World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Acidity calibre check in wine using machine learning

Raghupathi Kanala, Meghana Vijaya Raghavan, Jathin Kumar Gundala *, Sushruth Bommagoni and Arun Kumar Onteddu

Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Research and Reviews, 2025, 25(02), 213-221

Publication history: Received on 25 December 2024; revised on 01 February 2025; accepted on 04 February 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.2.0377

Abstract

The wine quality will help the winemakers to make decisions during production such as adjusting fermentation techniques or blending ratios to improve the quality. This in turn helps them to make profit while the wine seekers buy a top quality product.

The acidity calibre check project aims to develop a machine learning model to predict the acid quantity in wines based on various physicochemical features. The project involves data preprocessing, exploratory data analysis, feature selection, model training, and evaluation. To address the potential case of class imbalance in the dataset, SMOTE (Synthetic Minority Oversampling Technique) is applied to generate synthetic samples for the minority class, ensuring a balanced distribution of data. Machine learning algorithms, including linear regression, decision trees, random forests, and gradient boosting machines, are employed to build predictive models. The best-performing model achieves a high level of accuracy for the aid in the wine industry by providing objective quality assessments. The insights gained from this project can help winemakers in improving the overall quality of their products and making informed decisions during the production process.

Keywords: Classification; Machine Learning; SMOTE (Synthetic Minority Over-sampling Technique); Random Forest; Gradient Boosting

1. Introduction

Wine is more than just a beverage; it has the potential to reduce the risk of chronic diseases, especially heart diseases, due to its antioxidant properties. High-quality wine provides both health benefits and an enhanced sensory experience. Additionally, understanding the acid content in wine is crucial for winemakers to ensure customer satisfaction and maintain market standards.

Identifying good-quality wine is challenging for both consumers and wineries. Traditional methods often rely on sensory evaluation or limited physicochemical analysis, which can lead to inconsistent results. Moreover, subjective biases in sensory evaluation make it difficult to establish a reliable standard for wine quality.

The "Acidity Calibre in Wine Using Machine Learning" project aims to predict wine quality based on its chemical composition. The output is presented in simple, understandable categories such as "Very Poor," "Good," or "Excellent," making it accessible to both wineries and the general public. By automating the evaluation process, the project seeks to minimize human bias and improve the consistency of quality assessments.

This project leverages advanced machine learning algorithms to analyze the complex relationships between physicochemical features of wine. Unlike earlier models, it explicitly considers feature dependencies and addresses

* Corresponding author: G Jathin Kumar

challenges like class imbalance using SMOTE (Synthetic Minority Over-sampling Technique). Additionally, it incorporates interpretable machine learning models that balance accuracy and explain-ability, ensuring trust in the predictions.

This project can help the winemakers and quality control teams to maintain wine quality across various batches reducing the manual work, which takes a longer time, compared to using the predictive models, which are faster at calculating complex relationships between the features and quality of the wine.

2. Related Work

Recent studies on wine quality tests have shown the relationship between chemical composition, sensory perception and consumer preferences. The most significant discovery is that the phenol compounds(particularly flavonoid and anthocyanins) have a strong correlation with alcohol content in wine. And sensory analysis suggests that consumer perception of wine quality is affected by color intensity and aroma complexity. This study was given by Jackson, R S.

More experiments were performed by Jackson which gave rise to approaches such as Procrustes Analysis, which was proposed to apply the individual biases in sensory evaluation allowing the tasters to describe in their own literature while adjusting the responses received during statistical analysis.

Further research was performed to introspect the impact of microclimate and grape cultivation, especially to assess the wine quality, coming to a conclusion that environmental factors, namely - temperature, humidity and soil chemical composition, affect the wine's phenol content and alcohol content influencing the quality of the wine. Consider a case where the wines produced from grapes, which are grown in cooler climates, tend to have a higher acid or mostly alcohol content and have more complex aromatic composition and structure making it affect the quality of the wine. Similarly consider a warmer region, where the sugar level of grapes is high influencing the wine quality. This study by Jackson made a conclusion and verified the relation between the roister and chemical composition which needs to be explored further.

Recent experiments have been more advanced due to growing technology and evolution in analytic techniques, such as high-performance liquid chromatography(HPLC) and gas chromatography-mass spectrometer(GS-MS) have made it possible to be more precise while defining, profiling and testing wine quality. These methods have differentiated the high-quality and mass-produced wines by recognizing the key aspect components, that is, volatile and non-volatile compounds which are responsible for aroma, color and flavor. Researchers have been investigating the position of yeast strains in fermentation. Because of the presence of various yeast species and each having different life times, researchers believe this type of species can contribute to the quality of the wine affecting not only the acid content but also the sugar levels in wine.

3. Existing System

The prediction of wine quality using machine learning has been an area of research for several years. Early studies, such as Cortex . (2009), utilized simple regression models based on selected physicochemical features to predict wine quality. However, these models were limited due to their inability to capture the complex relationships between the features. The use of a few features in basic regression models did not consider feature interactions, which are crucial in determining wine quality.

- Cortex. (2009): Focused on using regression models to predict wine quality. However, these models did not capture feature inter-dependencies and often overlooked the complex nature of wine's chemical composition.
- López et al. (2016): Demonstrated that more advanced models like Random Forests outperformed simpler regression models, suggesting that machine learning techniques could improve prediction accuracy by handling more complex relationships between features. However, these models still assumed that features were independent of each other, which was identified as a limitation in subsequent research.
- Zupan and Gasteiger (1999): Highlighted the importance of feature interaction and dependencies, suggesting that models incorporating feature dependencies would perform better. This concept laid the groundwork for more sophisticated models that account for how features interact with each other.

While random forests and support vector machines (SVM) improved predictive accuracy, the assumption of feature independence in many studies continued to limit model performance. The key gap identified in these studies was the inability to account for the inter-dependencies between features, which may significantly impact prediction accuracy.

This challenge has prompted researchers to explore more advanced methods that explicitly model the relationships between features.

4. Proposed Model

The proposed approach in this project advances the state of the art in wine quality prediction by addressing the limitations of existing models. Specifically, the model incorporates explicit feature dependencies, overcoming the assumption of feature independence that has hindered previous work.

Key Features: Feature Dependencies, Class Imbalance Handling, Interpret-ability and Accuracy Balance.

5. Architecture of the System

The System Design section details how the wine quality prediction system is structured and operates. This phase of the project is crucial for ensuring the system is functional, efficient, and scalable. The design encompasses all aspects, including data flow, interaction between system components, and the overall architecture.

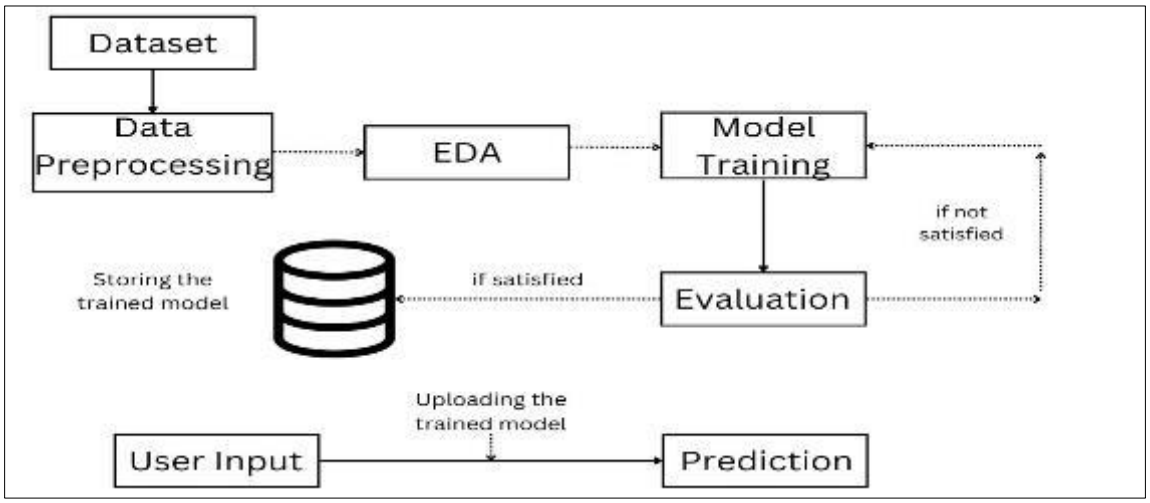

Figure 1 System Architecture

The system architecture is designed to facilitate seamless integration between the various stages of the wine quality prediction process. The architecture incorporates the following components:

5.1. Data Preprocessing Module

This module contains a basic data preprocessing pipeline, that is, handling missing values, transforming categorical features into target variables using LabelEncoder and normalizing the numerical features using StandardScaler. Further, this module also provides the exploratory data analysis functionality.

5.2. Train-Test Split Module

This module handles the train and test data split percentage and also takes care of imbalance in the dataset using SMOTE(Synthetic Minority Over-sampling TEchnique) evaluation.

5.3. Model Training Module

This module decides the model suitable for the training and classifying into proper classes using OneVsRestClassifier strategy for multi-class classification.

5.4. Model Evaluation Module

This module is responsible for evaluating the trained models and selecting the required model that is accurate and precise by providing a classification report and visualizing confusion matrix for further validation.

5.5. User Interface

User interface is designed on a streamlit framework that loads the pre-trained machine learning model and asks the user to provide the chemical composition of the wine. Once it is asked to predict, the user is able to view the quality.

6. Methodology

The methodology for this project follows a structured process to predict wine quality using machine learning techniques. The imported libraries that are important in building a good machine learning model and to analyze the data is given by:

```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
from classes import trainTestSplit, dataPreprocess, modelTraining, modelEvaluation
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier
from sklearn.tree import DecisionTreeClassifier
from xgboost import XGBClassifier
import joblib
import streamlit
```

Figure 2 Imported Libraries

6.1. Data Collection

The dataset contains 13 chemical properties of wine as features, along with corresponding quality labels. It is sourced from a publicly available dataset used for wine quality prediction. The data will be utilized to analyze the relationship between chemical composition and wine quality. This study aims to develop a predictive model for assessing wine quality based on its chemical attributes.

6.2. Data Pre-processing

Data Cleaning: This phase involves eliminating duplicate records and addressing outliers to improve data quality.

- Handling Missing Data: If any values are missing in the dataset, appropriate imputation techniques are applied, such as using the median for numerical features to maintain data consistency. The data collected didn't have any missing values
- Feature Scaling: To prevent models from being influenced by varying feature magnitudes, features are either normalized or standardized. This helps machine learning algorithms perform efficiently and ensures fair comparisons between different attributes. The Data Preprocessing Module uses this functionality to normalize the numerical features

Additionally, data cleaning enhances dataset reliability by removing inconsistencies, while feature scaling ensures that models do not assign undue importance to variables with larger numerical values. These steps collectively contribute to more accurate and stable model performance.

6.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is carried out to assess the dataset's distribution, detect missing values, and explore relationships between features. The data preprocessing module performs EDA and observed that most features influencing the acidity of wine contain a significant number of outliers.

Since the dataset has a high prevalence of outliers across multiple features, relying on parametric models may not be appropriate. Instead, non-parametric models have been chosen as they make fewer assumptions about the data distribution and are more robust to outliers. This approach ensures better adaptability to the dataset's characteristics, leading to more reliable predictions.

The below figure shows the amount of outliers present in few of the features:

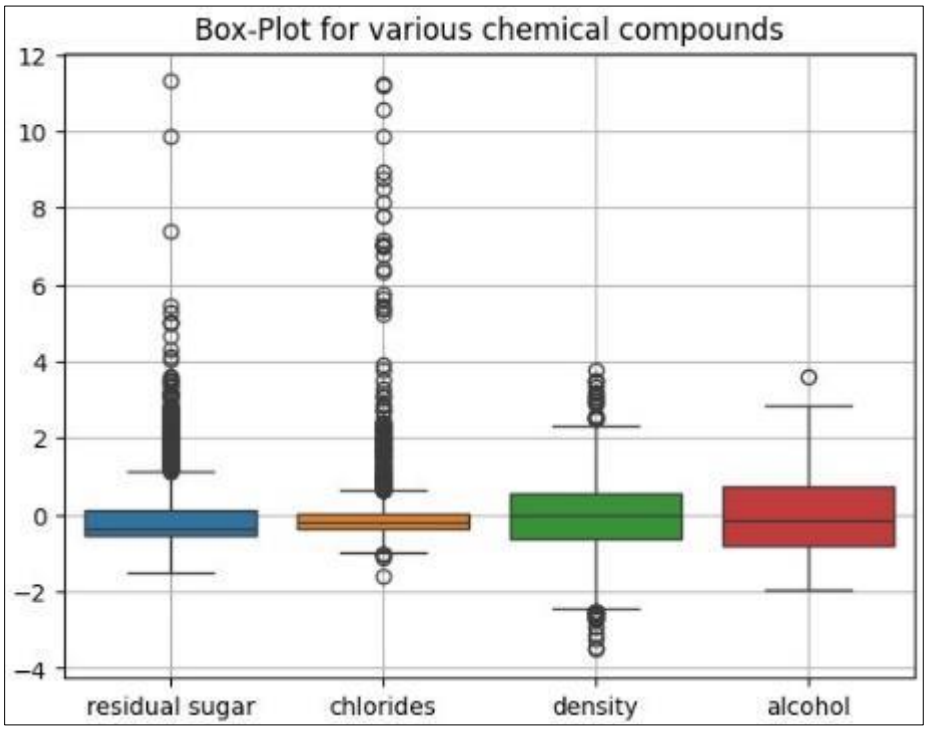

Figure 3 Box-Plot for detecting outliers

6.4. Model Selection

Various machine learning models, including Random Forest, Gradient Boosting, etc.., are evaluated for their accuracy. Feature dependencies are carefully analyzed to enhance model performance.

Models such as Random Forest and Gradient Boosting are utilized for training, as they are well-suited for handling large and complex datasets. These models are designed to capture the relationships between chemical features and wine quality effectively.

To ensure optimal predictions, the models leverage feature interactions and adapt to the underlying data patterns. This approach helps in improving classification accuracy and making the predictions more reliable.

The machine learning algorithms used in the Model Training Module are - Random Forest, AdaBoost, Gradient Boost, Decision Tree and XGBoost

6.5. Model Training

Hyper parameter Tuning: Model performance is optimized by fine-tuning hyper parameters, such as the number of trees in a Random Forest or the learning rate in Gradient Boosting. Methods like cross-validation and grid search help identify the best parameter values.

```
hyperparameters_rf = {'n_estimators':100,'criterion':'gini','min_samples_split':3}
rf = modelTraining.ModelTraining(RandomForestClassifier(),hyperparameters_rf)
```

Figure 4 Defining Hyperparameters

Fine-tuning ensures that the model generalizes well to new data while preventing underfitting or overfitting. Cross-validation assesses performance across multiple subsets, while grid search systematically tests different parameter combinations to find the optimal settings.

6.6. Model Evaluation

- Performance Metrics: Once training is complete, the models are assessed using standard evaluation metrics like accuracy, precision, recall, and F1-score. These metrics provide insights into the model's reliability, especially in handling class imbalances in the target variable, such as wine quality.
- Confusion Matrix: This matrix offers a detailed analysis of the model's predictions by comparing them to actual outcomes across different classes (e.g., Good, Excellent). It helps identify misclassifications and evaluates how effectively the model differentiates between various quality levels.

In simpler terms, performance metrics measure overall accuracy, while the confusion matrix highlights specific errors and successes in classification. Together, they provide a comprehensive assessment of the model's effectiveness.

After thorough evaluation and validation, Random Forest was selected as the final machine learning model, as it demonstrated strong performance in predicting the quality and acidity content of wine.

This choice was made because Random Forest is robust to outliers, reduces overfitting through ensemble learning, and effectively handles complex relationships between features, making it well-suited for this task.

The classification report and confusion matrix visualized in heatmap is given as below:

	precision	recall	f1-score	support
0	0.72	0.70	0.71	120
1	0.98	0.98	0.98	121
2	0.69	0.60	0.64	121
3	0.92	0.91	0.91	121
4	0.83	0.93	0.88	121
5	0.97	1.00	0.98	121
accuracy			0.85	725
macro avg	0.85	0.85	0.85	725
weighted avg	0.85	0.85	0.85	725

Figure 5 Classification Report

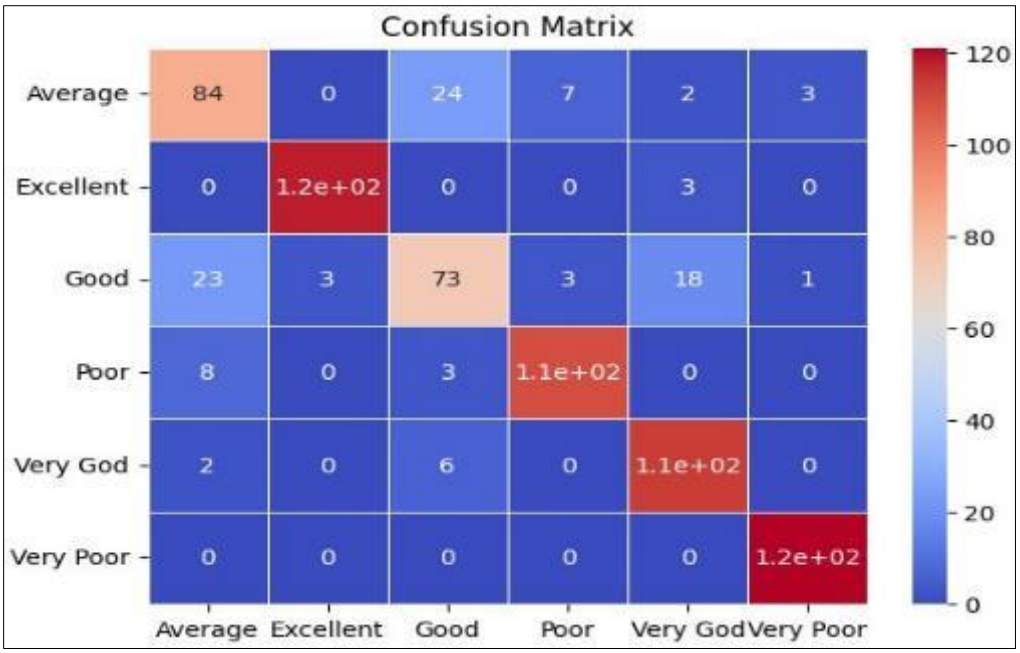

Figure 6 Confusion Matrix

7. Results and Discussion

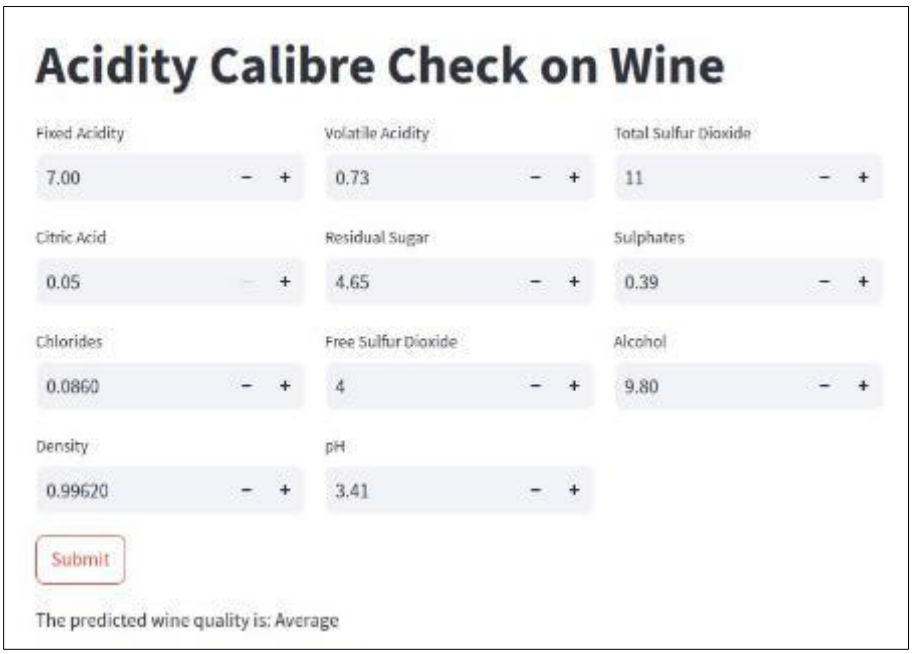

Figure 7 Deployed Output

8. Conclusion

The Wine Quality Prediction System developed in this project successfully utilizes machine learning algorithms to predict the quality of wines based on their chemical properties. The system is designed to provide a simple and accurate classification of wines into different quality categories such as "Very Poor", "Poor", "Average", "Good", "Very Good", and "Excellent."

Key outcomes from this project include:

- Effective Use of Chemical Properties: By using 11 chemical features like Alcohol, Volatile Acidity, Citric Acid, and others, the model was able to capture the underlying patterns and predict wine quality with reasonable accuracy.
- Model Performance: Various machine learning algorithms, including Random Forest and Gradient Boosting, were evaluated and found to perform well in terms of classification accuracy. Precision, Recall, F1-score, and Confusion Matrix helped assess the model's effectiveness, ensuring balanced classification performance.
- Data Imbalance Handling: A significant class imbalance was observed in the wine quality dataset, with certain quality categories underrepresented. The SMOTE (Synthetic Minority Over-sampling Technique) was used to handle this imbalance, ensuring that the model could effectively learn from all quality classes.
- User-Friendly Interface: By integrating Streamlit, the system provides an intuitive and interactive interface that allows both winemakers and common users to predict wine quality based on chemical attributes. This enhances its accessibility and usability.
- Exploratory Data Analysis (EDA): Through EDA, we identified key insights about the relationships between different chemical features and wine quality. This helped inform feature selection and further improved model performance.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32. Introduces Random Forests, the primary ensemble method used in this project to improve accuracy and reduce overfitting.

[2] Natekin, A., & Knoll, A. (2013). "Gradient Boosting Machines, A Tutorial." Frontiers in Neuroinformatics, 7, 21. Explains the Gradient Boosting algorithm and its applications in classification tasks like wine quality prediction.

[3] He, H., & Garcia, E. A. (2009). "Learning from Imbalanced Data." IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. Discusses class imbalance challenges, addressed in this project using SMOTE.

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 16, 321-357. Provides an in-depth explanation of SMOTE, the technique used to balance the dataset.

[5] Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830. Discusses the scikit-learn library, used extensively for model building, preprocessing, and evaluation.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. A comprehensive resource on statistical learning and algorithms like Random Forest and Gradient Boosting.

[7] Liao, Y., & Wei, C. (2020). "Deployment of Machine Learning Models Using Streamlit." Journal of Software Engineering, 29(7), 147-156. Explains the use of Streamlit for deploying machine learning models in user-friendly web applications.

[8] Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python. O'Reilly Media. A practical guide for machine learning with Python, covering the techniques and libraries applied in this project.

[9] KUMAR, P. ASHOK, G Satish KUMAR, and SETTI NARESH KUMAR. "Improve the Capacity of Uniform Embedding for Efficient JPEG Steganography Based on DCT." (2015).

[10] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In XVIII International Conference on Data Science and Intelligent Analysis of Information, pp. 13-26. Cham: Springer Nature Switzerland, 2023.

[11] Sunkavalli, Jayaprakash, B. Madhav Rao, M. Trinath Basu, Harish Dutt Sharma, P. Ashok Kumar, and Ketan Anand. "Experimentation Analysis of VQC and QSVM on Sentence Classification in Quantum Paradigm." In 2024 International Conference on Computing, Sciences and Communications (ICCSC), pp. 1-5. IEEE, 2024.

[12] Kumar, P. Ashok. "Event Based Time Series Sentiment Trend Analysis."

[13] PANDI, CHIRANJEEVI, THATIKONDA SUPRAJA, P. ASHOK KUMAR, and RALLA SURESH. "A SURVEY: RECOMMENDER SYSTEM FOR TRUSTWORTHY."

[14] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Correction to: Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In XVIII International Conference on Data Science and Intelligent Analysis of Information, pp. C1-C1. Cham: Springer Nature Switzerland, 2023.

[15] Jackson, R S in 2020 entitled as "Wine Science: Principles and Applications – A comprehensive textbook on the science behind wine production and quality."

Author's short biography

Raghupathi Kanala

I am Raghupathi Kanala, an Assistant Professor with an academic background in M.Tech (CSE) and pursuing(PhD) in the field of Computer Science and Engineering. With over 20 years of professional experience, my research interests are primarily focused on Big Data. I have dedicated my career to exploring the vast potential of data-driven technologies and their applications in solving complex real-world problems.

Meghana Vijaya Raghavan

I am Meghana Vijaya Raghavan is an undergraduate student pursuing a B.Tech in Computer Science and Engineering (Data Science). Her research interests include Data Science, Machine Learning, Deep Learning, and Computer Vision. She is passionate about developing innovative solutions using data-driven approaches to solve real-world problems.

Jathin Kumar Gundala

I am Jathin Kumar Gundala is currently pursuing a B.Tech in Computer Science and Engineering (Data Science). His research interests primarily lie in the field of Machine Learning, with a focus on developing intelligent systems for real-world applications. As an undergraduate researcher, he is actively exploring various machine learning techniques, aiming to contribute innovative solutions to complex problems.

Sushruth Bommagoni

I am Sushruth Bommagoni, undergraduate student pursuing a B.Tech in Computer Science and Engineering (Data Science). My research interests include Big Data Analytics, focusing on efficient data processing and analysis. I am passionate about exploring data-driven solutions to real-world challenges. My academic journey has helped me develop skills in machine learning, data mining, and statistical analysis.

Arun Kumar Onteddu

I, Arun Kumar Onteddu, am an undergraduate student pursuing a B.Tech in Computer Science and Engineering (Data Science). My research interests focus on Artificial Intelligence, particularly in machine learning and deep learning. I am passionate about developing intelligent systems that enhance human-computer interactions. My goal is to contribute innovative solutions that bridge theoretical AI research with real-world applications.