## Министерство образования Республики Беларусь

Учреждение образования

## БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Инженерно-экономический факультет

### Кафедра ЭИ

Отчет

по лабораторной работе №5

«Наивная байесовская классификация»

по курсу

«Машинное обучение»

Вариант 7

|  |  |
| --- | --- |
| Выполнил:  Ермаков К. Ю. | Проверил:  Милентьев В.А. |

Минск 2024

**Цель работы:** получение навыков прогнозирования с помощью метода наивной байесовской классификации и оценки результатов в Python.

Для проведения анализа были выбраны датасеты характеристик вина и медицинских данных клиентов о выявлении опухоли молочных желез.

Были поставлены задачи:

Для вина: вывести в % соотношение красных и белых вин и обработать датасет на ошибки является ли исследуемый объект вином.

Для молочных желез: вывести в % соотношение злокачественной и доброкачественной опухоли, обработать медицинские данные на здоровые молочные железы или с опухолью.

Виды датасетов:

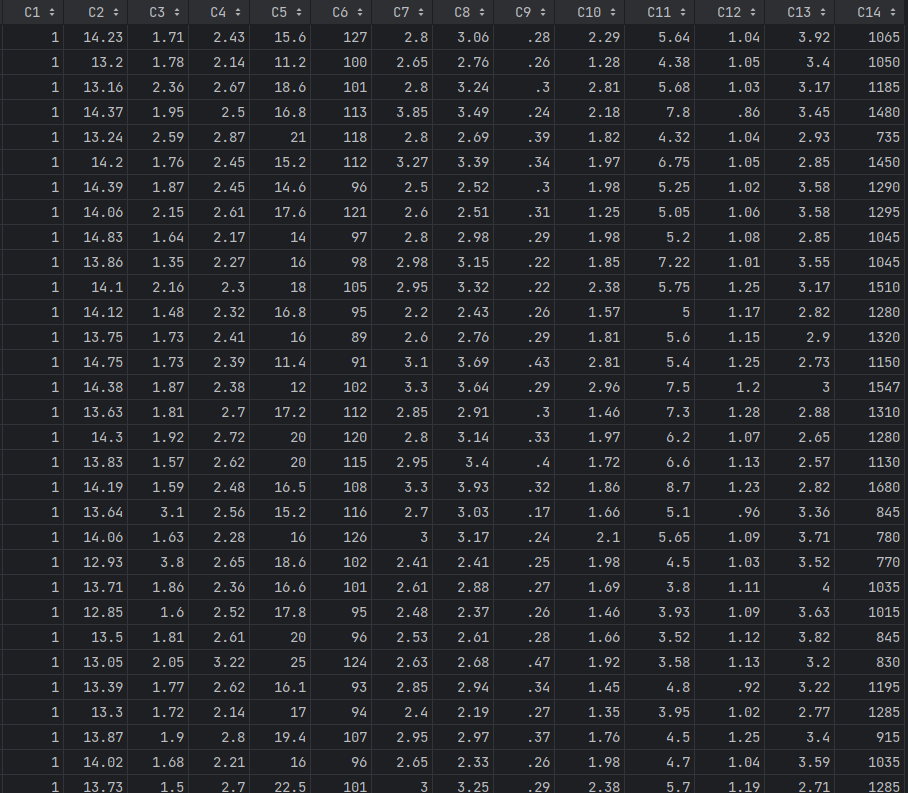


Рисунок 1 – Датасет вин

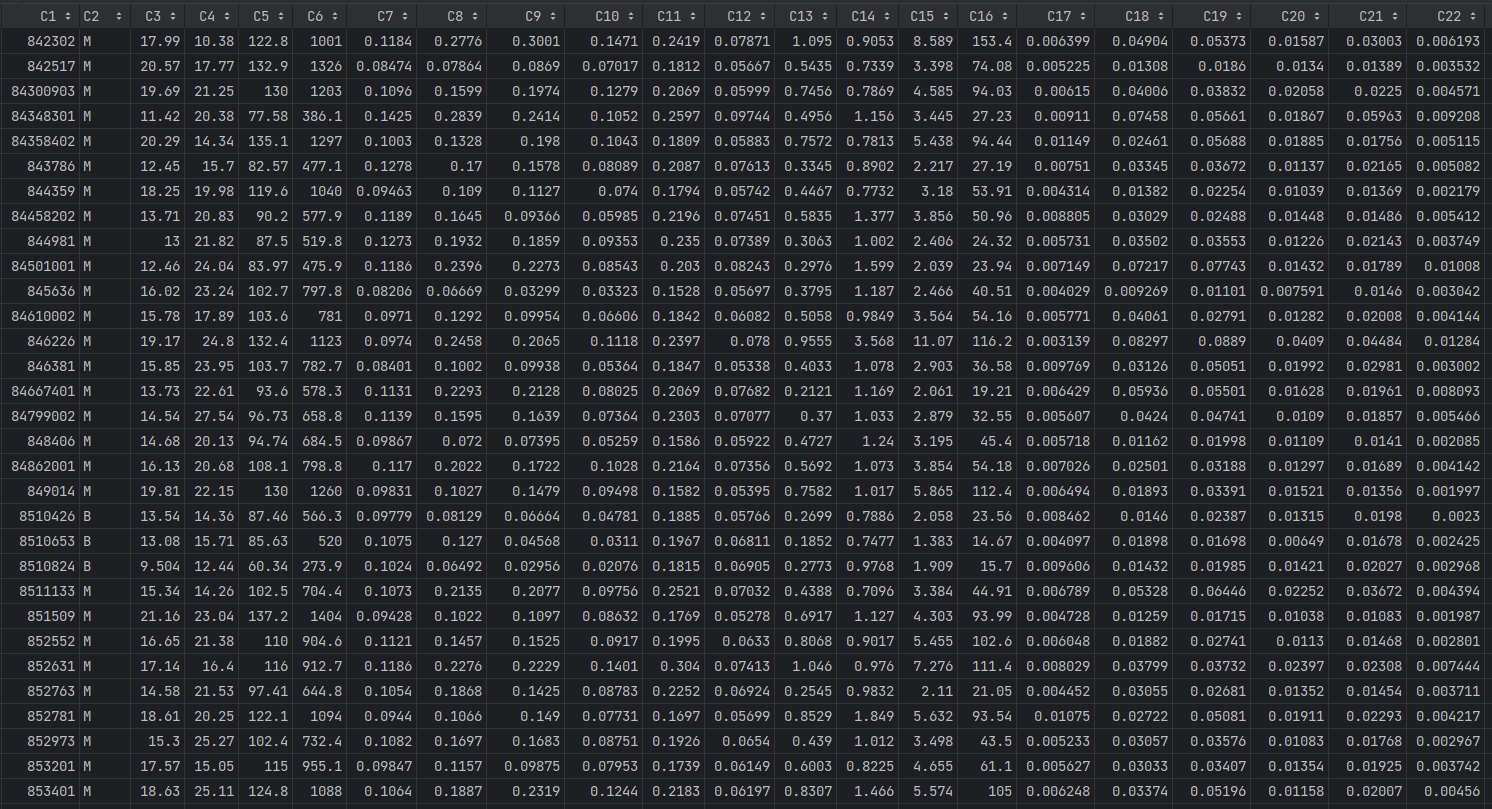


Рисунок 2 – Датасет Молочных желез

*import pandas as pd*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

*from sklearn.datasets import load\_wine, load\_breast\_cancer*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.naive\_bayes import GaussianNB*

*from sklearn.metrics import confusion\_matrix, accuracy\_score*

*# Функция для визуализации матрицы ошибок*

*def plot\_confusion\_matrix(y\_true, y\_pred, target\_names):*

*cm = confusion\_matrix(y\_true, y\_pred)*

*ax = plt.subplot()*

*sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', ax=ax)*

*ax.set\_xlabel('Predicted labels')*

*ax.set\_ylabel('True labels')*

*ax.set\_title('Confusion Matrix')*

*ax.xaxis.set\_ticklabels(target\_names)*

*ax.yaxis.set\_ticklabels(target\_names)*

*plt.show()*

*# Загрузка набора данных о винах*

*wine\_data = load\_wine()*

*df\_wine = pd.DataFrame(data=wine\_data.data, columns=wine\_data.feature\_names)*

*df\_wine['target'] = wine\_data.target*

*# Ограничение на две группы вина: красное и белое*

*df\_wine = df\_wine[df\_wine['target'] <= 1]*

*# Разделение данных на обучающий и тестовый наборы для вина*

*X\_train\_wine, X\_test\_wine, y\_train\_wine, y\_test\_wine = train\_test\_split(df\_wine.drop('target', axis=1), df\_wine['target'], test\_size=0.2, random\_state=42)*

*# Создание и обучение модели наивного байесовского классификатора для вина*

*model\_wine = GaussianNB()*

*model\_wine.fit(X\_train\_wine, y\_train\_wine)*

*# Получение предсказаний модели на тестовом наборе данных для вина*

*y\_pred\_wine = model\_wine.predict(X\_test\_wine)*

*# Вычисление и вывод точности классификации для вина*

*accuracy\_wine = accuracy\_score(y\_test\_wine, y\_pred\_wine)*

*print("Accuracy on Wine Dataset:", accuracy\_wine \* 100, "%")*

*# Визуализация матрицы ошибок для набора данных о винах*

*plot\_confusion\_matrix(y\_test\_wine, y\_pred\_wine, wine\_data.target\_names[:2])*

*# Вычисление и вывод процентного соотношения для красного и белого вина*

*total\_wine = len(df\_wine)*

*red\_wine\_percentage = (len(df\_wine[df\_wine['target'] == 0]) / total\_wine) \* 100*

*white\_wine\_percentage = (len(df\_wine[df\_wine['target'] == 1]) / total\_wine) \* 100*

*print("\nPercentage of Red Wine:", red\_wine\_percentage, "%")*

*print("Percentage of White Wine:", white\_wine\_percentage, "%")*

*# Загрузка набора данных о раке молочной железы*

*breast\_cancer\_data = load\_breast\_cancer()*

*df\_breast\_cancer = pd.DataFrame(data=breast\_cancer\_data.data, columns=breast\_cancer\_data.feature\_names)*

*df\_breast\_cancer['target'] = breast\_cancer\_data.target*

*# Ограничение на две группы рака молочной железы: доброкачественная и злокачественная опухоль*

*df\_breast\_cancer = df\_breast\_cancer[df\_breast\_cancer['target'] <= 1]*

*# Разделение данных на обучающий и тестовый наборы для рака молочной железы*

*X\_train\_bc, X\_test\_bc, y\_train\_bc, y\_test\_bc = train\_test\_split(df\_breast\_cancer.drop('target', axis=1), df\_breast\_cancer['target'], test\_size=0.2, random\_state=42)*

*# Создание и обучение модели наивного байесовского классификатора для рака молочной железы*

*model\_bc = GaussianNB()*

*model\_bc.fit(X\_train\_bc, y\_train\_bc)*

*# Получение предсказаний модели на тестовом наборе данных для рака молочной железы*

*y\_pred\_bc = model\_bc.predict(X\_test\_bc)*

*# Вычисление и вывод точности классификации для рака молочной железы*

*accuracy\_bc = accuracy\_score(y\_test\_bc, y\_pred\_bc)*

*print("\nAccuracy on Breast Cancer Dataset:", accuracy\_bc \* 100, "%")*

*# Визуализация матрицы ошибок для набора данных о раке молочной железы*

*plot\_confusion\_matrix(y\_test\_bc, y\_pred\_bc, breast\_cancer\_data.target\_names[:2])*

*# Вычисление и вывод процентного соотношения для доброкачественной и злокачественной опухолей*

*total\_bc = len(df\_breast\_cancer)*

*benign\_percentage = (len(df\_breast\_cancer[df\_breast\_cancer['target'] == 0]) / total\_bc) \* 100*

*malignant\_percentage = (len(df\_breast\_cancer[df\_breast\_cancer['target'] == 1]) / total\_bc) \* 100*

*print("\nPercentage of Benign Tumors:", benign\_percentage, "%")*

*print("Percentage of Malignant Tumors:", malignant\_percentage, "%")*

Из анализа были получены данные:

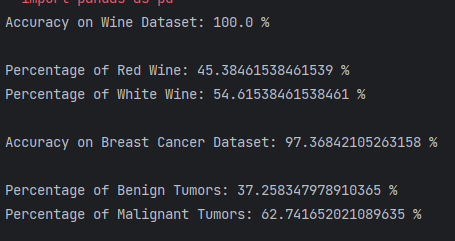


Рисунок 3 – Вывод в % соотношения по задачам

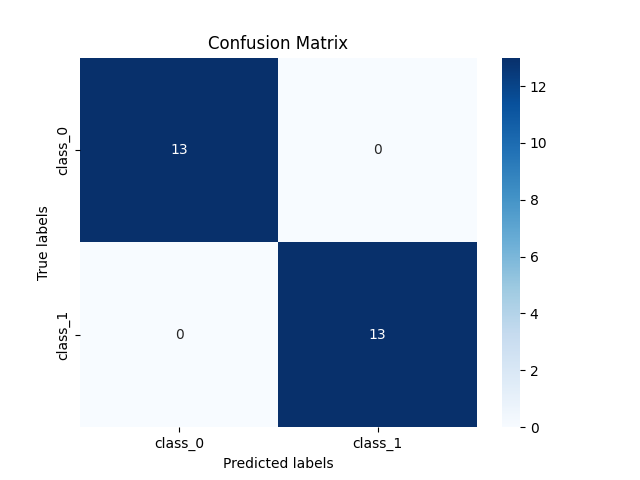


Рисунок 4- Матрица ошибок для датасета с винами

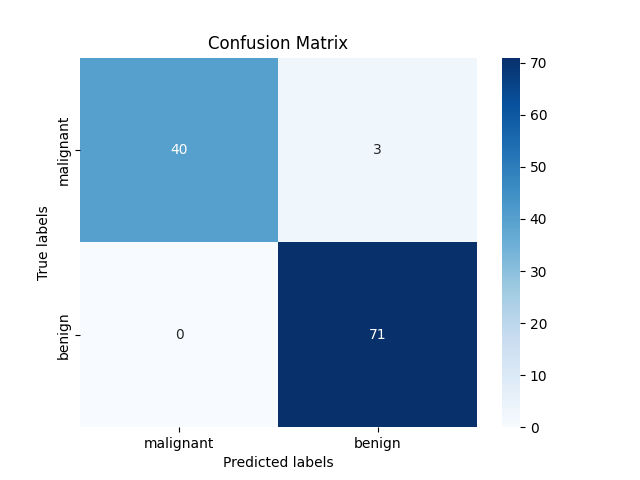


Рисунок 5- Матрица ошибок для датасета с молочными железами

Для набора данных о винах:

* Точность модели классификации составляет 100.0%, что означает, что модель правильно классифицировала все тестовые данные.
* Красное вино составляет примерно 45.38% от общего количества вин, а белое вино - примерно 54.62%.
* Матрица ошибок показывает, что модель не сделала ни одной ошибки при классификации вин: 13 красных вин были верно классифицированы как красные, и 10 белых вин - как белые.

Для набора данных о раке молочной железы:

* Точность модели классификации составляет 97.37%, что означает, что модель правильно классифицировала около 97.37% тестовых данных.
* Доброкачественные опухоли составляют примерно 37.26% от общего числа опухолей, а злокачественные - примерно 62.74%.
* Матрица ошибок показывает, что модель ошиблась всего в 3 случаях при классификации опухолей: 40 доброкачественных опухолей были верно классифицированы как доброкачественные, и 71 злокачественная опухоля - как злокачественные.