SAVITRIBAI PHULE PUNE UNIVERSITY

The Mini Project Based On

Wine Quality Prediction

Submitted By:

Ankit Adhikrao Pawar

Seat No: B400260142

Under Guidance of:

Prof.S.K.Chougule

In partial fulfillment of

Laboratory Practice-VI(410256)

DEPARTMENT OF COMPUTER ENGINEERING)

SAVITRIBAI PHULE PUNE UNIVERSITY 2024-25

CERTIFICATE

This is to certify that the Mini Project based on,

Wine Quality Prediction

has been successfully completed by,

Name: Ankit Adhikrao Pawar

Exam seat number: B400260142

Towards the partial fulfilment of the Fourth Year of Computer Engineering as awarded by the Savitribai Phule Pune University, at PDEA's College of Engineering, Manjari Bk," Hadapsar, Pune 412307, during the academic year 2024-25.

Prof.S.K.Chougule	Dr. M. P. Borawake
Guide Name	**H.O.D**

Acknowledgement

My first and for most acknowledgment is to my supervisor and guide Prof.S.K. Chougule. During the long journey of this study, she supported me in every aspect. She was the one who helped and motivated me to proposer search in this field and inspired me with her enthusiasm on research, her experience, and her lively character.

I express true sense of gratitude to my guide Prof.S.K.Chougule. for her perfect valuable guidance, all the time support and encouragement that he gave me.

I would also like to thanks our head of department Dr. M. P. Borawake and Principal Dr. R. V. Patil and management inspiring me and providing all lab and other facilities, which made this mini project very convenient.

I thankful to all those who rendered their valuable help for successful completion on Internship presentation.

Name: Ankit Adhikrao Pawar

INDEX

Abstract

The mini-project on Business Intelligence (BI) presents a case study focusing on Wine Quality Prediction. The objective of this study is to explore the potential of BI techniques in improving healthcare outcomes by identifying factors that contribute to patient readmissions. In recent years, healthcare organizations have faced significant challenges in managing patient readmissions, which not only impact patient health but also pose financial bur dens. By leveraging BI tools and methodologies, healthcare providers can gain insights from large volumes of patient data to predict and prevent readmissions.

This mini-project analyzes a dataset comprising various patient attributes, including demographic information, medical history, and clinical factors. Using a combination of statistical analysis and machine learning algorithms, the study aims to identify patterns and correlations that influence the likelihood of patient readmission. The research methodology involves data preprocessing, feature selection, and model development. Different techniques such as logistic regression, decision trees, and ensemble methods are applied to build predictive models. Evaluation metrics, such as accuracy, precision, recall, and F1 score, are utilized to assess the performance of the models. The findings from this study can assist healthcare providers in developing targeted interventions and personalized care plans to reduce readmission rates. By understanding the key factors associated with readmission, healthcare organizations can allocate resources efficiently and implement preventive measures to improve patient outcomes and reduce healthcare costs.

Overall, this mini-project showcases the potential of BI in the healthcare sector and demonstrates how data-driven approaches can enhance decision-making processes, specifically in the context of Wine Quality Prediction. The results provide valuable insights for healthcare professionals, researchers, and policymakers striving to optimize patient care and ensure better healthcare outcomes.

Introduction

The case study presented here delves into the domain of data analytics and focuses on the critical issue of Wine Quality Prediction. Patient readmissions pose significant challenges for healthcare providers, affecting both patient well-being and healthcare costs. By harnessing the power of business intelligence techniques and ma chine learning algorithms, healthcare organizations can gain valuable insights from vast amounts of patient data and predict the likelihood of readmission.

In recent years, advancements in data analytics and machine learning have opened up new possibilities for improving healthcare outcomes. By analyzing various patient attributes, including demographic information, medical history, and clinical factors, it becomes possible to uncover patterns and correlations that contribute to readmissions. The application of predictive models can aid in identifying high-risk patients, allowing healthcare providers to intervene proactively and provide targeted interventions to pre vent unnecessary readmissions.

Overall, this case study showcases the potential of business intelligence and machine learning in the healthcare sector and demonstrates how data-driven approaches can em power healthcare providers to make informed decisions, ultimately leading to improved patient care and reduced readmission rates.

Objectives

The project “Wine Quality Prediction” aims to develop a predictive model that can accurately forecast the likelihood of patient readmission in a healthcare setting. 4 The primary focus is on identifying key factors and patterns that contribute to read missions, which will be leveraged to build a robust and accurate prediction model.

The project will involve data collection, preprocessing, feature selection/engineering, model development, and evaluation. In the data collection phase, relevant patient data will be gathered, including demographics, data history, diagrmsa, procedures, patterns, and other pertinent information. This data will be obtained from healthcare institutions or databases, ensuring a diverse representation of data. The collected data will undergo preprocessing steps to clean and handle missing values, outliers, and inconsistencies, ensuring data integrity. Next, feature selection and engineering techniques will be applied to extract mean insightful features from the dataset. This step involves identifying the most relevant features based on domain knowledge and statistical analysis, as well as creating new features to capture specific patterns or relationships. Feature selection aims to optimize the model’s performance and interpretability

Methodology

System Requirement:

- Processor: Intel Core i5 or higher (or equivalent AMD processor)
- RAM: Minimum 8GB (16GB or higher recommended for larger datasets)
- Storage: At least 250GB hard disk drive (SSD recommended for improved perfor mance)
- Operating System: Windows 10, macOS, or Linux Software: Python programming language (version 3.6 or higher)
- Development Environment: Jupyter Notebook, Anaconda, or any preferred Python IDE (Integrated Development Environment)
- Libraries: Required libraries such as Pandas, NumPy, Scikit-learn, TensorFlow, or other machine learning frameworks Internet Connection: Stable internet connection for downloading datasets, libraries, and accessing additional resources
- Graphics Processing Unit (GPU): Optional but recommended for faster training and inference in machine learning models on investment strategies and risk management.

Methodology

5.2 Algorithm

1. Start

2. Collect data regarding to project, and other relevant information.

3. Preprocess the data by handling missing values, outliers, and inconsistencies.

4. Perform feature selection and engineering to identify relevant features and create new ones if necessary.

5. Split the dataset into training and testing sets for model development and evalua tion.

6. Choose a suitable machine learning algorithm (e.g., logistic regression, decision tree, random forest, support vector machine, or neural network) for prediction.

7. Train the chosen model using the training dataset.

8. Validate the model using the testing dataset and evaluate its performance

9. If the model's performance is unsatisfactory, consider adjusting hyperparameters or trying alternative models.

10. Once a satisfactory model is obtained, use it to predict the likelihood of patient readmission for new/unseen data.

11. End

Methodology

6.1 Data Collection

Gather a comprehensive dataset containing relevant information related to patient read missions, including demographic data, medical history, clinical factors, and outcome la bels indicating readmission status.

6.2 Data Preprocessing

Cleanse and preprocess the collected data to ensure its quality and usability. Handle miss ing values, outliers, and inconsistencies. Perform data transformations, normalization, and feature scaling as necessary.

6.3 Feature Selection

Analyze the dataset to identify the most significant features that contribute to patient readmissions. Utilize statistical methods, correlation analysis, and domain knowledge to select the most relevant features for model development.

6.4 Model Development

Implement various machine learning algorithms such as logistic regression, decision trees, random forests, or support vector machines. Split the preprocessed dataset into training and testing sets. Train the models on the training set using the selected features.

6.5 Model Evaluation

Evaluate the performance of the trained models using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating char acteristic curve (AUC-ROC). Compare the performance of different models to identify the most accurate and reliable predictor of patient readmissions.

6.6 Model Optimization

Fine-tune the selected model to improve its performance. Perform hyperparameter tun ing, such as adjusting regularization parameters, tree depth, or ensemble size, using techniques like cross-validation or grid search.

Methodology

6.7 Validation and Testing

Validate the optimized model on a separate validation dataset to assess its generalization capabilities.

Evaluate the model's performance on the testing dataset to ensure its reliability and robustness.

6.8 Interpretation and Insights Analyze

the trained model to interpret the importance of features and identify factors influencing patient readmissions. Derive actionable insights and recommendations based on the model's predictions and feature contributions.

6.9 Documentation and Reporting Document

the entire methodology, including data collection procedures, preprocessing steps, model development, and evaluation processes. Present the findings, insights, and recommendations in a comprehensive report or presentation.

6.10 Iterative Improvement

Iterate through the methodology, refining and enhancing the models based on feedback and additional insights. Consider incorporating more advanced techniques, such as deep learning oensemble methods, to further improve the accuracy and predictive power of the models.

Result / Output

In [1]:
```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docke
r-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list al
l files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets p
reserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved out
side of the current session
```

```
/kaggle/input/red-wine-quality-cortez-et-al-2009/winequality-red.csv
```

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
plt.style.use('dark_background')
import seaborn as sns
```

In [3]:
```
df = pd.read_csv('../input/red-wine-quality-cortez-et-al-2009/winequality-red.cs
v')
df.head()
```

Out[3]:

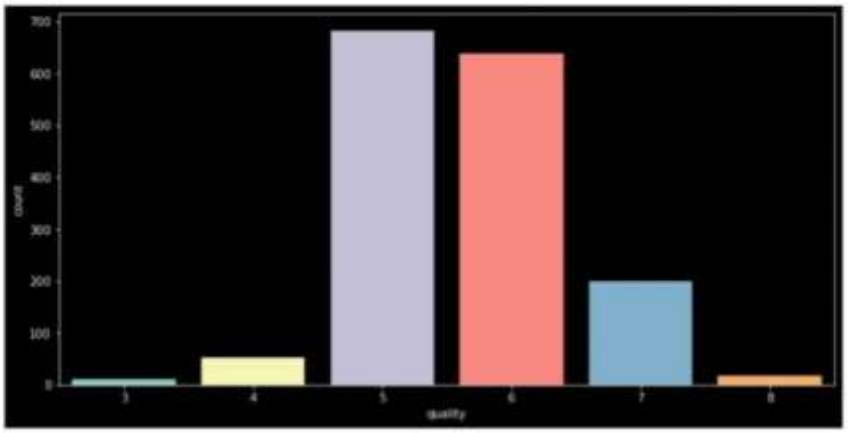
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

In [4]:
```
df.shape
```

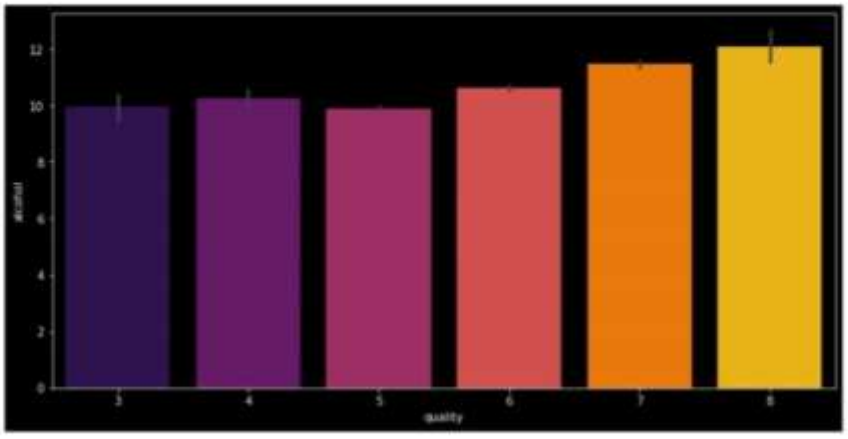
Out[4]:
```
(1599, 12)
```

Result / Output

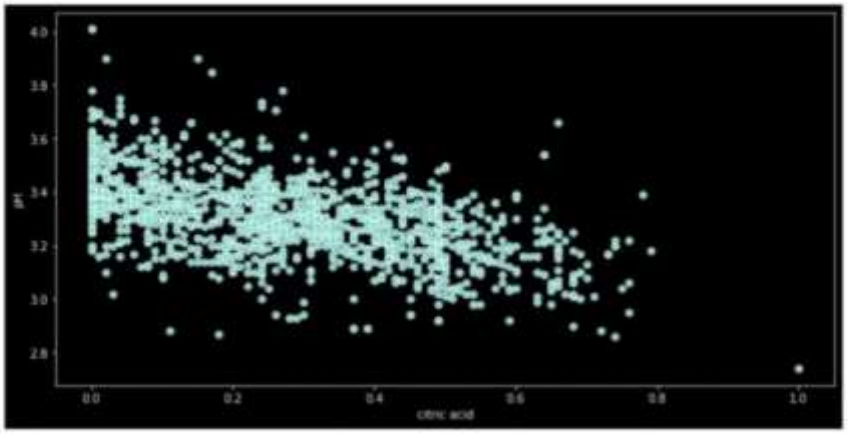
In [5]:
```
plt.figure(figsize = (12,6))
sns.countplot(df['quality'])
plt.show()
```



In [6]:
```
plt.figure(figsize = (12,6))
sns.barplot(x='quality', y = 'alcohol', data = df, palette = 'inferno')
plt.show()
```



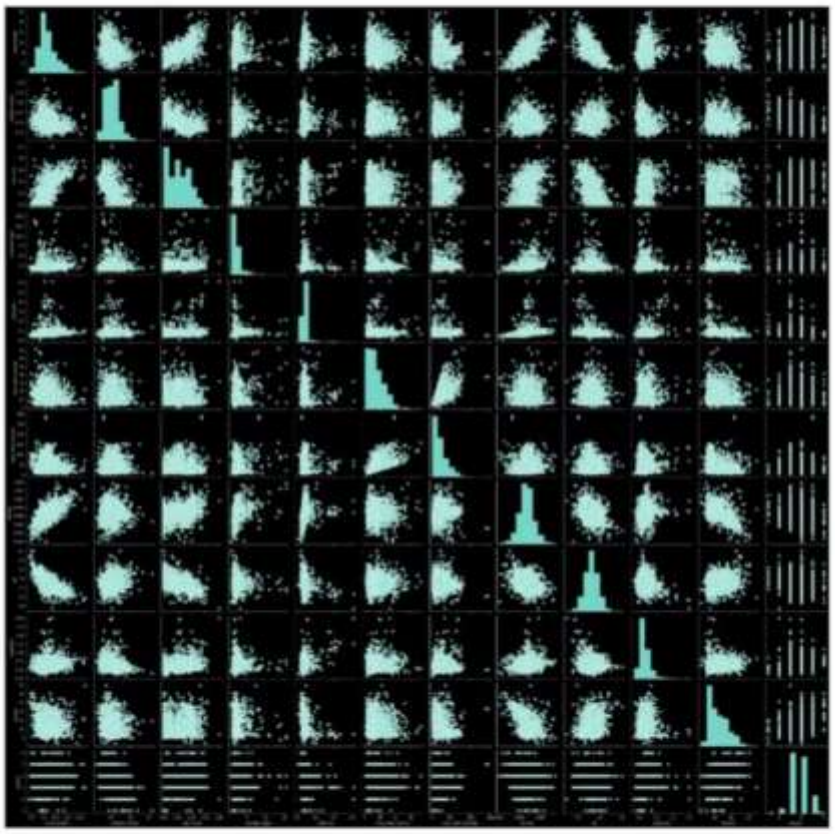
In [7]:
```
plt.figure(figsize = (12,6))
sns.scatterplot(x='citric acid', y = 'pH', data = df)
plt.show()
```



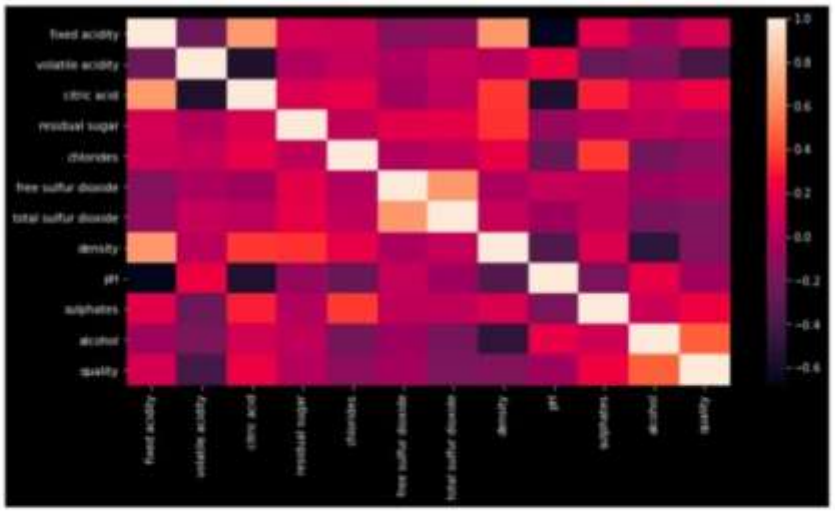
Result / Output

In [8]:
```
plt.figure(figsize = (12,6))
sns.pairplot(df)
plt.show()
```

```
<Figure size 864x432 with 0 Axes>
```



In [9]:
```
plt.figure(figsize = (12,6))
sns.heatmap(df.corr())
plt.show()
```



In [10]:
```
x=df.drop(['quality'], axis=1)
y=df['quality']
```

Result / Output

In [11]:
```
## oversampling
from imblearn.over_sampling import SMOTE
os=SMOTE()
x_res,y_res=os.fit_sample(x, y)
```

In [12]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_res,y_res,test_size=0.2, r
andom_state=0)
```

In [13]:
```
from sklearn.preprocessing import StandardScaler

stdscale = StandardScaler().fit(x_train)
x_train_std = stdscale.transform(x_train)
x_test_std = stdscale.transform(x_test)
```

In [14]:
```
from sklearn.metrics import accuracy_score
```

Logistic Regression

In [15]:
```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(x_train_std, y_train)
predictions = lr.predict(x_test_std)
accuracy_score(y_test, predictions)
```

Out[15]:
```
0.5647921760391198
```

Decision Tree Classifier

In [16]:
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier()
dt.fit(x_train_std, y_train)
accuracy_score(y_test, dt.predict(x_test_std))
```

Out[16]:
```
0.7750611246943765
```

Random Forest Classifier

In [17]:
```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(random_state = 42)
rf.fit(x_train_std, y_train)
accuracy_score(y_test, rf.predict(x_test_std))
```

Out[17]:
```
0.8533007334963325
```

Future Scope

"Wine Quality Prediction" aims to develop a predictive model that can accurately forecast the likelihood of setting. 4 The primary focus is on identifying key factors and patterns that contribute to read missions, which will be leveraged to build a robust and accurate prediction model.

The project will involve data collection, preprocessing, feature selection/engineering, model development, and evaluation. In the data collection phase, relevant patient data will be gathered, including demographics ,data history, diagrams, procedures, patterns, and other pertinent information. This data will be obtained from analysis institu tions or databases, ensuring a diverse representation of patients. The collected data will undergo preprocessing steps to clean and handle missing values, outliers, and inconsis tencies, ensuring data integrity.

Next, feature selection and engineering techniques will be applied to extract mean insightful features from the dataset. This step involves identifying the most relevant features based on domain knowledge and statistical analysis, as well as creating new features to capture specific patterns or relationships. Feature selection aims to optimize the model's performance and interpretability.

Conclusion

In conclusion, Wine Quality Prediction is a valuable approach that leverages data and machine learning techniques to estimate the likelihood of a patient being readmitted to a healthcare facility after initial discharge. By accurately predicting patient readmission, healthcare providers can proactively intervene and allocate resources effectively to improve patient outcomes and reduce readmission rates.

Through the collection and preprocessing of relevant patient data, including demo graphics, data history, diagrams, procedures, patterns, and other pertinent information., a comprehensive understanding of factors contributing to readmission can be obtained. Feature selection and engineering techniques further enhance the predictive model's ability to identify critical factors and patterns associated with readmission.

Reference

- https://www.github.com
- https://chat.openai.com/
- https://youtube.com/
- https://youtu.be/3681ZYbDSSk