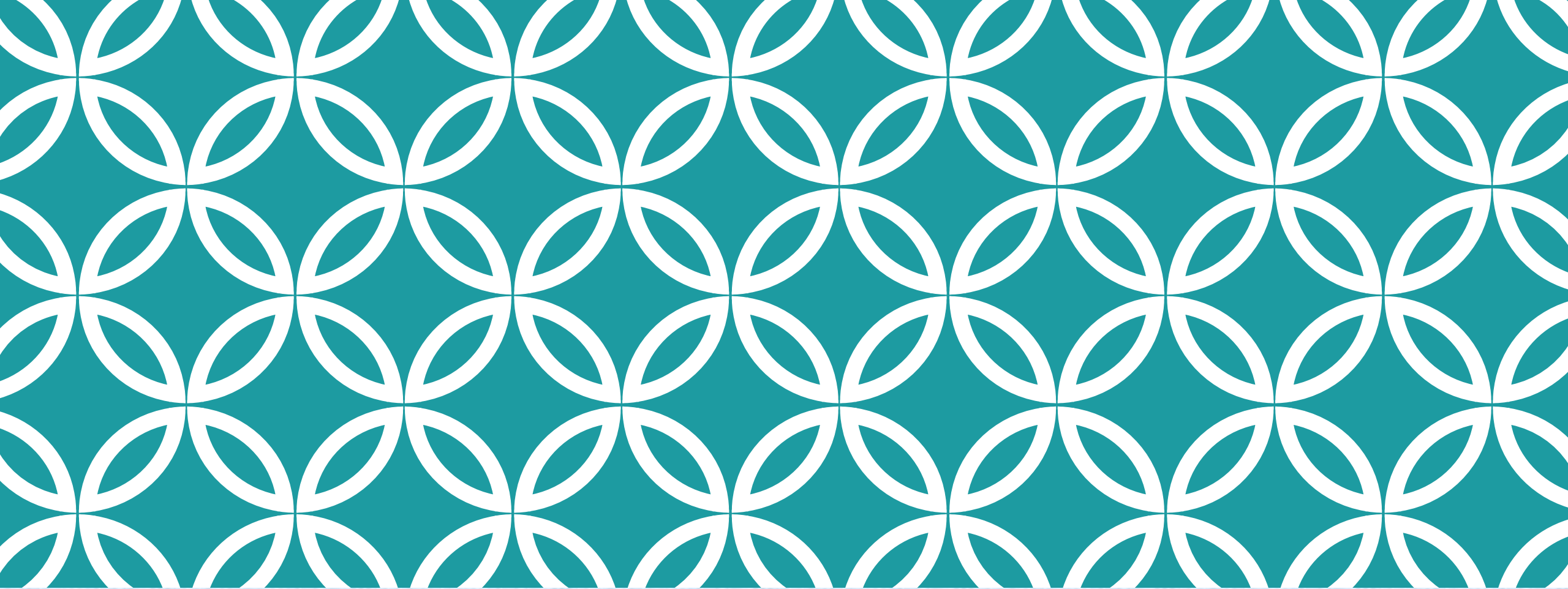
ОПИСАНИЕ НА ДАННИТЕ

В базата се съхраняват 9 основни характеристики за всеки запис:

- ❖ Пол
- ❖ Възраст
- ❖ Наличие на хипертония
- ❖ Наличие на сърдечно заболяване
- ❖ История на пушене
- ❖ Индекс на телесни мазнини
- ❖ Гликиран хемоглобин (HbA1c)
- ❖ Нива на глюкоза в кръвта
- ❖ Наличие на диабет

ОПИСАНИЕ НА ДАННИТЕ

Data Table - Orange										
Info										
100000 instances (no missing data)										
8 features										
Target with 2 values										
No meta attributes.										
Variables										
<input checked="" type="checkbox"/> Show variable labels (if present)										
<input type="checkbox"/> Visualize numeric values										
<input checked="" type="checkbox"/> Color by instance classes										
Selection										
<input checked="" type="checkbox"/> Select full rows										
Restore Original Order										
<input checked="" type="checkbox"/> Send Automatically										
	diabetes	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	
1	0	Female	80.00	0	1	never	25.19	6.6	140	
2	0	Female	54.00	0	0	No Info	27.32	6.6	80	
3	0	Male	28.00	0	0	never	27.32	5.7	158	
4	0	Female	36.00	0	0	current	23.45	5.0	155	
5	0	Male	76.00	1	1	current	20.14	4.8	155	
6	0	Female	20.00	0	0	never	27.32	6.6	85	
7	1	Female	44.00	0	0	never	19.31	6.5	200	
8	0	Female	79.00	0	0	No Info	23.86	5.7	85	
9	0	Male	42.00	0	0	never	33.64	4.8	145	
10	0	Female	32.00	0	0	never	27.32	5.0	100	
11	0	Female	53.00	0	0	never	27.32	6.1	85	
12	0	Female	54.00	0	0	former	54.70	6.0	100	
13	0	Female	78.00	0	0	former	36.05	5.0	130	
14	0	Female	67.00	0	0	never	25.69	5.8	200	
15	0	Female	76.00	0	0	No Info	27.32	5.0	160	
16	0	Male	78.00	0	0	No Info	27.32	6.6	126	
17	0	Male	15.00	0	0	never	30.36	6.1	200	
18	0	Female	42.00	0	0	never	24.48	5.7	158	
19	0	Female	42.00	0	0	No Info	27.32	5.7	80	
20	0	Male	37.00	0	0	ever	25.72	3.5	159	
21	0	Male	40.00	0	0	current	36.38	6.0	90	



АНАЛИЗ НА ДАННИТЕ



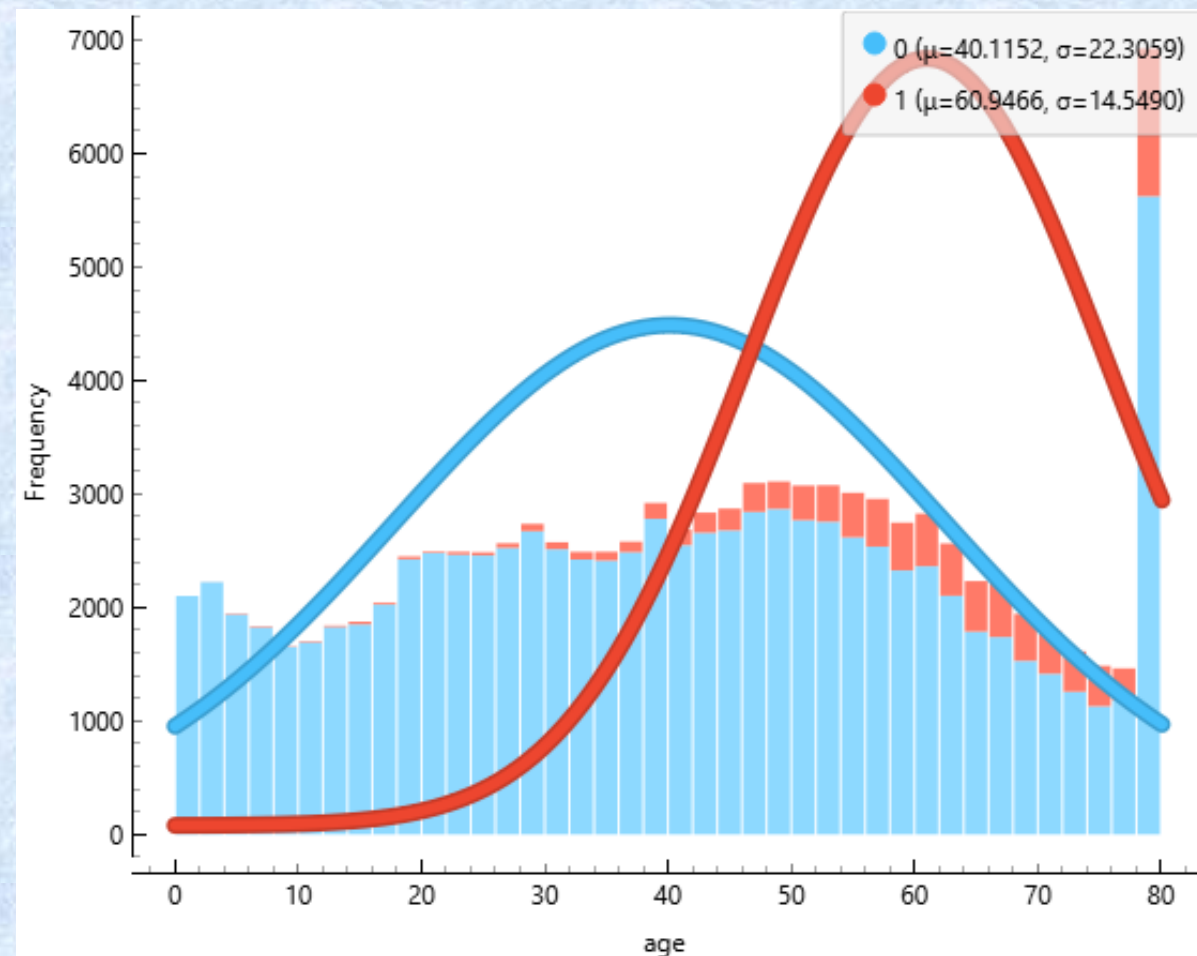
АНАЛИЗ НА ДАННИТЕ

Ще използваме няколко различни визуализации, с помощта на които ще проследим как и кои фактори влияят върху развитието на диабет.



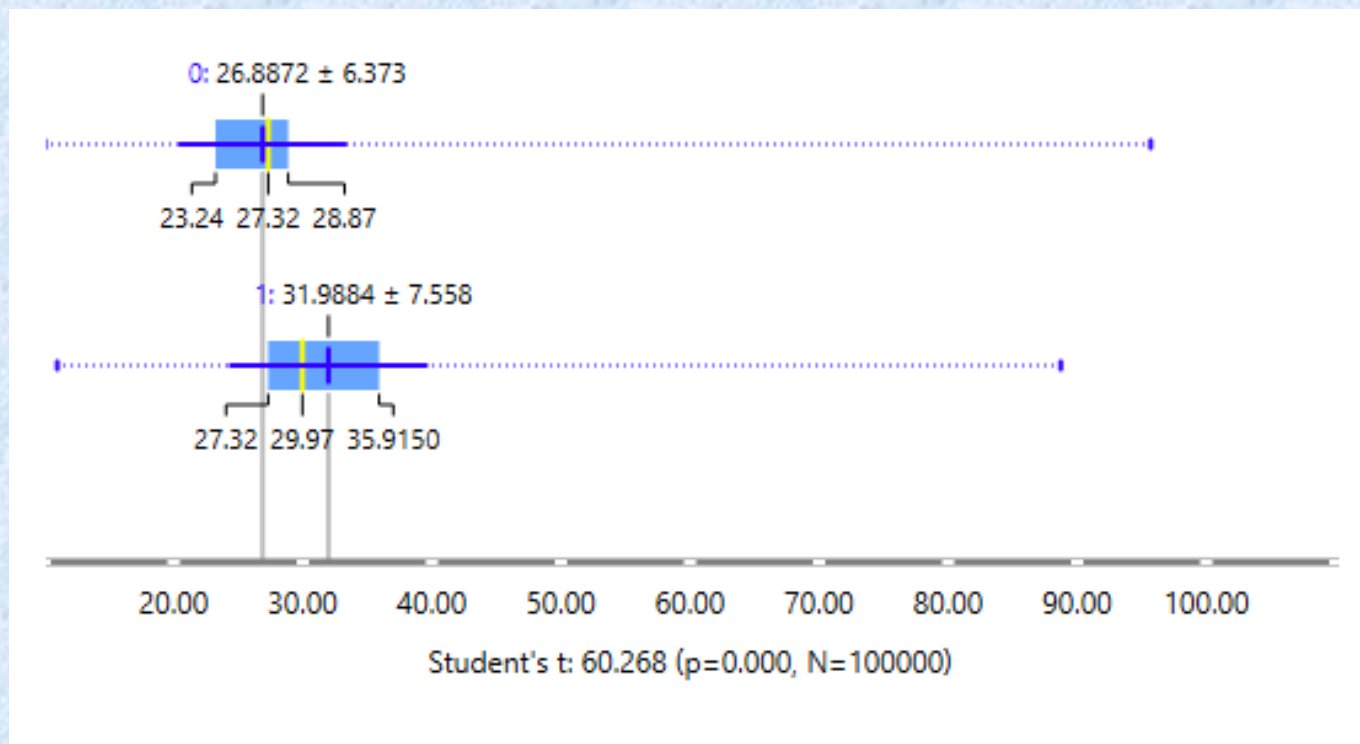
ВЪЗРАСТТА КАТО ФАКТОР ЗА РАЗВИТИЕ НА ДИАБЕТ

Според изложените данни с напредването на възрастта се увеличават и случаите на диабет. Единични случаи са регистрирани още в ранна детска възраст- между 2 и 4 години, но по- голям риск съществува за хората над 40 години.



ВЛИЯЕ ЛИ ИНДЕКСЪТ НА ТЕЛЕСНИ МАЗНИНИ ЗА РАЗВИТИЕТО НА ДИАБЕТ?

Оказва се, че хората, страдащи от диабет имат по- висок индекс на телесни мазнини. Това от своя страна ни навежда на мисълта, че хората с наднормено тегло са по-склонни да развият диабет, отколкото тези, които поддържат оптимално тегло.



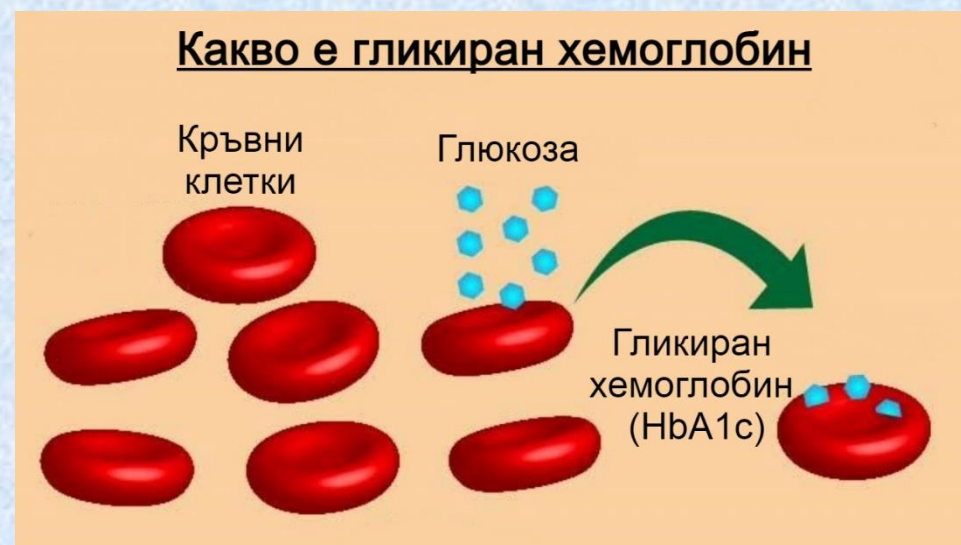
ВЛИЯНИЕ НА НИВАТА НА ГЛЮКОЗА В КРЪВТА И ГЛИКИРАН ХЕМОГЛОБИН

Какво е глюкоза? - Глюкозата е монозахарид, който е в основата на производство на енергия в организма. Тя е неразделна част от живота. Някои тъкани като мозъка, например, се нуждаят от постоянно снабдяване с нея. Глюкозата е известна и като „кръвна захар“, тъй като циркулира в кръвта ни като източник на лесно достъпна енергия.



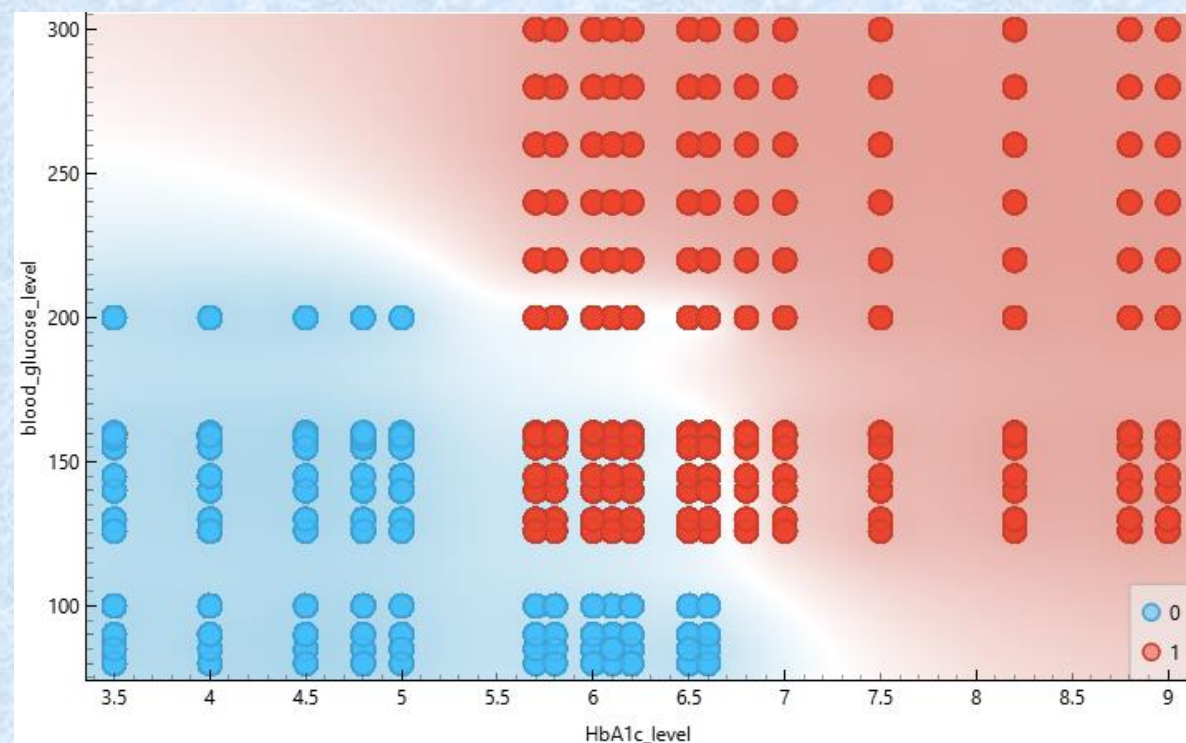
ВЛИЯНИЕ НА НИВАТА НА ГЛЮКОЗА В КРЪВТА И ГЛИКИРАН ХЕМОГЛОБИН

Какво е гликиран хемоглобин? - Гликираният хемоглобин (HbA1c) първоначално е идентифициран като "необичайна" форма на хемоглобин при пациенти със захарен диабет преди повече от 40 години . След това откритие са проведени множество проучвания, които показват, че нивото на гликираният хемоглобин е в тясна зависимост от средното ниво на кръвната захар. Тогава се поражда идеята, че измерването на HbA1c може да служи като един надежден маркер за оценка на гликемичния контрол при хората със захарен диабет.



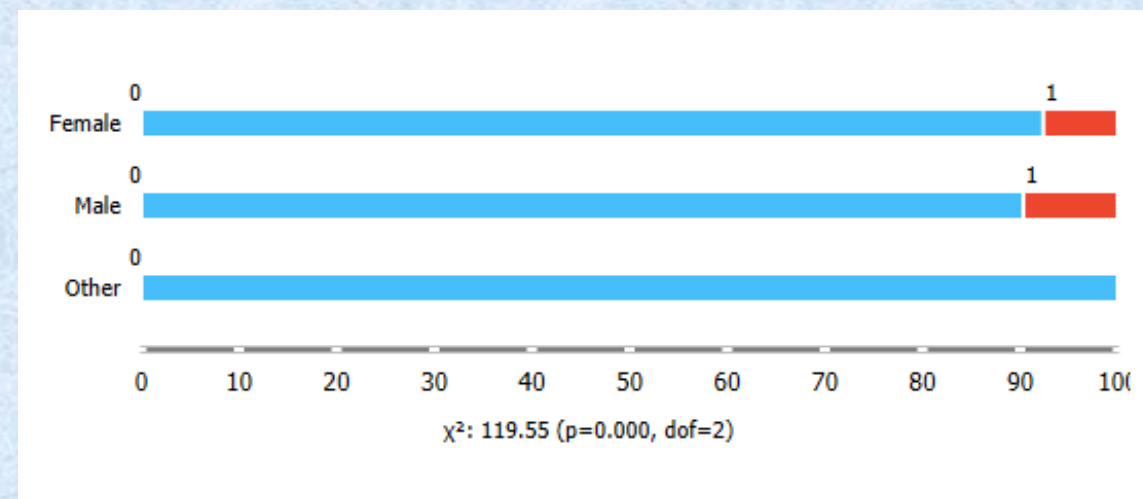
ВЛИЯНИЕ НА НИВАТА НА ГЛЮКОЗА В КРЪВТА И ГЛИКИРАН ХЕМОГЛОБИН

Чрез диаграма на разсейването проверяваме има ли връзка между двата фактора и развитието на диабет. От нея става ясно, че лицата с диабет имат по-високи стойности както на кръвната захар, така и на гликиран хемоглобин, докато тези без диабет са концентрирани при по-ниски стойности. Също така, графиката ни дава възможност да определим граничните стойности и преходната зона (около HbA1c ниво от 6 и глюкоза около 150), които също трябва да се вземат на предвид.



ПОЛЪТ КАТО ФАКТОР

Друг интересен показател е полът. Оказва се, че силният пол е по-податлив към развитие на болестта. Според данните, близо 10% от мъжете са засегнати от заболяването, докато при жените процентите са 8.



ХАРАКТЕРИСТИКИТЕ С НАЙ- ГОЛЯМО ЗНАЧЕНИЕ ЗА РАЗВИТИЕ НА ДИАБЕТ

За да проверим кои фактори играят ключова роля в развитието на диабет ще направим класификация на характеристиките. Първият метод Information Gain оценява колко информация дава всяка от характеристиките за предсказването на наличието на диабет. В нашия случай виждаме, че най- силно влияние оказва стойността на гликирания хемоглобин, следван от количеството глюкоза в кръвта. Вижда се, че най- малко влияние оказва полът.

		#	Info. gain	Gain ratio	ReliefF
1	N HbA1c_level		0.065	0.033	0.120
2	N blood_glucose_level		0.062	0.031	0.080
3	N age		0.055	0.028	0.042
4	N bmi		0.028	0.014	0.022
5	C hypertension	2	0.019	0.051	0.000
6	C smoking_history	6	0.014	0.007	0.000
7	C heart_disease	2	0.014	0.057	0.000
8	C gender	3	0.001	0.001	0.000

ХАРАКТЕРИСТИКИТЕ С НАЙ- ГОЛЯМО ЗНАЧЕНИЕ ЗА РАЗВИТИЕ НА ДИАБЕТ

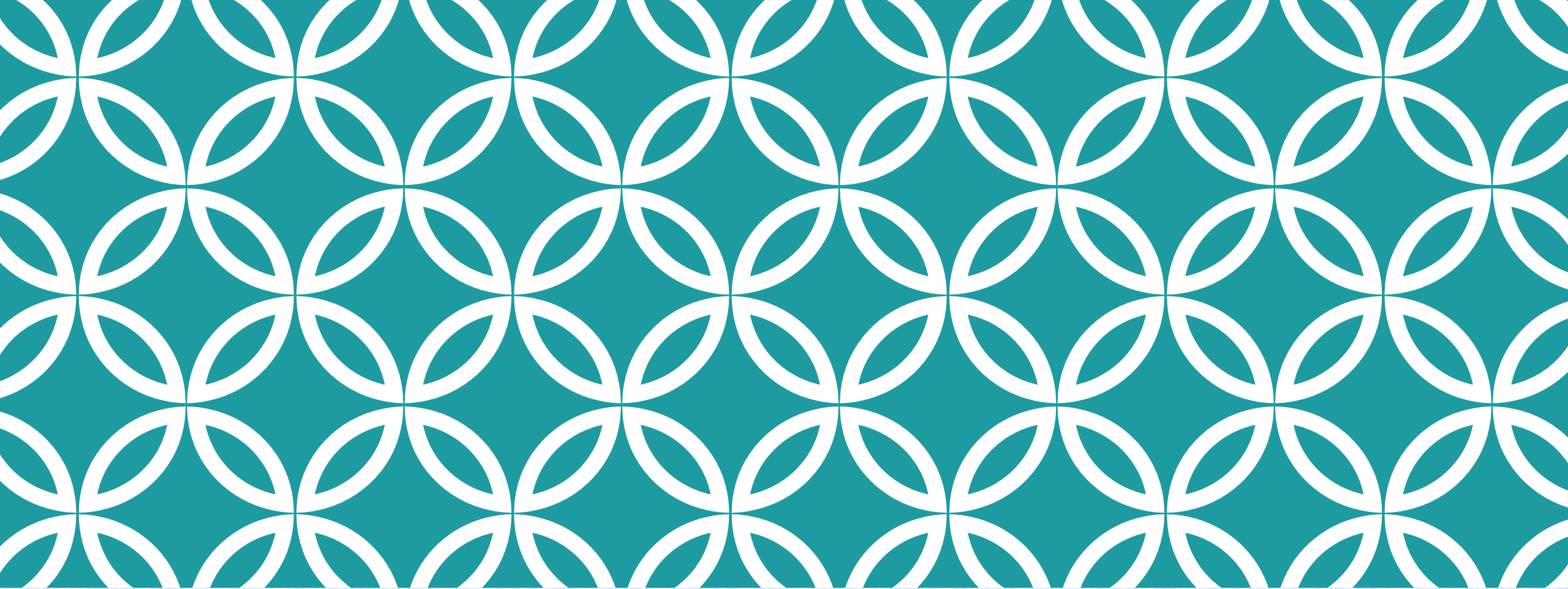
Вторият метод Gain Ratio, подобно на първия, показва каква част от информацията, която дадена характеристика носи, е полезна за класификация. Тук виждаме, че от най-голямо значение е наличието на сърдечно заболяване, следван от наличието на хипертония. Отново полът оказва най-малко влияние върху резултатите.

		#	Info. gain	Gain ratio	Relieff
1	N HbA1c_level		0.065	0.033	0.120
2	N blood_glucose_level		0.062	0.031	0.080
3	N age		0.055	0.028	0.042
4	N bmi		0.028	0.014	0.022
5	C hypertension	2	0.019	0.051	0.000
6	C smoking_history	6	0.014	0.007	0.000
7	C heart_disease	2	0.014	0.057	0.000
8	C gender	3	0.001	0.001	0.000

ХАРАКТЕРИСТИКИТЕ С НАЙ- ГОЛЯМО ЗНАЧЕНИЕ ЗА РАЗВИТИЕ НА ДИАБЕТ

Третият метод- Relief оценява всяка характеристика спрямо това, колко добре отличава примери с различни класове, като взема предвид съседни (близки) примери с еднакви и различни класове. Това е ефективен алгоритъм за работа с проблеми на класификация и може да се използва за оценка на релевантността на характеристиките спрямо целевата променлива. И тук стойността на гликирания хемоглобин има най- голямо значение.

		#	Info. gain	Gain ratio	ReliefF
1	N HbA1c_level		0.065	0.033	0.120
2	N blood_glucose_level		0.062	0.031	0.080
3	N age		0.055	0.028	0.042
4	N bmi		0.028	0.014	0.022
5	C hypertension	2	0.019	0.051	0.000
6	C smoking_history	6	0.014	0.007	0.000
7	C heart_disease	2	0.014	0.057	0.000
8	C gender	3	0.001	0.001	0.000



ОБУЧЕНИЕ НА МОДЕЛА ЗА ПРЕДСКАЗАНИЯ

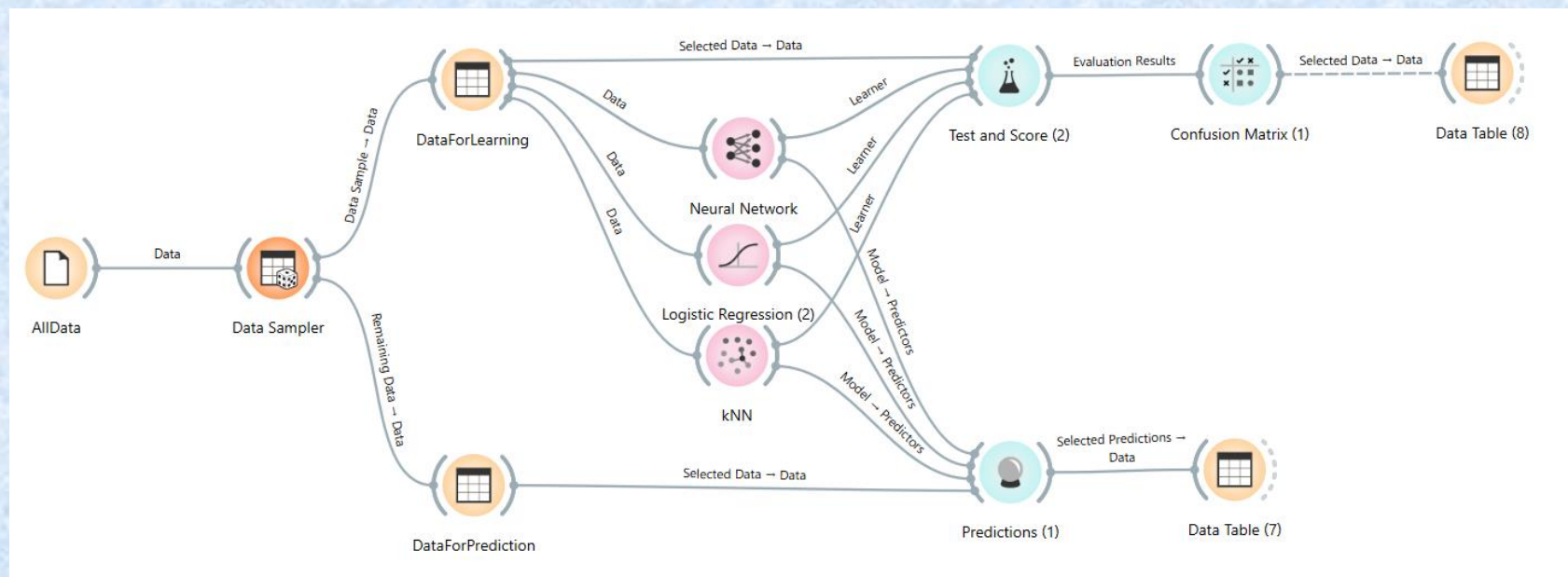


ОБУЧЕНИЕ НА МОДЕЛА ЗА ПРЕДСКАЗАНИЯ

Ще използваме три модела за предсказание, след което ще сравним резултати. В първия ще обучим невронна мрежа. Във втория подход ще използваме логистичната регресия, а в третия- k- методът на най-близките съседи (kNN). Целта на проучването е да се провери кой модел е по- точен.

ОБУЧЕНИЕ НА МОДЕЛА ЗА ПРЕДСКАЗАНИЯ

Разделяме данните на две части с помощта на компонента Data Sampler. Неговата роля е да отдели 70% от данните за обучение, а останалите- за предсказание. По-голямата част от данните се подават към моделите за обучение, а оттам- към Test & Score. Компонентът за тест от своя страна е свързан с матрица на объркването, която ни позволява да видим каква част от данните са сгрешени от модела.



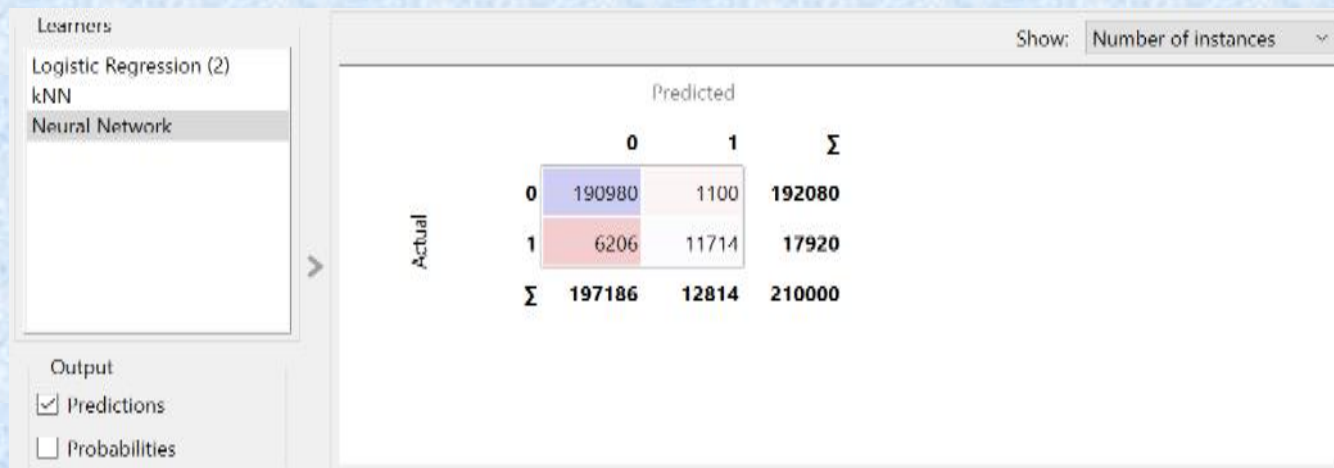
ОБУЧЕНИЕ НА МОДЕЛА ЗА ПРЕДСКАЗАНИЯ

Оказва се, че и трите модели са добре обучени- и получаваме над 95% точност. Те успяват да открият малко над 99% от общия брой заболяли (recall). Това означава, че почти всички случаи са идентифицирани. От данните на Test & Score можем да направим извода, че невронната мрежа е най- точна, а kNN е по- неефективна в сравнение с другите модели.

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression (2)	0.962	0.960	0.978	0.966	0.991	0.718
kNN	0.875	0.952	0.974	0.957	0.992	0.651
Neural Network	0.966	0.965	0.981	0.969	0.994	0.756

ОБУЧЕНИЕ НА МОДЕЛА ЗА ПРЕДСКАЗАНИЯ

От матрицата на объркването (Confusion Matrix) разбираме, че невронната мрежа правилно е определила близо 191 000 като здрави и над 11 000 като болни от диабет. Моделът е прави грешка при 1 100 здрави, които той определя като болни и малко над 6000 болни, които невронната мрежа е определила като здрави.



The screenshot shows a software interface for evaluating a model. On the left, under 'Learners', 'Neural Network' is selected. Below it, 'Output' is set to 'Predictions'. On the right, a 'Show:' dropdown is set to 'Number of Instances'. The main area displays a Confusion Matrix with 'Actual' on the y-axis and 'Predicted' on the x-axis. The matrix values are: True Positives (190980), False Positives (1100), False Negatives (6206), and True Negatives (11714). Row and column sums are also provided.

		Predicted		
		0	1	Σ
Actual	0	190980	1100	192080
	1	6206	11714	17920
Σ		197186	12814	210000

ИЗВОД

И трите модела, невронната мрежа, логистичната регресия и kNN, показват много близки стойности за ключовите метрики като точност, AUC, Recall и Precision. Това показва, че и трите подхода са ефективни в предсказването на диабет. В нашия случай може би по-добрият избор би бил моделът с невронна мрежа, понеже той има по-високи стойности на величината recall (0,994 срещу 0,992 от kNN и 0,991 от логистичната регресия). Нейната стойност е от изключително значение в медицината с оглед на това, че е важно да се минимизира пропускането на пациенти с диабет.

ЗАКЛЮЧЕНИЕ

Ранното диагностициране на заболяванията е от важно значение за тяхното навременно лечение и предотвратяване на усложнения. Такъв е и примерът с диабета. Използването на модели за предсказване, като невронни мрежи, например, позволява надеждно идентифициране на индивиди с повишен риск. Нашият модел постига висока точност с AUC от 0.969 и Recall от 0.994, което осигурява почти пълна идентификация на пациенти с диабет. Интегрирането на подобни модели в клиничната практика дава възможност на лекарите да предприемат проактивни мерки и да оптимизират здравните ресурси.

**Благодаря ви за
вниманието!**