Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра «Системы обработки информации и управления»

Отчет по лабораторной работе №2

«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных.»

«Курса «Технологии машинного обучения»»

Выполнила: студентка группы ИУ5-64 Светашева Ю.В

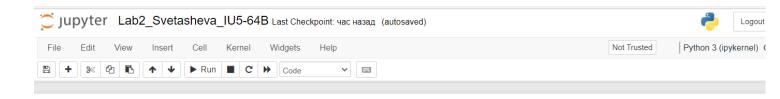
Подпись и дата:

Проверил: преподаватель каф. ИУ5 Антонов С.К. Подпись и дата:

Описание задания

- 1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
- 2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
 - о обработку пропусков в данных;
 - о кодирование категориальных признаков;
 - о масштабирование данных.

Текст программы и экранные формы



Для выполнения данной лабораторной работы будем использовать данные пассажиров Титаника

In [4]: import pandas as pd import numpy as np

Отметим, что при таком импорте DataSet необходимо разместить файл с DataSet в том же месте (в той же директории и папк e), что и проект Jupyter notebook.

In [5]: df = pd.read_csv('titanic.csv') In [5]: df.head() Out[5]

]:	Passengerld	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	0 1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
	1 2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.2833	C85	С
	2 3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
	3 4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
	4 5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Для начала удалим дубликаты, так как дублирующие записи искажают показатели DataSet. Но перед удалением дубликатов обяза тельно узнаем, сколько записей находится в DataSet.

In [6]: df.shape Out[6]: (891, 12)

Видно, что Pandas не нашел дубликатов.

In [7]: df = df.drop_duplicates() df.shape

Out[7]: (891, 12)

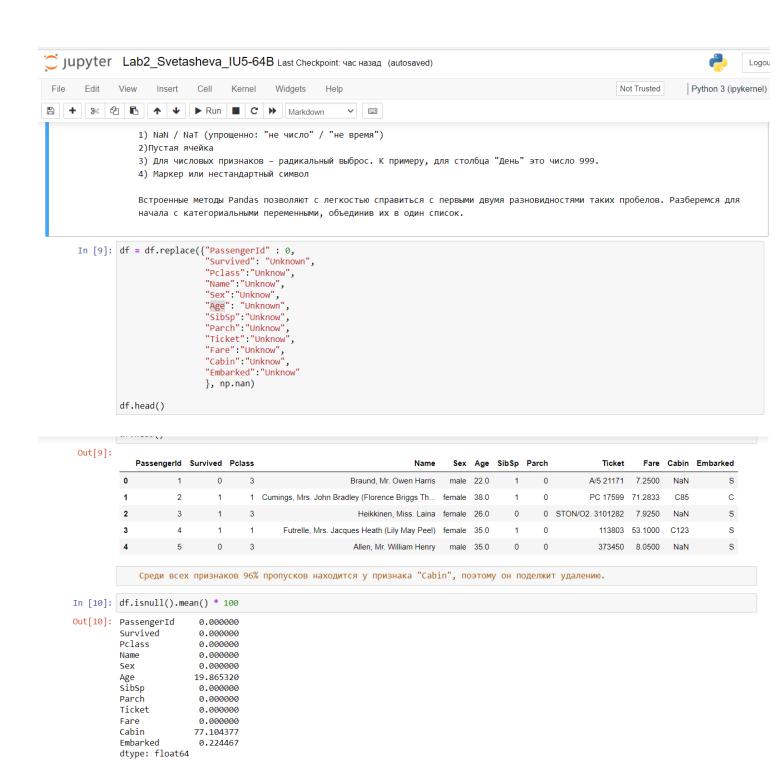
Обработка пропусков!

Помним,что если у признака более 70% пропусков, то такой признак удаляют. Поэтому проверим, насколько наши признаки полн

In [8]: column_values = df[['PassengerId','Survived','Pclass','Name','Sex','Age','SibSp','Parch','Ticket','Fare','Cabin','Embarked']].val unique_values = pd.unique(column_values) print(unique_values) 4

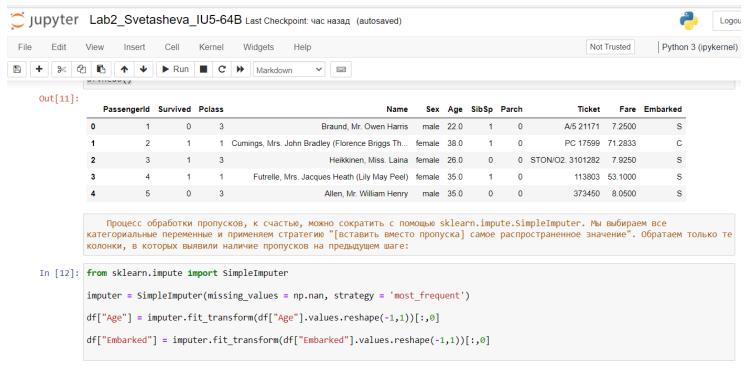
[1 0 3 ... 891 'Dooley, Mr. Patrick' '370376']

Существует несколько способов обозначить пропуски, и зачастую создатели датасета не описывают данные в достаточной мере, и определять, как обозначены пропуски, приходится вручную. Например:



In [11]: df = df.drop(columns=['Cabin'])

df.head()



Подобным образом заполняются пустоты в числовых переменных, только стратегия теперь - "вставить среднее значение".

```
In [13]: imputer = SimpleImputer(missing_values = np.nan, strategy = 'mean')

df["Age"] = imputer.fit_transform(df["Age"].values.reshape(-1,1))[:,0]
```

In [14]: df.head()

Out[14]:

	Passengerld	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.2833	С
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S