Wine Quality Prediction using ML

A REPORT

submitted by

Adithya S.T.(18MIS1025)

in partial fulfilment for the award

of

M. Tech. Software Engineering (Integrated)

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

MAY 2022

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

DECLARATION

I hereby declare that the project entitled **“Wine Quality Prediction using ML”** submitted by me to the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai 600127 in partial fulfilment of the requirements for the award of the degree of **Master of Technology - Software Engineering (Integrated)** is a record of bonafide work carried out by me**.** I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or university.

Signature

Adithya S.T.(18MIS1025)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

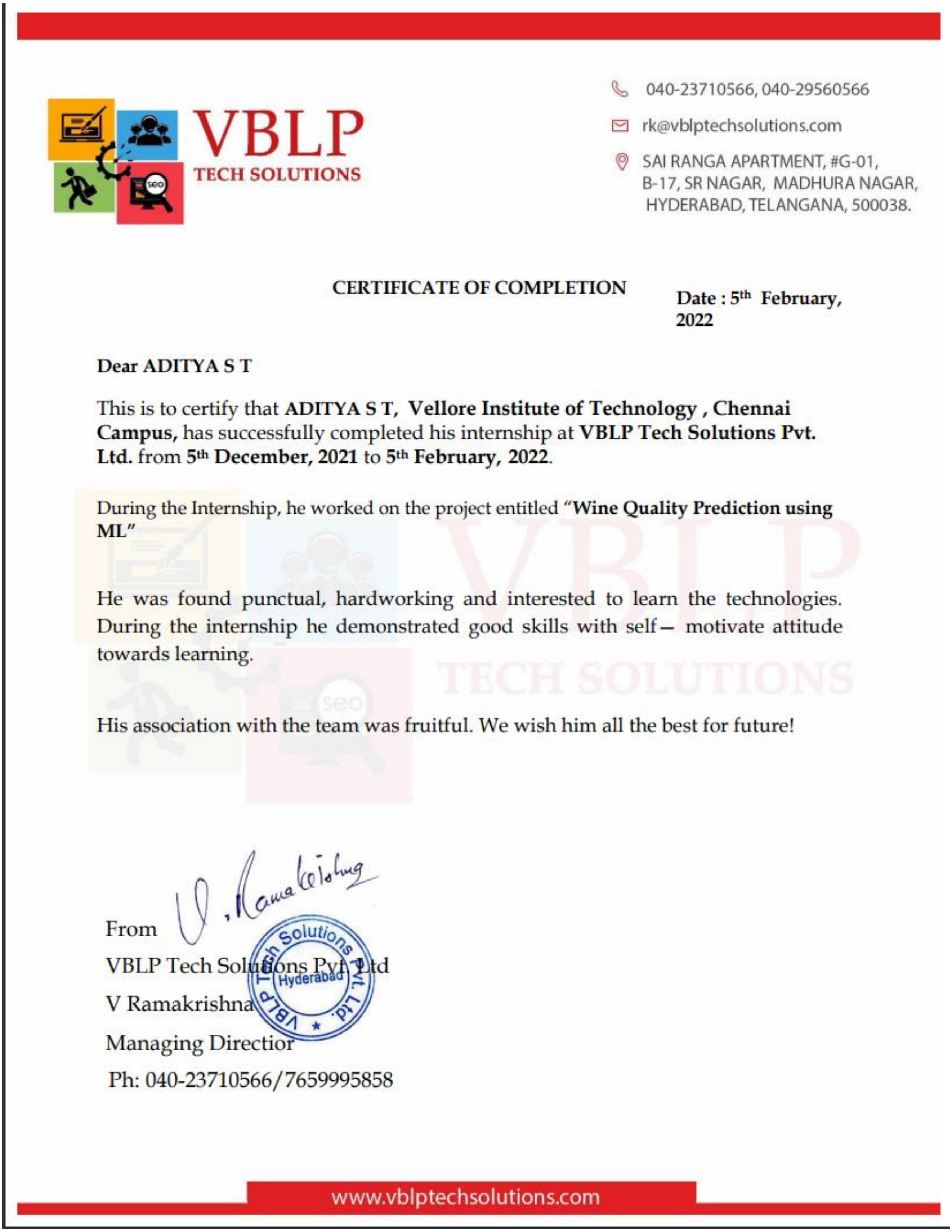
CERTIFICATE

The project report entitled "**Wine Quality Prediction using ML**" is prepared and submitted by **Adithya S.T.(Register No: 18MIS0125)**. It has been found satisfactory in terms of scope, quality and presentation as partial fulfilment of the requirements for the award of the degree of **Master of Technology – Software Engineering (Integrated)** in Vellore Institute of Technology, Chennai, India.

Examined by:

Examiner I **Examiner II**

Certificate of Merit:



VBLP
TECH SOLUTIONS

☎ 040-23710566, 040-29560566
✉ rk@vblptechsolutions.com
SAI RANGA APARTMENT, #G-01,
B-17, SR NAGAR, MADHURA NAGAR,
HYDERABAD, TELANGANA, 500038.

CERTIFICATE OF COMPLETION

Date : 5th February, 2022

Dear ADITYA S T

This is to certify that **ADITYA S T, Vellore Institute of Technology , Chennai Campus,** has successfully completed his internship at **VBLP Tech Solutions Pvt. Ltd.** from **5th December, 2021** to **5th February, 2022**.

During the Internship, he worked on the project entitled "**Wine Quality Prediction using ML**"

He was found punctual, hardworking and interested to learn the technologies. During the internship he demonstrated good skills with self— motivate attitude towards learning.

His association with the team was fruitful. We wish him all the best for future!

V. Ramakrishna

VBLP Tech Solutions Pvt. Ltd.
Hyderabad
*

From
VBLP Tech Solutions Pvt Ltd
V Ramakrishna
Managing Directior
Ph: 040-23710566/7659995858

www.vblptechsolutions.com

Link of the PDF:
https://1drv.ms/b/s!Am-eVTSjDSr6geB05kcXWT67oXQV6g?e=z3hANz

ACKNOWLEDGEMENT

Our project would not have been successful without the help of several people. We would like to thank the personalities who were part of our project in numerous ways, those who gave us outstanding support from the birth of the project.

We are also thankful to all the Company staff members and faculties of Computer Science and Engineering department who have co-operated in making our project a success. We would like to thank all our parents and friends who extended their help, encouragement and moral support either directly or indirectly in our project work.

Sincerely,
Adithya S.T.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Expansion
KNN	K-Nearest Neighbour
PCA	Principal component analysis
SVM	Support Vector Machine
SDLC	Software Development LifeCycle

ABSTRACT

Wines are being produced since thousands of years. But, it is a complex process to determine the relation between the subjective quality of a wine and its chemical composition. Industries use Product Quality Certification to promote their products and become concern for every individual who consumes any product. It is not possible to ensure quality with experts with such a huge demand of product as it will increase the cost. Wine-makers need a permanent solution to optimize the quality of their wine. This paper explores the space to easy out and make the whole process cost-effective and more trustworthy using machine learning. It allows to build a model with user interface which predicts the wine quality by selecting the important parameters of wine which play a significant role in determining the wines quality. Random forest algorithm is used in determining wines' quality whose correctness would further be escalated using KNN which makes our model dynamic. Output of this proposed model is used to determine the wines' quality on a scale of Good, Average or Bad. This proposed model can further be applied to several other products which need quality certification. Our prediction model provides ideal solution for the analysis of wine, which makes this whole process more efficient and cheaper with less human interaction.

CHAPTER 1

INTRODUCTION

1.1 Problem statement :

In recent years there is a modest increase in the wine consumption as it has been found that wine consumption has a positive correlation to the heart rate variability [1]. With the increase in the consumption wine industries are looking for alternatives to produce good quality wine at less cost. Different wines have different purposes. Although most of the chemicals are same for different type of wine based on the chemical tests, the quantity of each chemicals have different level of concentration for different type of wine. These days it is really important to classify different wine for quality assurance [2]. In the past due to lack of technological resources it become difficult for most of the industries to classify the wines based on the chemical analysis as it takes lot of time and also need more money. These days with the advent of the machine learning techniques it is possible to classify the wines as well as it is possible to figure out the importance of each chemical analysis parameters in the wine and which one to ignore for reduction of cost. The performance comparison with different feature sets will also help to classify it in a more distinctive way. In this paper an intelligent approach is proposed by considering genetic algorithm (GA) based feature selection as well as simulated annealing-based feature selection considering the nonlinear classifiers, linea classifiers and probabilistic classifiers to predict the quality in red wine as well as the white wine.

1.2 Motivation:

In the past few attempts have been made to use different machine learning approaches and feature selection techniques to the wine dataset. Er and Atasoy proposed a method to classify the quality of wines using three different classifier such as support vector machines, Random forest and k-nearest neighborhood. They have used principal component analysis for feature selection and they found good result using Random forest algorithm [3]. Chen et al proposed an approach that will predict the grade of wine using the human savory reviews. They have used hierarchical clustering approach and association rule algorithm to process the reviews and predict the wine grade and they found an accuracy of 85.25% while predicting the grade [4]. Appalasamy et al proposed a method to predict wine quality based on physiochemical test data. They have pointed out that classification approach helps to improve the quality of wine during the production [5].

1.3 Objective:

Beltrán et al proposed an approach to classify the wine based on aroma chromatograms and they have used PCA for dimensionality reduction and wavelet transform for feature extraction and classifiers such as neural network, linear discriminant analysis and support vector machine and found that support vector machine with wavelet transforms perform better than other classifiers [6]. Thakkar et al., used analytical hierarchy process (ahp) to rank the attributes and then used different machine learning classifiers such as support vector machine and random forest and they found accuracy of 70.33% using random forest and 66.54% using SVM [7]. Reddy and Govindarajulu used a user centric clustering approach to recommend the product. They have used red wine data set for the survey purpose. They have allocated relative voting to the attributes based on the literature review. Then they assigned weight to the attributes using Gaussian Distribution Process. They judged the quality based on the user preference group [8]. The above past work motivated us to try different feature selection algorithm as well as different classifiers to compare the performance metrics. This paper proposed GA based feature selection and SA based feature selection and used different classifiers such as PART, RPART, Bagging, C5.0, random forest, svm, lda, naïve bayes etc.

1.3.1 Proposed System:

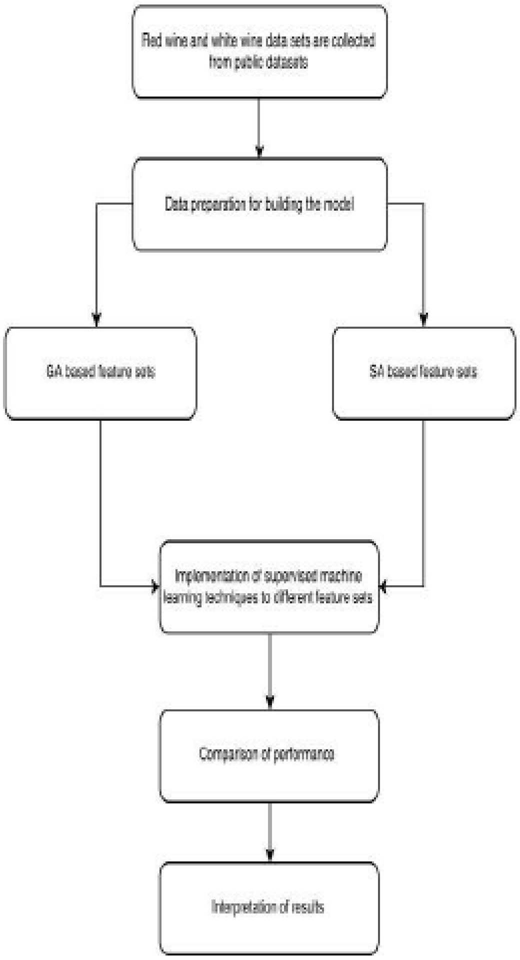
The wine data set is publicly available in the database of UCI. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. This data set contains the physiochemical variables as well as sensory variables; altogether there are 12 attributes [9]. We have used genetic algorithm (GA)-based feature sets for feature selection. Pledsoe first presented an adaptive optimization search methodology is called genetic algorithm and Holland mathematically presented the genetic algorithm-based approach by getting inspiration from Darwin’s theory of evolution.

1.3.2 Advantages of proposed system :

The parameters used to compare the performance and validations of classifier are as follows: accuracy, sensitivity, specificity, positive predictive value (ppv), negative predictive value (npv). The sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. The specificity is defined as the ratio of true negatives to the sum of false positives and true negatives. In our research we have used the Positive predictive value and negative predictive value to check the present and absent of one type of wine. So, the ppv is the probability that the one type of wine is present given a positive test result and npv is the probability that the one type of wine is absent given

CHAPTER 2
SYSTEM DESIGN

2.1SYSTEM ARCHITECTURE:


Red wine and white wine data sets are collected from public datasets
Data preparation for building the model
GA based feature sets
SA based feature sets
Implementation of supervised machine learning techniques to different feature sets
Comparison of performance
Interpretation of results

2.2Module description

Feature selection is an important part in machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model. In wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince classifiers from the features in each subset. The classifier consider the subset of feature with which the classification algorithm performs the best. To find the subset, the evaluator uses different search techniques like depth first search, random search, breadth first search or hybrid search. The filter method uses an attribute evaluator along with a ranker to rank all the features in the dataset. Here one feature is omitted at a time that has lower ranks and then sees the predictive accuracy of the classification algorithm. Weights or rank put by the ranker algorithms are different than those by the classification algorithm. Wrapper method is useful for machine learning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features.

Algorithms used in this project :-

Genetic algorithm works in similar way as its work on chromosomes by taking relevant genes to form new production and remove unhealthy or non-relevant genes. GA algorithm continuously iterate over dataset to look for non-relevant attributes by doing mutation, reproduction and fitness, only those attributes which has high fitness or related to more dataset values can be used for mutation and reproduction and unfitted values will be removed out.

Simulated annealing (SA) is a global search/selection method that makes small random changes (i.e. perturbations) to an initial (dataset values) candidate solution. If the performance value for the perturbed (new Data) value is better than the previous solution, the new solution (data/attribute) is accepted. If not, an acceptance probability is determined based on the difference between the two performance values and the current iteration of the search. From this, a sub-optimal solution can be accepted on the off-change that it may eventually produce a better solution or best attributes in subsequent iterations.

SVM Algorithm: Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

How do we find the right hyperplane?

Or, in other words, how do we best segregate the two classes within the data?

The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.

Random Forest Algorithm**:** it's an ensemble algorithm which means internally it will use multiple classifier algorithms to build accurate classifier model. Internally this algorithm will use decision tree algorithm to generate it train model for classification.

Naive Bayes: Naive Bayes which is one of the most commonly used algorithms for classifying problems is simple probabilistic classifier and is based on Bayes Theorem. It determines the probability of each features occurring in each class and returns the outcome with the highest probability.

2.3System Specification

2.3.1 Software Requirements

Functional requirements for a secure cloud storage service are straightforward:

1. The service should be able to store the user's data;

2. The data should be accessible through any devices connected to the Internet;

3. The service should be capable to synchronize the user's data between multiple devices (notebooks, smart phones, etc.);

4. The service should preserve all historical changes (versioning);

5. Data should be shareable with other users;

6. The service should support SSO; and

7. The service should be interoperable with other cloud storage services, enabling data migration from one CSP to another.

- **Operating System:** Windows
- **Coding Language**: Python 3.7
- **Script:**
- **Database :**

2.3.2 Hardware Requirements:

- **Processor** - Pentium –III
- **Speed** – 2.4 GHz
- **RAM** - 512 MB (min)
- **Hard Disk** - 20 GB
- **Floppy Drive** - 1.44 MB
- **Key Board** - Standard Keyboard
- **Monitor** – 15 VGA Colour

Cloud computing has three fundamental models, these are:

2.4 Detailed Design

UML is an acronym that stands for Unified Modeling Language. Simply put, UML is a modern approach to modeling and documenting software. In fact, it's one of the most popular business process modeling techniques.

It is based on diagrammatic representations of software components. As the old proverb says: "a picture is worth a thousand words". By using visual representations, we are able to better understand possible flaws or errors in software or business processes.

UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems. The need arose for a more unified way to visually represent those systems and as a result, in 1994-1996, the UML was developed by three software engineers working at Rational Software. It was later adopted as the standard in 1997 and has remained the standard ever since, receiving only a few updates.

GOALS:

The Primary goals in the design of the UML are as follows:

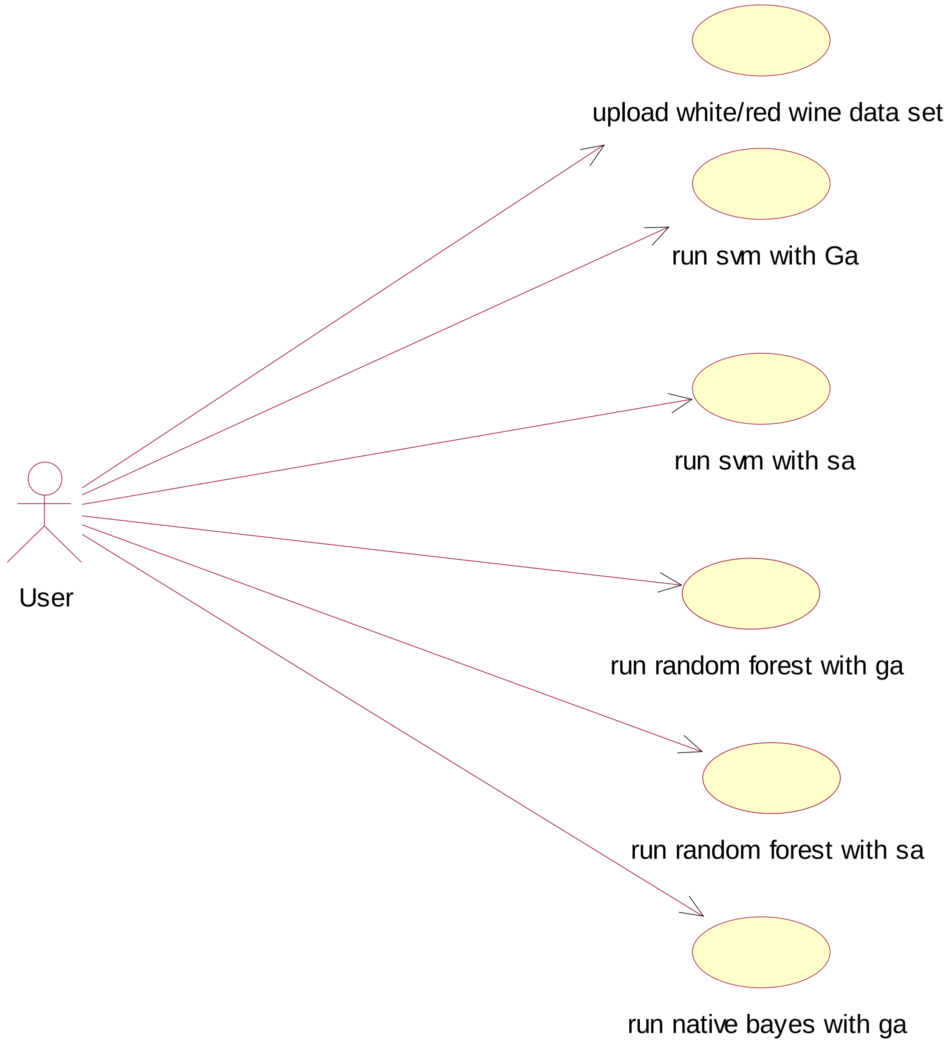
1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.

6 Support higher level development concepts such as collaborations, frameworks, patterns and components.

7 Integrate best practices.

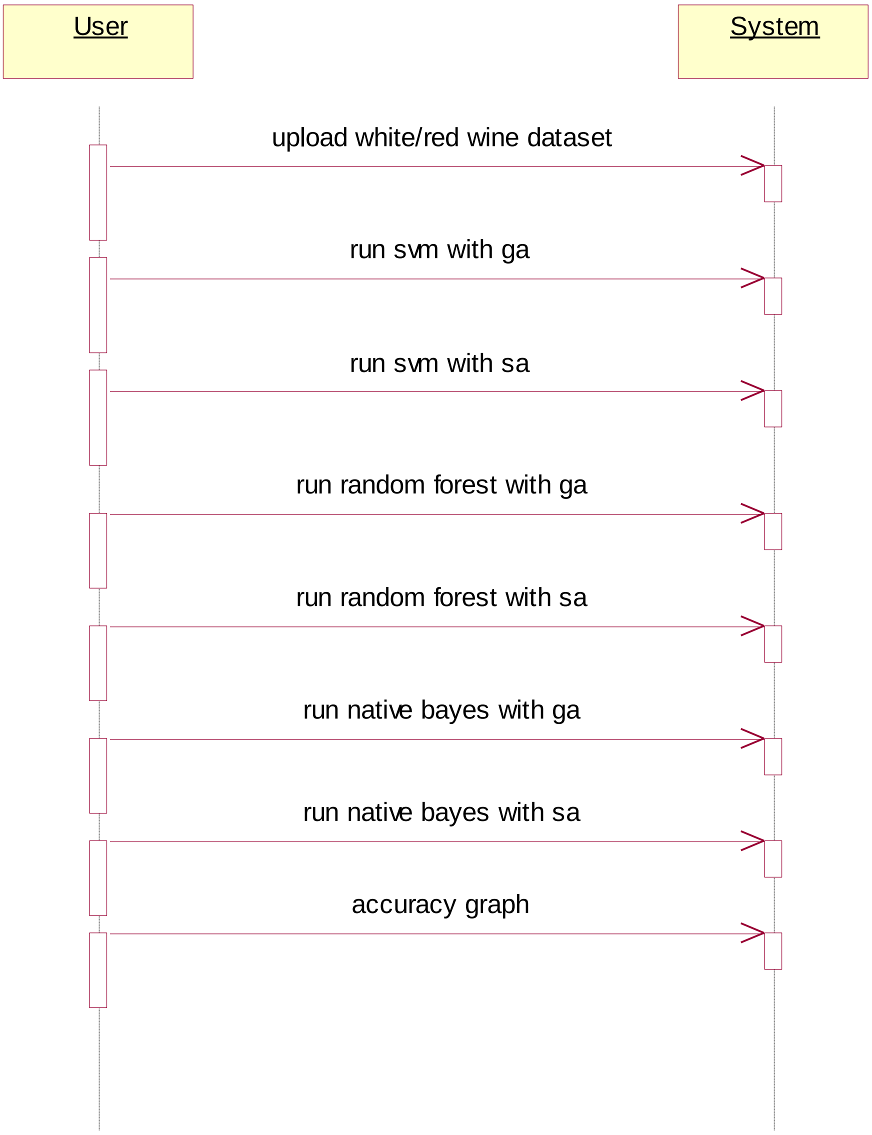
i. USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.


upload white/red wine data set
run svm with Ga
run svm with sa
User
run random forest with ga
run random forest with sa
run native bayes with ga

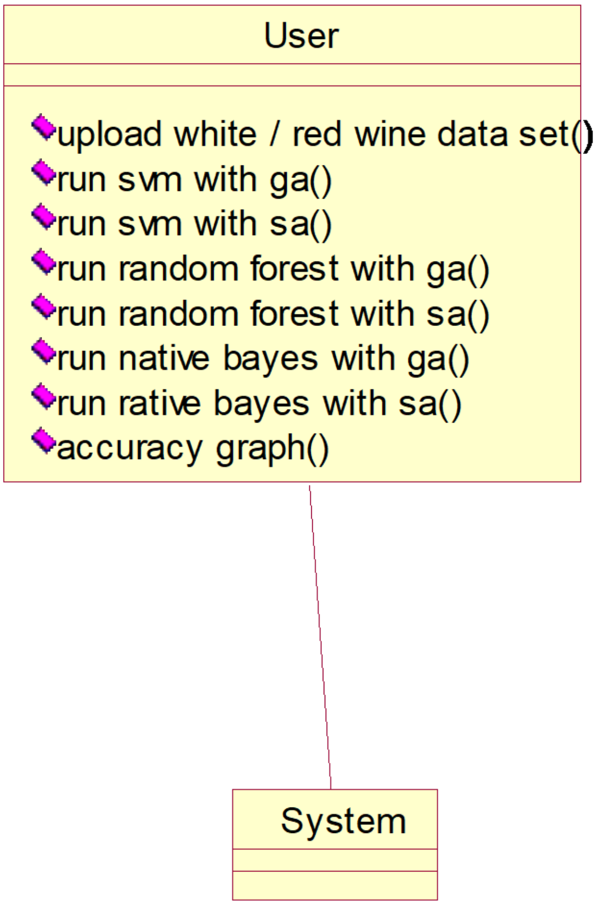
ii. SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



iii. CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

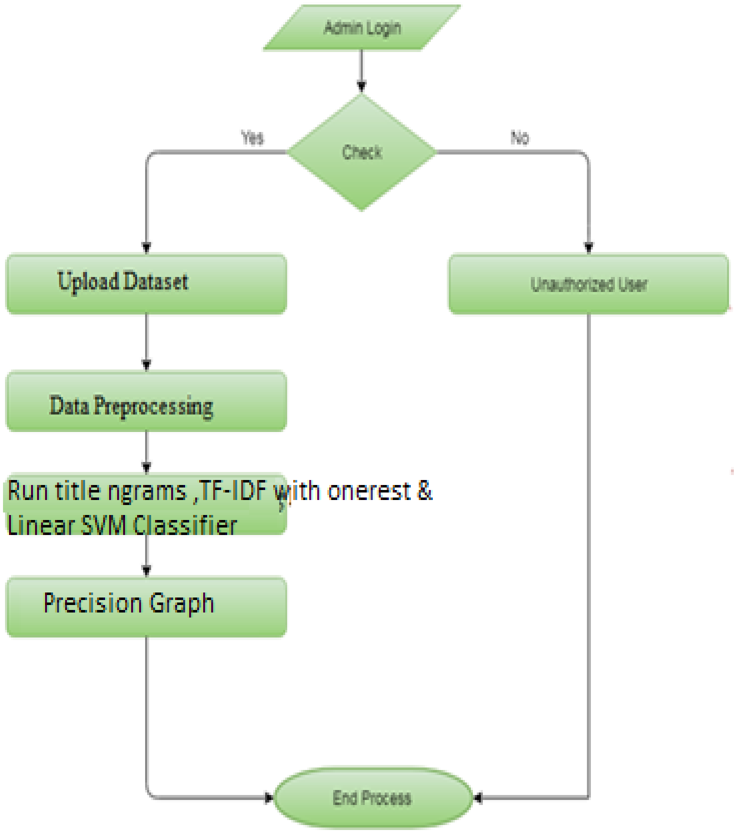

User
upload white / red wine data set()
run svm with ga()
run svm with sa()
run random forest with ga()
run random forest with sa()
run native bayes with ga()
run rative bayes with sa()
accuracy graph()
System

iv. Data Flow diagram :-

Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation.

Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow..

DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchy of detailed diagrams. DFD has often been used due to the following reasons:



v.Component Diagram :-

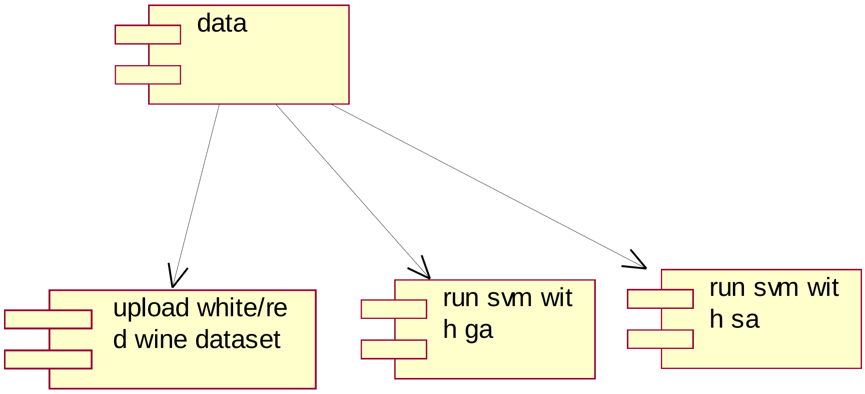
Component diagram is a special kind of diagram in UML. The purpose is also different from all other diagrams discussed so far. It does not describe the functionality of the system but it describes the components used to make those functionalities.

Thus from that point of view, component diagrams are used to visualize the physical components in a system. These components are libraries, packages, files, etc.

Component diagrams can also be described as a static implementation view of a system. Static implementation represents the organization of the components at a particular moment.

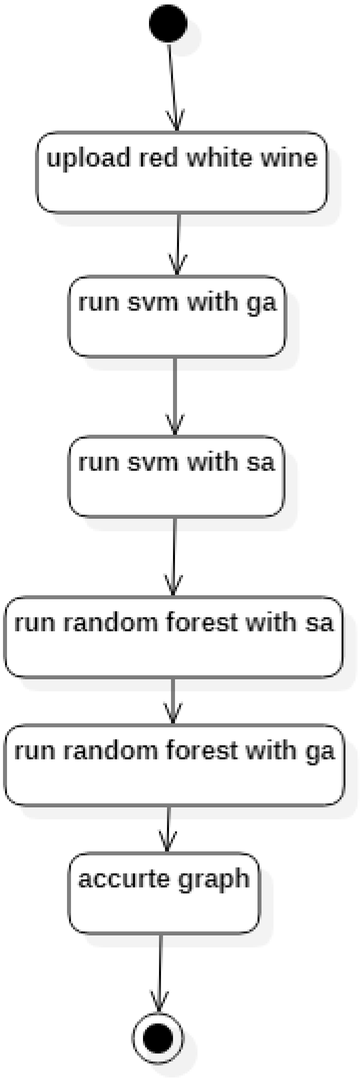
A single component diagram cannot represent the entire system but a collection of diagrams is used to represent the whole.

UML Component diagrams are used in modeling the physical aspects of object-oriented systems that are used for visualizing, specifying, and documenting component-based systems and also for constructing executable systems through forward and reverse engineering. Component diagrams are essentially class diagrams that focus on a system's components that often used to model the static implementation view of a system.



vi.ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



CHAPTER – 3

TEST RESULTS

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the
Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application
.it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must
be accepted. Invalid Input : identified classes of invalid
input must be rejected. Functions : identified functions
must be exercised.
Output : identified classes of application outputs must
be exercised. Systems/Procedures: interfacing systems or procedures
must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot –seel into it. The test provides inputs and responds to outputs without considering how the software works.

3.1Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

3.2 Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g., components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

3.3Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER-4

CONCLUSION

This paper mentioned about potential of genetic algorithm as well as simulated annealing algorithm for feature selection as well as the potentials of the classifiers to predict accurately based on the new feature sets. The feature selection algorithm provided a clear idea about the importance of the attributes for prediction of quality, which was time consuming and expensive when done in the traditional way. We have also compared the performance metrics of linear, nonlinear, and probabilistic based classifiers and it was found that these classifiers performed well with the new feature sets. We have found that the SA based feature sets performed better than the GA based feature sets. We have also found that the SVM classifier performed better compared to all other classifiers for red wine and white wine data sets. We have found accuracy ranging from 95.23% to 98.81% with different feature sets. In future we can try other performance measures and other machine learning techniques for better comparison on results. This analysis will help the industries to predict the quality of the different type of wines based on certain attributes and also it will helpful for them to make good product in the future.

References

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the 7th International Conference on World Wide Web, 1998.

[1] J. Furnkranz and E. Hullermeier. Preference Learning: A Tutorial Introducton, DS 2011, Espoo, Finland, Oct 2011.

[2] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[J]. 2003.

[3] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.

[4] Swingler K. Applying neural networks: a practical guide[M].Morgan Kaufmann, 1996.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In Proceedings of the 2010 International AAAI Conference on Weblogs and Social Media, 2010.

[7] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google's pagerank algorithm. Journal of Informetrics, 1(1), 2007.

[8] D. J. Cook and L. B. Holder. Mining Graph Data. John Wiley & Sons, 2006.

[9] B. S. Demoll and D. Mcfarland. The Art and Science of Dynamic Network Visualization. Journal of Social Structure, Volume 7, 2005.

[10] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. Journal of the American Society for Information Science and Technology, 60(11), 2009.

[11] G. Dutton. Improving locational specificity of map data - a multi-resolution, metadata-driven approach and notation. International Journal of Geographical Information Science, 10(3), 1996.

[12] D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010.