

Введение в машинное обучение

Н.В. Артамонов

7 мая 2025 г.

Содержание

1 Введение в Python	1
1.1 Pandas	1
1.2 Визуализация	5
2 Preprocessing	8
3 Снижение размерности	8
4 Кластеризация	9
5 Регрессия	11
5.1 k-NN	11
5.2 Валидация моделей	17
6 Классификация	21
6.1 k-NN	21
6.2 Валидация моделей	22

1 Введение в Python

1.1 Pandas

#1. Загрузите датасет `countries`. Вычислите описательные статистики для каждой переменной.

#2. Загрузите датасет `sleep75`.

1. вычислите размер датасета (число наблюдений & число переменных)
2. Заполните следующую таблицу со значениями переменных

index	sleep	totwrk	age	male
0				
5				
100				
700				

3. Вычислите корреляционную матрицу для следующих переменных:
sleep, totwrk, age
4. Заполните следующую таблицу

Desc.Stat	sleep	totwrk	age	hrwage
max				
min				
mean				
median				
st.dev				
var				
1st quartile				
3rd quartile				

Замечание: 1st/3rd квантили – 25%/75% квантили соответственно.

5. Сколько наблюдения соответствуют следующим условиям
 - (a) sleep>3000
 - (b) totwrk<2000
 - (c) age>40
 - (d) age<30
6. Сколько наблюдений с условием totwrk=0? Кто эти люди?
7. Есть ли в датасете пропущенные наблюдения? Сколько их?

#3. Загрузите датасет Electricity.

1. вычислите размер датасета (число наблюдений & число переменных)
2. заполните следующую таблицу со значениями переменных

index	cost	q	pl	pk	pf
1					
15					
48					
87					

3. Вычислите корреляционную матрицу для следующих переменных:
cost, q, pl, pk, pf
4. Заполните следующую таблицу

Desc.Stat	cost	q	pl	pk	pf
max					
min					
mean					
median					
st.dev					
var					
1st quartile					
3rd quartile					

Замечание: 1st/3rd квантили – 25%/75% квантили соответственно.

5. Сколько наблюдения соответствуют следующим условиям
 - (a) $\text{cost} > 40$
 - (b) $q < 5000$
 - (c) $q > 4000$
 - (d) $20 < \text{cost} < 50$
6. Есть ли в датасете пропущенные наблюдения? Сколько их?

#4. Загрузите датасет `wage2`.

1. вычислите размер датасета (число наблюдений & число переменных)
2. заполните следующую таблицу со значениями переменных

index	wage	hours	IQ	educ	exper	age
1						
25						
179						
800						

3. Вычислите корреляционную матрицу для следующих переменных: `wage`, `hours`, `IQ`, `educ`, `exper`
4. Заполните следующую таблицу

Desc.Stat	wage	hours	IQ	educ	exper	wage
max						
min						
mean						
median						
st.dev						
var						
1st quartile						
3rd quartile						

Замечание: 1st/3rd квантили – 25%/75% квантили соответственно.

5. Сколько наблюдения соответствуют следующим условиям
 - (a) `wage > 1000`
 - (b) `age < 40`
 - (c) `exper > 10`
 - (d) `100 < IQ < 130`
6. Есть ли в датасете пропущенные наблюдения? Сколько их?

#5. Загрузите датасет **Labour**. Создайте новый датасет, содержащий log-переменные из исходного датасета.

#6. Загрузите датасет **Electricity**. Создайте новый датасет, содержащий log-переменные из исходного датасета.

1.2 Визуализация

#7. Загрузите датасет **sleep75**.

1. нарисуйте гистограммы для переменных `sleep`, `totwrk`, `age`, `hrwage`, `educ`
2. нарисуйте гистограмму с накоплением для `sleep` относительно `male`
3. нарисуйте гистограмму с накоплением для `totwrk` относительно `south`
4. нарисуйте гистограмму с накоплением для `totwrk` относительно `smsa`
5. нарисуйте диаграмму рассеяния `sleep` vs `totwrk`
6. нарисуйте диаграмму рассеяния `sleep` vs `totwrk` с группировкой по `male`
7. нарисуйте диаграмму рассеяния `sleep` vs `age`
8. нарисуйте диаграмму рассеяния `sleep` vs `age` с группировкой по `south`
9. нарисуйте диаграмму рассеяния `sleep` vs `edu`
10. нарисуйте диаграмму рассеяния `sleep` vs `edu` с группировкой по `smsa`
11. визуализируйте корреляционную матрицу для следующих переменных: `sleep`, `totwrk`, `age`

#8. Загрузите датасет **Labour**.

1. нарисуйте гистограммы для каждой переменной
2. нарисуйте гистограммы для log-переменных `output`, `capital`, `labour`, `wage`

3. нарисуйте диаграммы рассеяния output vs других переменных
4. нарисуйте диаграммы рассеяния $\log(\text{output})$ vs log других переменных
5. визуализируйте корреляционную матрицу для всех переменных
6. визуализируйте корреляционную матрицу для log-переменных

#9. Загрузите датасет **Electricity**.

1. нарисуйте гистограммы для переменных cost, q, pf, pk, pl
2. нарисуйте гистограммы для log-переменных cost, q, pf, pk, pl
3. нарисуйте диаграммы рассеяния cost vs других переменных
4. нарисуйте диаграммы рассеяния $\log(\text{cost})$ vs log других переменных
5. визуализируйте корреляционную матрицу для всех переменных
6. визуализируйте корреляционную матрицу для log-переменных

#10. Загрузите датасет **diamonds**.

1. нарисуйте гистограммы для переменных price, carat
2. нарисуйте гистограммы для log-переменных price, carat
3. нарисуйте гистограмму с накоплением для price относительно cut
4. нарисуйте гистограмму с накоплением для carat относительно clarity
5. нарисуйте гистограмму с накоплением для $\log(\text{price})$ относительно color
6. нарисуйте гистограмму с накоплением для $\log(\text{carat})$ относительно color
7. нарисуйте диаграмму рассеяния price vs carat
8. нарисуйте диаграмму рассеяния log-price vs log-carat
9. нарисуйте диаграмму рассеяния log-price vs log-carat с группировкой по cut

10. нарисуйте диаграмму рассеяния $\log(\text{price})$ vs $\log(\text{carat})$ с группировкой по `color`
11. нарисуйте диаграмму рассеяния $\log(\text{price})$ vs $\log(\text{carat})$ с группировкой по `clarity`

#11. Загрузите датасет **Diamond**.

1. нарисуйте гистограммы для переменных `price`, `carat`
2. нарисуйте гистограммы для \log -переменных `price`, `carat`
3. нарисуйте гистограмму для `price` с группировкой относительно переменной `certification`
4. нарисуйте гистограмму для `carat` с накоплением относительно `clarity`
5. нарисуйте гистограмму для $\log(\text{price})$ с накоплением относительно `colour`
6. нарисуйте гистограмму для $\log(\text{carat})$ с накоплением относительно `colour`
7. нарисуйте диаграмму рассеяния `price` vs `carat`
8. нарисуйте диаграмму рассеяния $\log(\text{price})$ vs $\log(\text{carat})$
9. нарисуйте диаграмму рассеяния $\log(\text{price})$ vs $\log(\text{carat})$ с группировкой по `certification`
10. нарисуйте диаграмму рассеяния $\log(\text{price})$ vs $\log(\text{carat})$ с группировкой по `colour`
11. нарисуйте диаграмму рассеяния $\log(\text{price})$ vs $\log(\text{carat})$ с группировкой по `clarity`

#12. Загрузите датасет **countries**.

1. Постройте гистограммы для всех переменных
2. Постройте диаграмму рассеяния Население vs ВВП д/н
3. Постройте диаграмму рассеяния ИРЧП vs ВВП д/н
4. Постройте диаграмму рассеяния Безработица vs ВВП д/н

2 Preprocessing

Замечание: рассмотрите следующие преобразования переменных:

- квантильное (для гауссового распределения)
- Box-Cox
- Yeo-Johnson

#1. Загрузите датасет **Labour**

1. Нарисуйте гистограммы для каждой переменной в уровнях и после стандартных преобразований
2. Нарисуйте диаграммы рассеяния в уровнях и после стандартных преобразований

#2. Загрузите датасет **diamonds**. Для переменных **price**, **carat**, **x**, **y**, **z**

1. Нарисуйте гистограммы для каждой переменной в уровнях и после стандартных преобразований
2. Нарисуйте диаграммы рассеяния в уровнях и после стандартных преобразований

3 Снижение размерности

#1. Загрузите датасет **Labour**.

1. Визуализируйте данные в главных компонентах (рассмотрите 2D и 3D визуализацию)
2. Визуализируйте данные, используя метод t-SNE (рассмотрите 2D и 3D визуализацию)
3. Вычислите накопленные дисперсии главных компонент.

#2. В условиях предыдущей задачи проведите визуализацию и вычислите накопленные дисперсии главных компонент после (нелинейного) преобразования данных (квантильное, Box-Cox, Yeo-Johnson)

#3. Загрузите датасет `sleep75` и удалите переменные с пропущенными значениями.

1. Визуализируйте данные в главных компонентах (рассмотрите 2D и 3D визуализацию)
2. Визуализируйте данные, используя метод t-SNE (рассмотрите 2D и 3D визуализацию)
3. Вычислите накопленные дисперсии главных компонент.

#4. В условиях предыдущей задачи проведите визуализацию и вычислите накопленные дисперсии главных компонент после (нелинейного) преобразования данных (квантильное, Box-Cox, Yeo-Johnson)

#5. Загрузите датасет `diamonds` и удалите категориальные переменные.

1. Визуализируйте данные в главных компонентах (рассмотрите 2D и 3D визуализацию)
2. Визуализируйте данные, используя метод t-SNE (рассмотрите 2D и 3D визуализацию)
3. Вычислите накопленные дисперсии главных компонент.

#6. В условиях предыдущей задачи проведите визуализацию и вычислите накопленные дисперсии главных компонент после (нелинейного) преобразования данных (квантильное, Box-Cox, Yeo-Johnson)

4 Кластеризация

Важно обязательно проводим предварительную обработку данных:

- удаление пропущенных значений
- нормировка
- преобразование категориальных признаков

#1. Для набора данных `countries` проведите разбиение на кластеры следующими методами:

Число кластеров	Метод
3	k-средних
4	k-средних
5	k-средних
3	иерархическая
4	иерархическая
5	иерархическая

Визуализируйте разбиение на кластеры на диаграмме рассеяния в переменных датасета

#2. Для набора данных `countries` найдите «оптимальное» число кластеров для метода

1. k-средних
2. иерархической кластеризации

относительно метрик: Silhouette, Calinski-Harabasz, Davies-Bouldin

#3. Из набора данных `sleep75` возьмите переменные `sleep`, `totwrk`, `age`, `educ` и проведите разбиение на кластеры следующими методами:

Число кластеров	Метод
3	k-средних
4	k-средних
5	k-средних
3	иерархическая
4	иерархическая
5	иерархическая

Визуализируйте разбиение на кластеры на диаграмме рассеяния в переменных датасета

#4. Из набора данных `sleep75` возьмите переменные `sleep`, `totwrk`, `age`, `educ` и найдите «оптимальное» число кластеров для метода

1. k-средних
2. иерархической кластеризации

относительно метрик: Silhouette, Calinski-Harabasz, Davies-Bouldin

#5. Для набора данных **Labour** проведите разбиение на кластеры следующими методами:

Число кластеров	Метод
3	k-средних
4	k-средних
5	k-средних
3	иерархическая
4	иерархическая
5	иерархическая

Визуализируйте разбиение на кластеры на диаграмме рассеяния в переменных датасета

#6. Для набора данных **Labour** найдите «оптимальное» число кластеров для метода

1. k-средних
2. иерархической кластеризации

относительно метрик: Silhouette, Calinski-Harabasz, Davies-Bouldin

5 Регрессия

5.1 k-NN

#1. Для набора данных **sleep75** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
sleep	totwrk, age, south, male

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	totwrk	age	south	male
0	2160	32	1	0
1	1720	24	0	1
2	2390	44	0	1

вычислите прогноз **sleep** по каждой модели

#2. Для набора данных **sleep75** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
sleep	totwrk, age, south, male, smsa, yngkid, marr

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	totwrk	age	south	male	smsa	yngkid	marr
0	2150	37	0	1	1	0	1
1	1950	28	1	1	0	1	0
2	2240	26	0	0	1	0	0

вычислите прогноз **sleep** по каждой модели

#3. Для набора данных **wage2** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
wage	age, IQ, south, married, urban

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	age	IQ	south	married	urban
0	36	105	1	1	1
1	29	123	0	1	0
2	25	112	1	0	1

вычислите прогноз **wage** по каждой модели

#4. Для набора данных **wage2** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
$\log(\text{wage})$	age, IQ, south, married, urban

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	age	IQ	south	married	urban
0	36	105	1	1	1
1	29	123	0	1	0
2	25	112	1	0	1

вычислите прогноз **wage** по каждой модели

#5. Для набора данных **wage1** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
wage	exper, female, married, smsa

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	exper	female	married	smsa
0	5	1	1	1
1	26	0	0	1
2	38	1	1	0

вычислите прогноз **wage** по каждой модели

#6. Для набора данных **wage1** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
$\log(\text{wage})$	exper, female, married, smsa

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	exper	female	married	smsa
0	5	1	1	1
1	26	0	0	1
2	38	1	1	0

вычислите прогноз **wage** по каждой модели

#7. Для набора данных **Labour** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
output	capital, labour

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	capital	labour
0	2.970	85
1	10.450	60
2	3.850	105

вычислите прогноз **output** по каждой модели

#8. Для набора данных **Labour** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
$\log(\text{output})$	$\log(\text{capital}), \log(\text{labour})$

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	capital	labour
0	2.970	85
1	10.450	60
2	3.850	105

вычислите прогноз **output** по каждой модели

#9. Для набора данных **Labour** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
output	capital, labour, wage

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	capital	labour	wage
0	2.970	85	36.98
1	10.450	60	33.82
2	3.850	105	40.23

вычислите прогноз **output** по каждой модели

#10. Для набора данных **Labour** рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
log(output)	log(capital), log(labour), log(wage)

1. подгоните на исходном датасете модель k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

2. Рассмотрим трёх людей с характеристиками

index	capital	labour	wage
0	2.970	85	36.98
1	10.450	60	33.82
2	3.850	105	40.23

вычислите прогноз **output** по каждой модели

5.2 Валидация моделей

#11. Набор данных **sleep75** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
sleep	totwrk, age, south, male

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#12. Набор данных **sleep75** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
sleep	totwrk, age, south, male, smsa, yngkid, marr

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#13. Набор данных **wage2** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
wage	age, IQ, south, married, urban

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#14. Набор данных **wage2** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
$\log(\text{wage})$	age, IQ, south, married, urban

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#15. Набор данных **wage1** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
wage	exper, female, married, smsa

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#16. Набор данных **wage1** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
$\log(\text{wage})$	exper, female, married, smsa

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#17. Набор данных **Labour** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
output	capital, labour, wage

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

#18. Набор данных **Labour** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
$\log(\text{output})$	$\log(\text{capital})$, $\log(\text{labour})$, $\log(\text{wage})$

и следующие модели

№	Модель
1	линейная регрессия
2	k-NN с $k = 5$, веса 'uniform'
3	k-NN с $k = 5$, веса 'distance'
4	k-NN с $k = 10$, веса 'uniform'
5	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик R^2 , MSE, MAE, MAPE. Какая модель предпочтительней?

6 Классификация

6.1 k-NN

#1. Для набора данных **sleep75** рассмотрим переменные

Зависимая/таргетная	объясняющие/признаки
male	sleep, totwrk, age, south

Рассмотрим трёх людей с характеристиками

index	sleep	totwrk	age	south
0	2900	2160	32	1
1	3120	1720	24	0
2	2850	2390	44	0

Постройте прогноз для **male** методом k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

#2. Для набора данных **sleep75** рассмотрим переменные

Зависимая/таргетная	объясняющие/признаки
smsa	sleep, totwrk, age, south, male, yngkid, marr

Рассмотрим трёх людей с характеристиками

index	sleep	totwrk	age	south	male	yngkid	marr
0	2900	2150	37	0	1	0	1
1	3120	1950	28	1	1	1	0
2	2850	2240	26	0	0	0	0

Постройте прогноз для **smsa** методом k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

#3. Для набора данных **default** рассмотрим переменные

Зависимая/таргетная	объясняющие/признаки
default	age, income, ownrent, selfempl

Рассмотрим трёх людей с характеристиками

index	age	income	ownrent	selfempl
0	37	2000	0	1
1	42.5	5250	1	0
2	29	2916	0	0

Постройте прогноз для **default** методом k-NN с параметрами

№	k	веса
1	5	uniform
2	5	distance
3	10	uniform
4	10	distance

6.2 Валидация моделей

#4. Набор данных **sleep75** разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
male	sleep, totwrk, age, south

и следующие модели

№	Модель
1	Логистическая регрессия
2	SVM
3	k-NN с $k = 5$, веса 'uniform'
4	k-NN с $k = 5$, веса 'distance'
5	k-NN с $k = 10$, веса 'uniform'
6	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик Ассигасы, ROC. Какая модель предпочтительней?

#5. Набор данных `sleep75` разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
smsa	sleep, totwrk, age, south, male, yngkid, marr

и следующие модели

№	Модель
1	Логистическая регрессия
2	SVM
3	k-NN с $k = 5$, веса 'uniform'
4	k-NN с $k = 5$, веса 'distance'
5	k-NN с $k = 10$, веса 'uniform'
6	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик Ассурасы, ROC. Какая модель предпочтительней?

#6. Набор данных `default` разбейте на обучающую и тестовую часть в соотношении 80:20.

Рассмотрим задачу прогнозирования для переменных

зависимая/target	объясняющая/предикторы/features
smsa	sleep, totwrk, age, south, male, yngkid, marr

и следующие модели

№	Модель
1	Логистическая регрессия
2	SVM
3	k-NN с $k = 5$, веса 'uniform'
4	k-NN с $k = 5$, веса 'distance'
5	k-NN с $k = 10$, веса 'uniform'
6	k-NN с $k = 10$, веса 'distance'

Проведите валидацию моделей относительно метрик Ассурасы, ROC. Какая модель предпочтительней?