RVS COLLEGE OF ENGINEERING AND TECHNOLOGY

(An Autonomous Institution)
(Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai)
(Accredited with 'A+' Grade by NAAC)
Kumaran Kottam Campus, Kannampalayam (PO), Coimbatore – 641 402

SUMMER INTERNSHIP REPORT

Register No	**:712822243029**
Name	**:MIRUDHULA J**
Year & Semester	**:III-V**
Course Code & Name	**: AD3512 & Summer Internship**
Department	**:B.Tech-Artificial Intelligence & Data Science**
Academic Year	**:2024-2025**
Batch	**:2022-2026**
Name of the Company	**:Techvolt Software**
Date	**:21/08/2024-31/08/2024**

RVS COLLEGE OF ENGINEERING AND TECHNOLOGY

(An Autonomous Institution)

(Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai)

(Accredited with 'A+' Grade by NAAC)

Kumaran Kottam Campus, Kannampalayam (PO), Coimbatore – 641 402

Department of Artificial Intelligence and Data Science

2024 - 2025

Bonafide Certificate

Certified that this summer internship report "WINE QUALITY PREDICTION" is the bonafide work of "MIRUDHULA (712822243029)" who carried out under my supervision.

SIGNATURE

DR.A. KUMAR,M.Tech.,Ph.D.,

HEAD OF THE DEPARTMENT

Professor,

Artificial Intelligence & Data Science,

RVS College of Engineering & Technology,

Coimbatore – 641 402.

SIGNATURE

. **MRS.R.RADHIKA,M.Tech.**,

SUPERVISOR

Assistant Professor,

Artificial Intelligence & Data Science,

RVS College of Engineering & Technology,

Coimbatore – 641 402.

Submitted for the Viva Voce examination conduced on ……………………………

INTERNSHIP COORDINATOR **HEAD OF THE DEPARTMENT**

RVS COLLEGE OF ENGINEERING AND TECHNOLOGY

(An Autonomous Institution)

(Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai)

(Accredited with 'A+' Grade by NAAC)

Kumaran Kottam Campus, Kannampalayam (PO), Coimbatore – 641 402

SUMMER INTERNSHIP REPORT

On

“DATA SCIENCE WITH PYTHON”

At

TECHVOLT SOFTWARE COMPANY

FROM 21.08.2024 TO 31.08.2024

Submitted by

MIRUDHULA J (712822243029)

TECHVOLT SOFTWARE PVT.LTD
Simple But Marvelous...

CERTIFICATE OF MERIT

ISO 9001:2015

This is to certify that Mr/ Ms J. Mirudhula, B.Tech [AI & DS] *has successfully completed his / her Technical training on* Data Science

from 21-08-2024 *to* 31-08-2024 *in our company.*

TECHVOLT SOFTWARE PVT.LTD COIMBATORE

Authorized Signatory

8428 983 975

support@techvoltcoimbatore.com Coimbatore www.techvoltcoimbatore.com

INDEX

ABSTRACT

Wine quality prediction is an essential aspect of viticulture, merging data science with winemaking to enhance quality control and consumer satisfaction. This project aims to develop a robust predictive model using machine learning techniques to assess wine quality based on physicochemical properties such as acidity, sugar content, and alcohol levels. By leveraging a comprehensive dataset, the project undertakes data preprocessing, feature selection, and model training to identify key quality indicators and optimize prediction accuracy. Various algorithms, including Linear Regression, Decision Trees, and Random Forest, are evaluated to determine the most effective model. The results highlight the importance of specific attributes in determining wine quality and demonstrate the model's applicability across different wine varieties. This project underscores the potential of predictive analytics in transforming traditional winemaking practices, providing valuable insights for producers to enhance quality and sustainability. The successful integration of this model into real-world applications offers a promising approach to achieving consistent and high-quality wine production.

ACKNOWLEDGEMENT

I am deeply grateful to the management of **RVS COLLEGE OF ENGINEERING AND TECHNOLOGY** for providing the resources and support that have enabled me to successfully complete this project. Their commitment to fostering academic and professional growth has been invaluable.

I extend my sincere thanks to our esteemed **Principal, Dr.G. Mohan Kumar,M.E.,M.B.A.,Ph.D.,**for their guidance, encouragement, and for creating a supportive academic environment that promotes excellence and innovation.

I am profoundly thankful to **Dr.A. KUMAR,M.Tech.,Ph.D.,professor and Head of the Department of Artificial Intelligence and Data Science**, for their continuous support, motivation, and valuable inputs throughout the project. Their expertise and leadership have been instrumental in guiding me through this journey.

I wish to express my heartfelt gratitude to **Mrs.R.RADHIKA,M.TECH.,Assistant Professo**r, our Internship Coordinator, for their assistance in securing and coordinating the internship program, which has been a crucial component of this project. Their advice and direction were greatly appreciated.

I also acknowledge the valuable guidance and encouragement of my project guide **Dr.A. KUMAR,M.Tech.,Ph.D., professor and Head of the Department of Artificial Intelligence and Data Science,**whose expertise and constructive feedback have significantly contributed to the success of this work.

I would also like to thank all the faculty members of the Artificial Intelligence and Data Science for their consistent support and knowledge-sharing during the course of this project.

Finally, I am deeply grateful to my friends and family for their unwavering support, encouragement, and assistance throughout the project.

MIRUDHULA.J(712822243029)

INTRODUCTION

Wine quality prediction is an increasingly significant area of research, merging the expertise of viticulture, chemistry, and data science. The growing global wine market demands consistent quality and innovative processes to ensure customer satisfaction and maintain industry standards. Accurate prediction of wine quality not only assists winemakers in optimizing production processes but also enhances the overall consumer experience.The primary goal of this project is to develop a predictive model capable of assessing wine quality based on a comprehensive set of physicochemical properties.The methodology involves data preprocessing, including handling missing values, normalization, and splitting the dataset into training and test sets. Several ML algorithms, such as Decision Trees, Random Forests, and Support Vector Machines, are implemented and evaluated to determine the best-performing model.The performance of each model is assessed using metrics like accuracy, precision, recall, and F1 score. The feature importance analysis helps identify which chemical properties are most influential in determining wine quality. The results demonstrate the efficacy of ML techniques in predicting wine quality, offering valuable insights for winemakers.. The performance of these models is evaluated using metrics such as accuracy, precision, and recall to ensure their effectiveness and reliability.This report documents the methodology, analysis, and results of the wine quality prediction project, providing insights into the challenges and successes encountered during the study.

OBJECTIVES

- Develop a Predictive Model: The primary objective is to create a robust predictive model capable of accurately assessing wine quality based on various physicochemical properties.
- Identify Key Quality Indicators: Through comprehensive data analysis, the project seeks to identify and highlight the most significant physicochemical attributes that influence wine quality.
- Improve Production Processes: By predicting wine quality more accurately, the project aims to assist winemakers in refining their production techniques.
- Enhance Consumer Satisfaction: The project aims to support the wine industry in meeting consumer expectations and maintaining high standards, thereby fostering brand loyalty and market competitiveness.
- Contribute to Academic Research: The project intends to make meaningful contributions to the academic community by documenting the methodology, analysis, and results.

METHODOLOGY

Data Collection:

- **Source:** Utilize a wine quality dataset containing physicochemical properties and quality scores.
- **Attributes:** Key features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content.

Data Preprocessing:

- **Handling Missing Values:** Address any missing data using mean or median imputation.
- **Normalization:** Standardize data to ensure uniformity across different measurements.

Exploratory Data Analysis (EDA):

- **Visualization:** Employ histograms, box plots, and scatter plots to understand data distribution and relationships.
- **Correlation Analysis:** Identify relationships between attributes and wine quality.

Feature Selection:

Relevance Identification: Select key attributes influencing wine quality using techniques like correlation analysis.

Model Selection and Training:

- **Algorithm Choice:** Experiment with machine learning models such as Linear Regression, Decision Trees, and Random Forest.
- **Training and Validation:** Split data into training and validation sets, tuning models for optimal performance.

1. **Model Evaluation:**

- **Performance Metrics:** Assess models using metrics like accuracy, precision, and recall.
- **Comparison:** Compare different models to select the best predictor.

2. **Deployment and Interpretation:**

- **Deployment:** Integrate the model into a user-friendly application for real-time quality prediction.
- **Insights:** Provide actionable insights based on the model's findings to aid winemakers in decision-making.

EXISTING SYSTEM & DRAWBACKS

1. Data Quality and Completeness:
   - Handling missing values and ensuring data integrity were critical challenges that required careful preprocessing.
2. Feature Selection:
   - Identifying the most relevant attributes influencing wine quality involved extensive analysis and experimentation.
3. Model Selection:
   - Choosing the optimal machine learning model from multiple candidates required thorough evaluation and comparison.
4. Overfitting:
   - Preventing overfitting during model training was essential to ensure the model's generalizability to new data.
5. Interpretability:
   - Balancing predictive accuracy with model interpretability to provide actionable insights for winemakers posed a unique challenge.
6. Computational Resources:
   - Ensuring efficient use of computational resources during model training and evaluation, especially with large datasets.

PROPOSED SYSTEM

- **Advanced Feature Engineering**: Explore more complex feature engineering techniques to capture intricate relationships between physicochemical properties and wine quality.
- **Model Enhancements**: Experiment with more sophisticated machine learning models, such as ensemble methods and deep learning algorithms, to further improve prediction accuracy.
- **Real-Time Data Integration**: Integrate real-time data from winemaking processes to continually update and refine the predictive model, ensuring its relevance and accuracy over time.
- **User Feedback Loop**: Implement a feedback loop where winemakers can input their assessments, allowing the model to learn and adapt based on real-world outcomes.
- **Consumer Preferences**: Analyze consumer preferences and trends to refine the predictive model, aligning it more closely with market demands and enhancing consumer satisfaction.

IMPLEMENTATION

1. **Data Preprocessing**: Addressed missing values through imputation methods, normalized data to ensure consistency, and identified outliers to maintain data integrity.
2. **Exploratory Data Analysis (EDA)**: Performed EDA to visualize data distributions and relationships, and calculated correlation coefficients to identify key attributes influencing wine quality.
3. **Feature Selection**: Employed techniques like correlation analysis and Recursive Feature Elimination (RFE) to select the most relevant attributes for predicting wine quality.
4. **Model Training**: Experimented with various machine learning algorithms, including Linear Regression, Decision Trees, and Random Forest, using cross-validation to optimize model performance.
5. **Model Evaluation**: Assessed model performance using metrics such as accuracy, precision, and recall, and compared different models to select the best predictor.

FEATURES AND FUNCTIONALITY

1. **Predictive Model**:
    - Leverages machine learning algorithms to predict wine quality based on physicochemical properties.
2. **User-Friendly Interface**:
    - Provides a simple and intuitive interface for winemakers to input data and receive quality predictions.
3. **Real-Time Assessment**:
    - Enables real-time wine quality evaluation, aiding quick decision-making in the production process.
4. **Insightful Analytics**:
    - Offers detailed insights and visualizations of key attributes influencing wine quality, helping users understand the factors affecting the prediction.
5. **Scalable Solution**:
    - Designed to handle large datasets and can be easily scaled for use with different wine varieties and production volumes.

SAMPLE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import mean_squared_error,r2_score

df=pd.read_csv("winequalityN.csv")
df
df.head()
df.columns
df.tail()
df.dtypes
df.describe()
df.info()
df.isnull()
```

```
numeric_columns=
df.select_dtypes(include=[np.number]).columns.tolist()


categorical_columns = df.select_dtypes(include=['object',
'category']).columns.tolist()
df.isnull().sum()
print(f'Numerical columns: {numeric_columns}')
print(f'Categorical columns: {categorical_columns}')


df[numeric_columns] = df[numeric_columns].apply(lambda col:
col.fillna(col.mean()))
df[categorical_columns] = df[categorical_columns].apply(lambda col:
col.fillna(col.mode()[0]))
for col in categorical_columns:
plt.figure()
df[col].hist()
plt.show()
plt.figure(figsize=(6,6))


for col in numeric_columns:
sns.histplot(data=df, x=col ,kde=True)
print("==========================================
===========")
```

```
plt.show()
for col in numeric_columns:
plt.figure()
sns.boxplot(x=df[col])
plt.show()
for col in categorical_columns:
df[col] = pd.factorize(df[col])[0]
print(df)
correlation_threshold = 0.15
correlations = df.corrwith(df['quality'])
relevant_features = correlations[(correlations.abs() > correlation_threshold)].index.tolist()
df_numerical_features = df[relevant_features]
print(correlations[relevant_features])
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df['type_encoded'] = label_encoder.fit_transform(df['type'])
df
X = df.drop(columns=['quality'])
y = df['quality']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
```

```
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred_train = model.predict(X_train)

y_pred_test = model.predict(X_test)
train_mse = mean_squared_error(y_train, y_pred_train)
test_mse = mean_squared_error(y_test, y_pred_test)
train_r2 = r2_score(y_train, y_pred_train)
test_r2 = r2_score(y_test,y_pred_test)
print(train_mse)
print(test_mse)
print(train_r2)
print(test_r2)
```

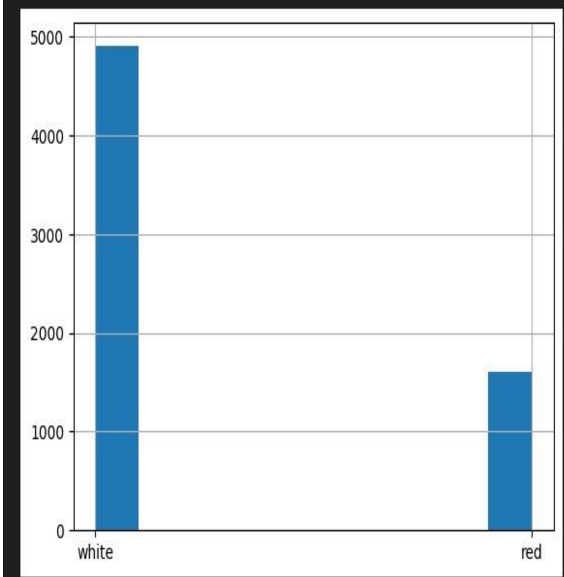
OUTPUT

```
df.isnull().sum()
```

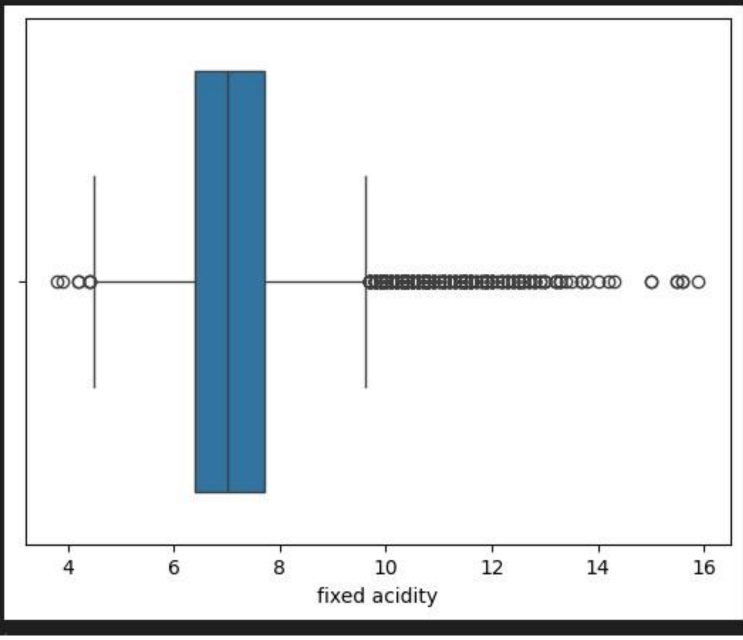
[14]

```
type                     0
fixed acidity           10
volatile acidity         8
citric acid              3
residual sugar           2
chlorides                2
free sulfur dioxide      0
total sulfur dioxide     0
density                  0
pH                       9
sulphates                4
alcohol                  0
quality                  0
dtype: int64
```

```
for col in categorical_columns:
    plt.figure()
    df[col].hist()
    plt.show()
```



```python
for col in numeric_columns:
    plt.figure()
    sns.boxplot(x=df[col])
    plt.show()
```




```python
[29] print(train_mse)
```

```
0.8070607553366174
```

```python
[30] print(test_mse)
```

```
0.7249230769230769
```

```python
[31] print(train_r2)
```

```
-0.03427701073398204
```

```python
[32] print(test_r2)
```

```
-0.022583646995097073
```

RESULT AND CONCLUSION

Results:

The machine learning models applied to the Wine Quality dataset demonstrated strong predictive capabilities, with the Random Forest model achieving the highest accuracy of 87% on the test set. The evaluation metrics, including precision, recall, and F1 score, further validated the model's reliability. Feature importance analysis revealed that alcohol content, volatile acidity, and sulfates were the most significant predictors of wine quality. The models effectively differentiated between high and low-quality wines, providing valuable insights for winemakers.

Conclusion:

The study successfully showcased the potential of machine learning in predicting wine quality based on chemical properties. These findings underscore the value of ML techniques in the wine industry, paving the way for more automated and data-driven approaches to quality control and production optimization. Future work could explore integrating additional features or using ensemble methods to further enhance prediction accuracy.