## Artificial Intelligence and Machine

## Learning

Project Report

Semester-IV (Batch-2022)

**WINE QUALITY PREDICTION**

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Mr. Animesh Amul Arora (2210991256)

Amrit pal Singh (2210991255)

Ashmeet Monga(2210991406)

Ayush Garg(2210991423)

**Department of Computer Science and Engineering**

Chitkara University Institute of Engineering & Technology,

Chitkara University, Punjab

**Title:** WINE QUALITY ANALYSIS

**Abstract:**

**Introduction:**

Wine quality is a critical factor for both consumers and producers. Analyzing the quality of wine involves understanding the chemical properties that contribute to taste, aroma, and overall acceptability. This project aims to predict wine quality using machine learning techniques based on various physicochemical tests.

**Data Collection and Preprocessing:**

Data Collection:

The dataset for this project, sourced from the UCI Machine Learning Repository, includes two datasets: one for red wine and another for white wine. Both datasets comprise several physicochemical properties that affect the quality of wine. Here are the key attributes collected:

Fixed Acidity: Primarily tartaric acid, which affects the wine’s taste and preservation.

Volatile Acidity: Acetic acid concentration, contributing to vinegar taste at high levels.

Citric Acid: Adds freshness and flavor.

Residual Sugar: Remaining sugar post-fermentation; influences sweetness.

Chlorides: Salt content; affects taste.

Free Sulfur Dioxide: Prevents microbial growth and oxidation.

Total Sulfur Dioxide: Sum of bound and free forms.

Density: Related to alcohol and sugar content.

pH: Measure of acidity.

Sulphates: Contribute to wine’s preservation and taste.

Alcohol: Affects the wine’s body and sensory profile.

Quality: Subjective score ranging from 0 to 10.

Data preprocessing:

Data Cleaning: Checked for missing values and outliers.

Normalization: Scaled the features to ensure they have a similar range.

Feature Engineering: Created new features and selected the most relevant ones.

**Exploratory Data Analysis (EDA):**

Correlation Analysis: Identified correlations between different features and quality.

Distribution Analysis: Analyzed the distribution of each feature.

Box Plots: Visualized the spread of quality scores for red wines.

**Supervised Machine Learning Algorithms:**

Multiple supervised learning algorithms are implemented and evaluated to predict disease occurrence based on the collected data. Algorithms such as Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), Naive Bayes, and Neural Networks are among the models examined in this investigation. Each algorithm is trained on a subset of the medical dataset and evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Feature selection and importance analysis techniques are employed to identify the most significant predictors of disease onset.

**Model Evaluation and Comparison:**

The performance of each predictive model is rigorously evaluated using cross-validation techniques and hyperparameter tuning. Visualizations, including heatmaps and feature importance plots, are generated to illustrate the relationships between various health parameters and disease likelihood. Results indicate the efficacy of different AIML algorithms in predicting disease occurrence, with particular emphasis on model accuracy, sensitivity, and specificity. Comparative analysis reveals the strengths and limitations of each model, aiding in the selection of the most suitable algorithm for disease prediction tasks

**Conclusion:**

The project successfully identified and predicted wine quality using physicochemical properties. The Random Forest model showed the highest accuracy among the models tested. Further improvements can be made by including more advanced feature engineering and hyperparameter tuning.

**Future Work:**

Incorporate Additional Data: Include sensory data for a more holistic analysis.

Advanced Techniques: Implement more sophisticated models like ensemble learning.

Feature Importance Analysis: Further investigate which features contribute most to wine quality prediction.

**In artificial intelligence and machine learning, there are several types of learning approaches, each with its own characteristics and applications. The main types of learning in artificial intelligence and machine learning include:**

**Supervised Learning:**

In supervised learning, the algorithm learns from labeled data, where each training example is associated with a corresponding target label or outcome.

The goal of supervised learning is to learn a mapping from input features to the correct output labels based on the provided training data.

Supervised learning tasks include classification, where the algorithm predicts a categorical label, and regression, where the algorithm predicts a continuous numerical value.

Examples of supervised learning algorithms include decision trees, linear regression, logistic regression, support vector machines (SVM), neural networks, and ensemble methods like random forests and gradient boosting.

**Unsupervised Learning:**

In unsupervised learning, the algorithm learns from unlabeled data, where the training data does not have corresponding output labels or target values.

The goal of unsupervised learning is to find hidden patterns or structures within the data, without explicit guidance or supervision.

Unsupervised learning tasks include clustering, where the algorithm groups similar data points together, and dimensionality reduction, where the algorithm reduces the number of features while preserving important information.

Examples of unsupervised learning algorithms include k-means clustering, hierarchical clustering, principal component

**In supervised learning, algorithms can be categorized into several types based on the nature of the task they perform. Here are the main types of algorithms used in supervised learning:**

**Random Forest Regression:**

Random forest regression is an ensemble learning method that combines multiple decision trees to make predictions.

It constructs a large number of decision trees during training and outputs the average prediction of individual trees (regression) or the mode (classification).

Random forests reduce overfitting and improve generalization by aggregating the predictions of multiple weak learners (decision trees) trained on random subsets of the data.

Random forest regression is effective for handling high-dimensional data and capturing complex nonlinear relationships.

**Support Vector Machine (SVM):**

Support vector machine is a supervised learning algorithm used for classification and regression tasks.

It constructs a hyperplane or set of hyperplanes in a high-dimensional space to separate classes or approximate the relationship between input features and target variables.

SVM aims to maximize the margin between the nearest data points (support vectors) of different classes, making it robust to outliers and effective in high-dimensional spaces.

SVM can handle both linear and nonlinear relationships through the use of different kernel functions such as linear, polynomial, and radial basis function (RBF) kernels.

**XG - Boost:**

XG Boost (Extreme Gradient Boosting) is an ensemble learning method that uses a gradient boosting framework for classification and regression tasks.

It sequentially adds weak learners (decision trees) to the ensemble, with each tree trained to correct the errors of the previous trees.

X Boost optimizes a differentiable loss function by gradient descent, enhancing the performance of the model with each iteration.

XG Boost is known for its efficiency, scalability, and ability to handle sparse data, making it a popular choice for structured data problems and winning numerous machine learning competitions.

**LOGISTIC REGRESSION:**

Logistic Regression is a linear model used for classification tasks. In this project, it's employed to predict the quality of wine based on its physicochemical properties.

Predict the probability of wine belonging to different quality categories (0-10).

Multi-class classification using One-vs-Rest (OvR) or multinomial logistic regression.

**pip install is a command-line tool used in Python for installing Python packages from the Python Package Index (PyPI) or other repositories. It simplifies the process of installing and managing external libraries and dependencies for Python projects.**

**In my project, you've utilized several essential libraries for data manipulation, visualization, and machine learning tasks. Here is a brief overview of each library:**

**Pandas:**

pandas is a powerful data manipulation and analysis library in Python.

It provides data structures like DataFrame and Series, making it easy to work with structured data.

pandas is widely used for tasks such as data cleaning, transformation, and exploratory data analysis.

**NumPy:**

numpy is a fundamental package for scientific computing in Python.

It provides support for multidimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.

numpy is essential for numerical computing tasks and serves as the backbone for many other scientific Python libraries.

**Matplotlib:**

matplotlib is a plotting library in Python that provides a MATLAB-like interface for creating static, interactive, and publication-quality visualizations.

It offers a wide range of plotting functions and customization options for creating various types of plots, including line plots, scatter plots, histograms, and more.

**Seaborn (SNS):**

seaborn is a statistical data visualization library built on top of matplotlib.

It provides a higher-level interface for creating attractive and informative statistical graphics, with built-in support for complex visualizations like heatmaps, pair plots, and categorical plots.

seaborn simplifies the process of creating visually appealing plots with minimal code.

**Scikit-learn (Sklearn):**

scikit-learn is a machine learning library in Python that provides simple and efficient tools for data mining and data analysis.

It includes a wide range of algorithms for classification, regression, clustering, dimensionality reduction, and model evaluation.

scikit-learn is designed to be user-friendly and easily integrated into existing Python workflows, making it a popular choice for machine learning projects.

**Xg boost:**

Xg boost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting machines.

It is widely used for supervised learning tasks such as classification and regression, offering high performance and accuracy.

Xg boost is known for its speed, flexibility, and effectiveness in handling large datasets, making it a popular choice in machine learning competitions and real-world applications.

By using pip install along with the names of these libraries, I can easily install them in my Python environment and leverage their functionality to build and enhance your project.

**Some Screenshot’s of My project of importing data and analyzing using graph:**

**Importing data:**



