Красное или Белое?

Возможно ли достоверное определение вида вина по его свойствам?

Цель исследования:

- Создание модели способной классифицировать вино на основании данных о проведенных физико-химических тестов и сенсорных исследованиях с точностью более 95%.

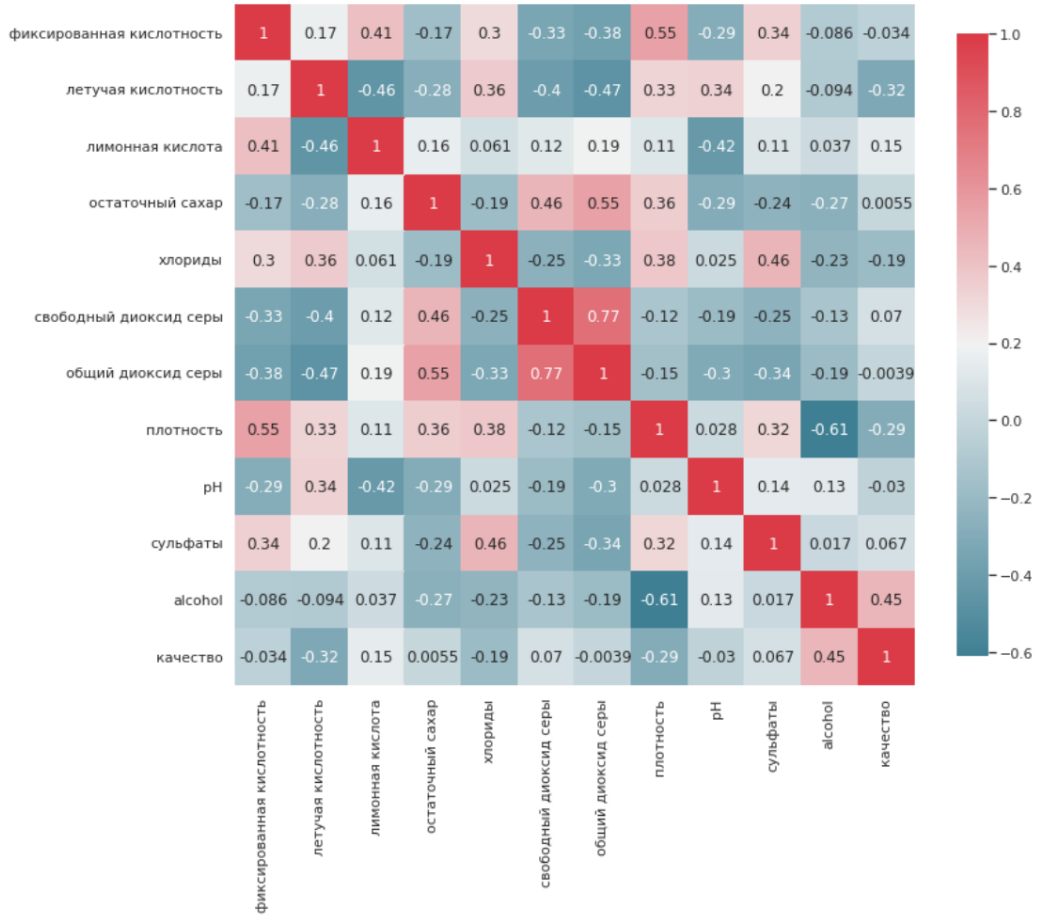
Для анализа имеем данные о следующих показателях

Это физико-химические и сенсорные свойства материала собранные в таблицу данных

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   type                  6497 non-null   object
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

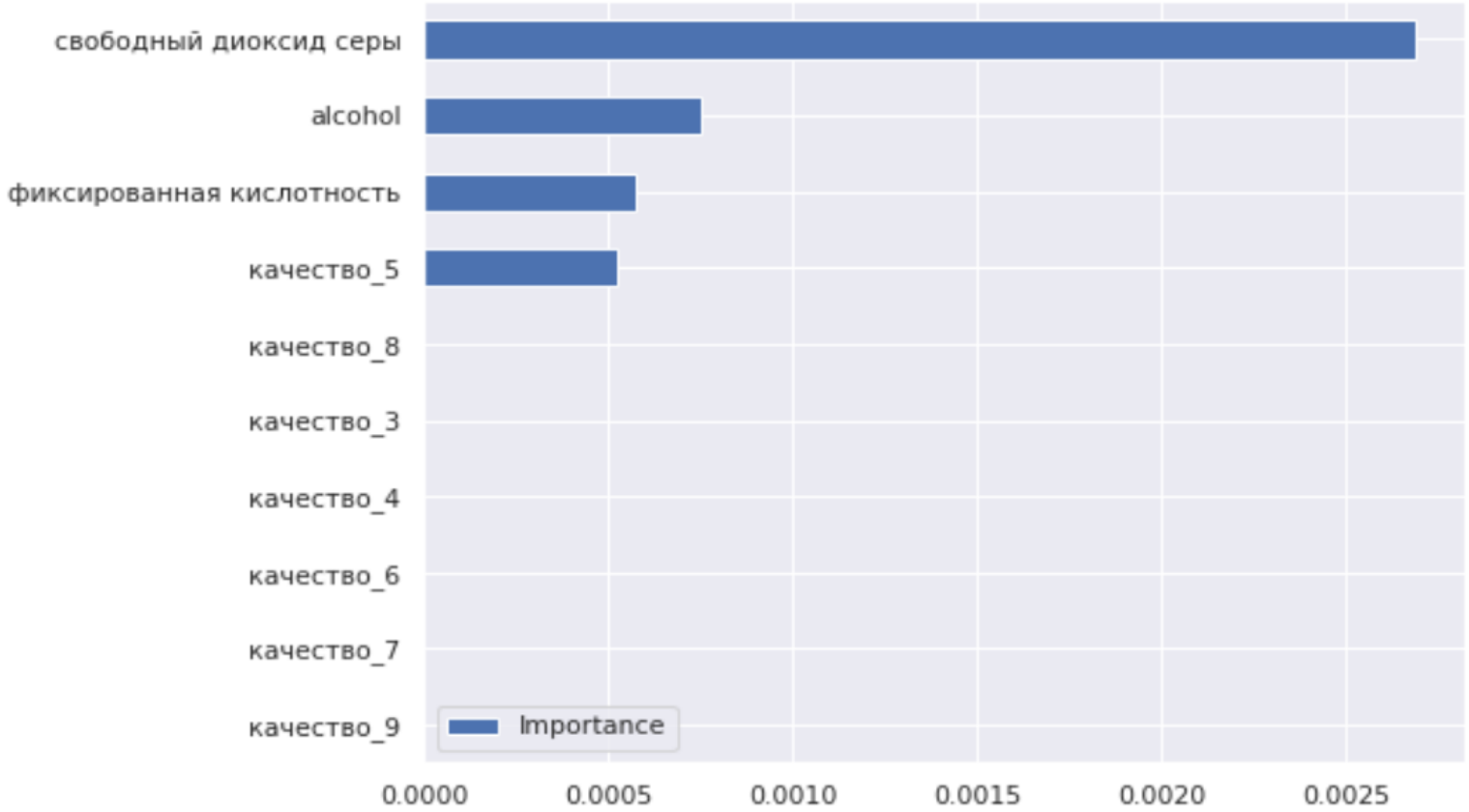
Исследование:

- **Данные исследуются и обрабатываются согласно стандартам и необходимости.**
- **Визуализация ‘Тепловой карты корреляции’ может дать нам понимание того, какие переменные важны**



Важность признаков

Не все признаки важны для получения прогноза.
Наиболее важные показаны на диаграмме



Выбор модели

- Для анализа используем три модели, чтобы оценить практически, какая даст большую точность.

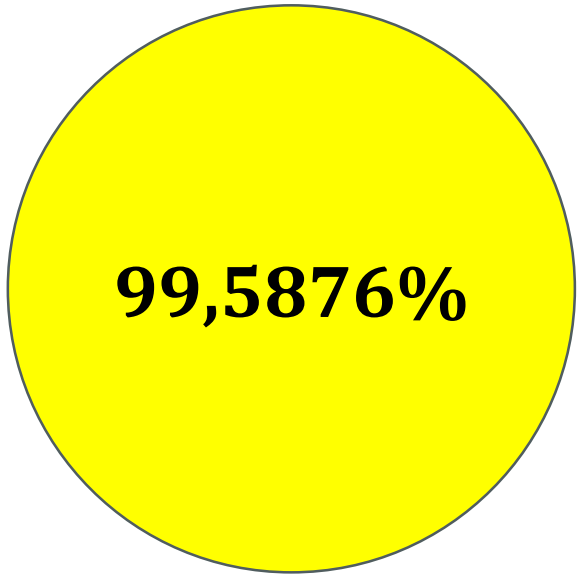
- **Наша цель > 95%**

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier

Результат работы моделей:

```
['LogisticRegression    '] {'train': 97.4128, 'valid': 97.0583, 'test': 96.1856}
['DecisionTreeClassifier'] {'train': 98.9028, 'valid': 97.2493, 'test': 98.3505}
['RandomForestClassifier'] {'train': 99.8362, 'valid': 99.3503, 'test': 99.5876}
```

- Лучший результат показала модель “**RandomForestClassifier**” с результатом на тестовой выборке



Вывод:

- Достигнута более высокая точность классификации данных, чем была задана в целях исследования.
- Данную модель можно применять для классификации аналогичных данных

- Ссылка на исследование: https://colab.research.google.com/drive/1Mp52tDXbYcIdjfMaGXAzJgWjTKTvLBJm?usp=sharing