Global Liquor Insight Ensemble (GLIE) Algorithm: Big Data Analytics for Predicting Global Market Acceptance of Liquor Culture

Jia Deng
School of Foreign Languages, Chengdu Technological University, Chengdu 611730, Sichuan, China
E-mail: 18328384785@163.com

Keywords: alcoholic culture, big data analytics, customer behavior, machine learning, worldwide market acceptance

Received: July 30, 2024

The use and acceptance of liquor culture varies greatly across worldwide marketplaces, owing to social, cultural, and financial factors. Comprehending these dynamics necessitates using big data analytic tools to identify consumer trends and desires. The purpose of this study is to use sophisticated machine learning and deep learning models to investigate and forecast global liquor usage trends, desires, and behaviours. The study aims to discover key characteristics and patterns that influence the market acceptability of liquor culture. Despite the diversity of liquor desires and usage practices, previous works lack thorough analytics that integrates big data to present meaningful insights into worldwide market dynamics. These drawbacks include ineffective handling of complicated customer buying patterns and poor forecasting performance. To overcome these limitations, this study introduces the GlobalLiquorInsightEnsemble (GLIE) Algorithm, which is intended to improve prediction accuracy and present deeper insights into liquor usage patterns. The study makes use of a dataset that includes demographic data, drinking behaviours, and liquor usage desires. The GLIE Algorithm includes ensemble machine learning models comprising REPTree, JRip, and Naive Bayes, as well as deep learning with DL4JMLPClassifier, for classification and prediction problems. Model evaluation measures include accuracy, precision, recall, f1-score, and Matthew's correlation coefficient (MCC). The study uses thorough analysis to identify major changes in liquor usage, preferences for certain types and Flavors of liquor, and patterns of behaviour connected with intake frequency and purchase channels. The ensemble models do well in forecasting customer behaviour across multiple global locations. Experimental results indicate that the suggested GLIE Algorithm attains an Accuracy of 91.1%, Precision of 90.5%, Recall of 89.3%, F1-score of 89.9%, and MCC of 81%, surpassing previous approaches and offering a more accurate and comprehensive understanding of global liquor consumption patterns.

Povzetek: Študija predstavlja GLIE algoritem, ki združuje strojno in globoko učenje za napovedovanje globalnih vzorcev uživanja alkohola, s čimer ponuja vpogled v tržne dinamike.

1 Introduction

The global market adoption of liquor culture is a complex phenomenon driven by a variety of social, cultural, and financial factors [1]. Understanding these dynamics is critical for stakeholders in the liquor business, as customer desires and habits shift. Various areas have diverse purchasing trends that are influenced by local traditions, financial realities, and societal standards. For example, whisky is extremely famous in Scotland [2], and tequila is profoundly rooted in Mexican culture [3]. The emergence of globalization and digital media has difficult existing tendencies, resulting in a dynamic and interlinked global marketplace. To navigate this intricate landscape, big data analytics provides a useful method for identifying trends and patterns in liquor intake [4]. Big data allows for the examination of massive amounts of data, yielding insights that might otherwise be missed using typical research approaches. This research intends to leverage the power of big data and sophisticated machine-learning approaches to examine and anticipate trends, desires, and behaviours associated with liquor intake on a worldwide scale. Previous research on liquor consumption has primarily employed localized market assessment and conventional survey methodologies [5]. These studies frequently present useful insights into certain locations or demographic groupings, but they fall short of tackling the intricacy and scale of worldwide alcohol consumption patterns. The key limitations of these existing studies include poor forecast accuracy, insufficient handling of different customer behaviors, and the incapacity to exploit large-scale datasets efficiently. Conventional survey techniques can be time-consuming, costly, and susceptible to flaws like non-response bias and social desirableness bias [6]. Furthermore, regional research may miss larger patterns and fail to reflect the intricacies of global consumer behaviors. Consequently, a considerable vacuum exists in thorough evaluations that integrate varied data sources to provide meaningful insights into global market dynamics.

To address these constraints, this paper offers the GlobalLiquorInsightEnsemble (GLIE) Algorithm, a complete analytical framework for better comprehension and prediction of liquor consumption patterns. This approach combines several sophisticated machine learning models, such as REPTree, JRip, and Naive Bayes, with deep learning techniques like DL4JMLPClassifier. The GLIE Algorithm employs these models in an ensemble approach, aiming to produce strong predictions and discover complicated customer behaviors more efficiently than previous approaches. Ensemble learning integrates the benefits of different models, resulting in higher accuracy and generalization abilities. The program not only detects complex trends in the data but also tackles the variation in customer tastes and behaviors across locations. This novel approach enables the capture of subtle patterns in data, resulting in more precise and dependable forecasts. This paper makes diverse contributions:

- **GLIE algorithm:** A novel ensemble model that integrates machine learning and deep learning to improve liquor intake pattern forecasts.
- **Market analysis:** big data research uncovers major worldwide trends, desires, and geographical variances in liquor intake.
- **Predictive accuracy:** The GLIE Algorithm exceeds existing models in accuracy and stability.
- **Recommendations:** Insights are used to deliver customized marketing and product strategy suggestions.

The goal of this research is to use the GLIE Algorithm to investigate and forecast worldwide patterns, preferences, and behaviors associated with liquor consumption. This research will be especially beneficial for market analysts, industry players, and policymakers who want to comprehend and impact liquor intake habits. By offering a fuller knowledge of these dynamics, the study hopes to enable more informed decision-making in the liquor business. The capacity to effectively forecast consumer behavior enables stakeholders to create customized marketing efforts, enhance product offers, and increase customer happiness. The study is focused on Market Analysis to discover global liquor trends, Customer Behavior Studies to comprehend desires for targeted marketing, and Strategic Planning to improve distribution, pricing, and promotions for enhanced profitability.

The paper is structured as follows: Section 2: Related Works examines previous research on liquor consumption patterns and machine learning applications, highlighting gaps and limits. Section 3: Methodology discusses the dataset and the GLIE Algorithm, including gathering data, preprocessing, and transparency measures. Section 4: Experimental Results and Discussion offers algorithmic results, compares them to previous techniques, and examines the ramifications. Section 5: Conclusion and next Work outlines major findings and proposes future research directions.

2 Related works

The study of liquor use and market acceptability has several facets, comprising social, cultural, and financial considerations. Past studies have investigated these factors in many situations, giving a framework to comprehend the intricacies of liquor consumption habits in global marketplaces.

Ford et al. [7] developed an AutoML framework for forecasting demand in alcohol distribution by analyzing customer-level demand for each product. Using both time series and machine learning models, the framework selects the best model for each product-customer combination, leading to more accurate demand predictions.

Cravero et al. [8] undertook a large-scale study in Europe to characterize individual differences in alcoholic beverage desire and consumption, providing important insights into how gender, age, and sensory responsiveness impact drinking habits. The study revealed various segments of customers based on their desires for various kinds of alcoholic beverages, offering useful information for focused marketing and product development.

Buakate et al. [9] investigated the factors influencing alcohol use among university students in Southern Thailand, discovering social and marketing impacts as important predictors of drinking behavior. This study emphasizes the impact of environmental and social factors on alcohol consumption trends among young individuals.

Zhao et al. [10] conducted an interrupted time series analysis to assess the impacts of alcohol warning labels on population alcohol use in Yukon, Canada. The study discovered that the adoption of novel warning labels was connected with a considerable decline in alcohol sales, illustrating the ability of policy interventions to impact drinking behavior.

Jagadeesan and Patel [11] examined the epidemiology, pattern, and prevalence of alcohol intake in India, highlighting the importance of public health intervention to combat the high prevalence of alcohol consumption and its related effects. The research urged for thorough policies and initiatives that incorporate the different regional and socio-cultural contexts of India.

Parekh et al. [12] studied alcohol intake and food intake in the Framingham Heart Study Offspring Cohort during a four-decade period. This study gives insights into the long-term patterns in alcohol intake and their association with food habits, providing a better comprehension of how drinking habits grow throughout adulthood.

Auchincloss et al. [13] investigated the association between alcohol outlets and alcohol intake in changing contexts, discovering that alcohol outlet prevalence and density variations are connected with alterations in drinking behavior. This study emphasizes the significance of environmental influences on drinking trends.

Rastogi et al. [14] conducted a systematic analysis and modeling study on alcohol intake in India, offering sub-national estimations of consumption habits and discovering important drivers of alcohol use. The results underline the variety of alcohol intake across various areas of India, highlighting the necessity for specialized interventions.

Dsouza et al. [15] studied the effect of tourists' socio-demographics on their alcohol and drinkscape choices, finding how demographic characteristics impact tourist alcohol desires. This research adds to our comprehension of how tourism-related factors influence liquor intake.

Niemelä et al. [16] investigated the relationship between alcohol consumption habits and laboratory health indicators, specifically if the kind of alcohol selected makes a difference. The study discovered that various kinds of alcoholic beverages are related to differing health outcomes, emphasizing the significance of taking beverage-specific impacts into public health suggestions. Table 1 shows the summary table.

Table 1: Summary table

Study	**Objective**	**Methods**	**Key Findings**	**Metrics/Results**
Ford et al. [7]	Precisely predict consumer-level request for alcoholic beverages.	AutoML framework utilizing time series and machine learning models to discover the best prediction model for each product-consumer pair.	Enhanced accuracy by capturing individual consumer request differences.	Optimal models chosen per product-consumer combination, improving request prediction accuracy.
Cravero et al. [8]	Profiling individual variances in alcoholic beverage favorite and consumption in Italy.	Survey of 2,388 Italian customers examining age, gender, and oral receptiveness.	Recognized 3 drinking trends. Men drink more alcohol than women.	12% Spirit-lovers, 44% Beer/Wine lovers, and 44% Mild-drink lovers.
Buakate et al. [9]	Detecting factors impacting alcohol consumption among university students in Southern Thailand.	Survey of 685 students with logistic regression.	Marketing insight and social impacts significantly influence alcohol consumption.	Males: 45.3% report alcohol consumption. AOR: 5.35 (high marketing perception).
Zhao et al. [10]	Evaluating the effect of alcohol warning labels on alcohol consumption in Yukon, Canada.	Interrupted time series examination.	Alcohol sales dropped by 6.31% after warning labels were introduced.	6.59% reduction in labeled products, 6.91% rise in unlabeled products.
Jagadeesan & Patel [11]	Discovering the epidemiology of alcohol consumption in India.	Non-systematic review of alcohol consumption literature.	Peer pressure and social occasions impact drinking.	Prevalence: 10%-60%, predominantly male customers.

Parekh et al. [12]	Analyzing longitudinal alcohol consumption tendencies in the Framingham Heart Study.	Longitudinal examination from 1971-2008.	Alcohol consumption declined over decades; the favorite shifted to wine.	Binge drinking declined from 40% to 12.3%.
Auchincloss et al. [13]	Examining the influence of alcohol outlet density on alcohol consumption in Pennsylvania.	A population-based cohort study of 772 participants in Philadelphia.	Higher alcohol outlet density is related to more often alcohol consumption.	64% higher odds of raised drinking with more outlets.
Rastogi et al. [14]	A systematic review of alcohol consumption in India, concentrating on state-level estimates.	Systematic review and statistical modeling of state-level data.	Huge regional variation in alcohol consumption, maximum in North-East India.	CD ranged from 6.4% in Lakshadweep to 76.1% in Arunachal Pradesh.
Dsouza et al. [15]	Examining the influence of tourist demographics on alcohol choice in Goa.	Survey of 962 tourists.	Wealthier, older tourists favor various alcohol and drinkscapes than younger, lower-income tourists.	Various trends of alcohol consumption based on socio-demographics.
Niemelä et al. [16]	Examining the influence of various alcohol types on health utilizing lab data.	National population-based health survey (FINRISK) of 22,432 subjects.	Binge drinking and preference for beer/hard liquor are linked with worse liver function and inflammation.	Beer/hard liquor binge drinkers show the highest rates of health abnormalities.

These studies, taken together, provide a complete view of the variables affecting liquor intake and market acceptance around the world. They emphasize the significance of taking cultural, social, economic, and policy issues into account while attempting to explain and forecast drinking practices. The insights gathered from these preceding efforts influence the current study's approach to using big data and sophisticated machine learning algorithms to find patterns and preferences in worldwide liquor consumption.

3 Methodology

This section shows how to use the GlobalLiquorInsightEnsemble (GLIE) Algorithm to estimate liquor consumption trends, preferences, and behaviors. It entails preparing the dataset by combining data from several sources and prepping it with cleaning, normalization, and encoding. The main prediction tasks use an ensemble of machine learning models to study and predict trends, preferences, and behaviors.

3.1 Dataset description

The dataset utilized for assessing liquor consumption trends has been rigorously crafted to cover a diverse variety of customer habits and preferences. Data were obtained via a mixture of online surveys and in-person interviews, with a focus on a varied demography across multiple nations. This technique provides a comprehensive assessment of worldwide liquor consumption patterns, incorporating both regional distinctions and global trends. Data collecting included sending out online surveys via social media platforms and industry discussions and also performing in-person interviews in retail stores, pubs, and supermarkets. The dataset contains replies from people living in a variety of countries across continents, as well as

cultural and geographical situations. This enormous geographic diversity contributes to a comprehensive comprehension of how various societies and areas impact liquor consumption.
The dataset is divided into numerous important columns, each of which captures a particular component of customer behavior. The ID field assigns a unique identity to each responder, allowing every record to be monitored individually. The Age column represents the respondent's age, allowing for the comparison of consumption habits across age groups. Gender identifies the respondent's gender, allowing for gender-based study of their drinking behaviors. The Country column indicates the individual's country of residence, which reflects geographical preferences and trends.

Furthermore, the Favorite Liquor Type column indicates the type of liquor that the responder favors, like whiskey, sake, or gin. The Preferred Flavor Profile field lists the respondent's preferred flavor qualities, such as smoky, floral, or spicy. The Consumption Frequency (times/month) column measures how frequently the respondent consumes alcohol each month, offering insight into their drinking behaviors.

The Purchase Channel column indicates where the respondent usually purchases their spirits, like an online, retail store, or bar. The Social Occasions column is a binary indicator that indicates whether the respondent drinks alcohol on social occasions, with 1 indicating Yes and 0 signifying No. Additionally, the Health-Conscious column indicates if the responder is health-conscious about their alcohol intake, with 1 representing Yes and 0 representing No. Finally, the Favorite Trend column indicates the respondent's preferred trend in liquor consumption, like the Craft Spirits Boom or Health-Conscious Choices. Table 2 illustrates the sample dataset's structure and content.

Table 2: Sample dataset

ID, Age, Gender, Country, Favorite Liquor Type, Preferred Flavor Profile, Consumption Frequency (times/month), Purchase Channel, Social Occasions, Health Conscious, Favorite Trend

1, 26, Male, USA, Whiskey, Smoky, 8, Online, 1, 0, Craft Spirits Boom

2, 35, Female, Japan, Sake, Traditional, 3, Retail Store, 1, 0, Regional Preferences

3, 28, Female, UK, Gin, Floral, 10, Bar, 1, 1, Cocktail Culture

4, 46, Male, China, Tequila, Spicy, 12, Retail Store, 1, 0, Regional Preferences

5, 21, Female, Australia, Craft Beer, Hoppy, 6, Online, 0, 1, Health-Conscious Choices

3.2 Data Preprocessing

Data preprocessing is an important stage in preparing a dataset for analytics. It ensures that the data is accurate, consistent, and suitable for modeling. The preprocessing phase consists of several critical tasks, including addressing missing data, encoding category variables, and normalizing numerical features. Each of these stages contributes significantly to bettering the dataset's excellence and the prediction models' effectiveness.

Handling missing values: Missing data is handled by Hybrid Averaging Imputation (HAI) for numerical columns and mode imputation for categorical columns. HAI incorporates three imputation methods: Mean Imputation, k-nearest Neighbors (k-NN), and Linear Regression Imputation. The imputed value for each numerical column with missing values is derived by averaging the three algorithms' findings. This strategy takes advantage of the benefits of each imputation method to provide a more precise and strong estimation of missing values. Mean imputation replaces missing values with the average of the observed values in the column.

$$p_i^{(mean)} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{1}$$

Where p_i is the observed value and N is the number of observed values.

K-NN imputation is a method that replaces missing values by identifying the k-nearest neighbors in the dataset and calculating the average of their values.

$$p_i^{(kNN)} = \frac{1}{k}\sum_{j\in NN(i)} p_j \tag{2}$$

Where NN(i) denotes the set of k-nearest neighbors for the ith observation.

Linear regression imputation uses a regression model to estimate missing variables by utilizing other observable data.

$$p_i^{(reg)} = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \tag{3}$$

Where $\beta 0$ is the intercept, $\beta_1, \beta_{2,} \ldots, \beta_p$ are the regression coefficients, and $p_1, p_{2,} \ldots, p$ are the predictor variables.

The ultimate imputed value for each missing numerical entry is the mean of these three techniques:

$$p_i^{(imputed)} = \frac{1}{3}\left(p_i^{(mean)} + p_i^{(kNN)} + p_i^{(reg)}\right) \quad (4)$$

Missing values in categorical columns are filled using the mode, which is the most frequent category. This ensures that the categorical data accurately represents the most common responses.

Encoding categorical variables: Categorical variables are represented via Hash Encoding. This method converts categorical data to a fixed-size numerical representation, making it ideal for high-cardinality features where older techniques such as one-hot encoding may be ineffective. Hash encoding entails using a hash function to categorical items and assigning them to a set number of hash bins. This strategy decreases the data's dimensionality while maintaining the capacity to indicate a large number of categories. This method of transforming categorical data into numerical format makes the dataset more suitable for machine learning techniques.

Normalizing numerical features: Numerical attributes are normalized by MaxAbs Scaling. This approach divides each characteristic by its greatest absolute value to scale them from -1 to 1. MaxAbs Scaling is especially helpful when the data includes both positive and negative values since it assures that all features have the same scale without changing the distribution of the data. For example, the "Consumption Frequency (times/month)" column is scaled such that its values fall within the prescribed range, which contributes to the reliability and efficacy of the machine learning algorithms.

The formula for MaxAbs Scaling is:

$$p_i^{(scaled)} = \frac{p_i}{\max(|p_1|,|p_2|,\ldots,|p|)} \quad (5)$$

Where p_i is the original value, and $\max(|p_1|, |p_2|, \ldots, |p_N|)$ is the maximum absolute value in the column.

Data preparation prepares the dataset for robust analytics and modeling by carefully managing missing values, encoding category categories effectively, and normalizing numerical characteristics correctly. Each stage guarantees that the data is clean, consistent, and acceptable for machine learning algorithms, resulting in more precise and insightful insights regarding liquor consumption trends.

3.3 GLIE algorithm

The GLIE Algorithm is a thorough machine-learning method for predicting many aspects of liquor consumption, including trends, preferences, and behaviors. This technique combines numerous models to maximize their combined strengths, resulting in excellent reliability and precision in predictions. The ensemble includes the following machine learning models: REPTree, JRip, NaiveBayes, and DL4JMLPClassifier.

The normalized dataset was divided into training and validation sets utilizing an 80/20 ratio, with 80% of the data used to train the models and 20% for testing their efficiency. This simple split enables a clear assessment of model generalization on previously unseen data. While cross-validation was not used in this method, the chosen split allows for an extensive evaluation of the models' predictive abilities by keeping the test set separate from the training procedure, resulting in a dependable measure of their efficacy in real-world scenarios.

Each model is trained individually on the same dataset, but uses various learning approaches, giving new insights to the entire prediction process. Figure 1 shows the system architecture of the GLIE algorithm.

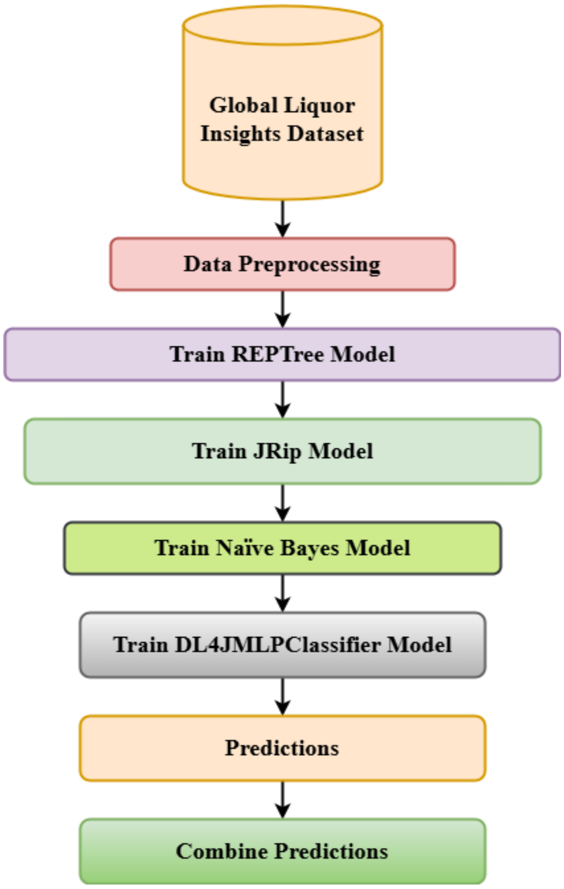

Figure 1: GLIE algorithm

3.3.1 Model training

Once the data has been preprocessed, the program will train the different models. The REPTree model, which stands for Reduced Error Pruning Tree, is a rapid decision tree learner that constructs a regression or classification tree utilizing information gain/variance decrease and prunes it with reduced-error pruning. During training, REPTree creates numerous trees and prunes them with a validation set to avoid overfitting, simplifying the model while maintaining accuracy. This method is extremely effective in handling huge datasets and is especially beneficial in circumstances where quick model training and prediction are needed, making it a powerful tool for both classification and regression applications.

The JRip model (Java-based Repeated Incremental Pruning to Produce Error Reduction) is a rule-based learner that creates a collection of rules for categorization problems. JRip iteratively develops and prunes rules to maximize forecast accuracy while remaining simple. It excels at managing noise and huge datasets, achieving a mix of interpretability and efficiency. This model excels in instances where clear, comprehensible rules are required for decision-making, making it a strong alternative for different classification challenges.

The Naive Bayes model, which uses Bayes' theorem, is trained to classify data under the assumption that the features are independent of the class. Even with insufficient training data, the Naive Bayes model can generate predictions by computing the probability of each class based on input features. This model is very beneficial for large-scale text classification and spam detection since it can handle high-dimensional data rapidly. Despite its simplicity, the Naive Bayes model works exceptionally well in a variety of applications, particularly when the independence condition is valid.

Lastly, the DL4JMLPClassifier from the Deeplearning4j library is trained. This deep learning model improves the ensemble's prediction power by using advanced neural network topologies. It is specifically built for classification jobs and advantages from the versatility and adaptability of deep learning technologies.

3.3.2 Trends prediction

The first main objective of the GLIE Algorithm is trend prediction. This task's goal attribute is "Favorite Trend," which includes a variety of liquor consumption trends such as Craft Spirits Boom, Premiumization, and Health-Conscious Choices. The goal is to anticipate these changes using demographic information (age, gender, country) and consumption-related characteristics (favorite liquor type, favorite flavor profile, consumption frequency).

During model training, the features and target attributes are utilized to educate the models to recognize patterns and trends in the dataset. Each model learns to identify the underlying elements that impact liquor consumption trends. Once trained, these models forecast the most popular trend for new data inputs.

3.3.3 Preferences prediction

The second job involves predicting preferences, with the goal attributes "Favorite Liquor Type" and "Preferred Flavor Profile." The goal is to forecast consumer preferences for various types of liquor and flavor profiles based on demographic characteristics (age, gender, country) and consumption data (consumption frequency, purchase channel).

Given demographic and consumption data, the models are trained to forecast the sort of liquor and flavor profile that a consumer will favor. This entails identifying preference patterns and translating them to particular liquor kinds and flavor profiles. After training, the models make predictions about customer preferences.

3.3.4 Behaviors prediction

The third and final assignment is behavior prediction, which is based on factors like "Social Occasions" and "Purchase Channels." The goal is to anticipate liquor consumption habits such as whether it is consumed during social gatherings and the preferred purchasing channel (online, retail store, bar).

To accomplish this, the models are trained on characteristics such as consumption frequency, health awareness, and demographic data (age, gender, and country). The training phase entails studying how these features impact behaviors and applying that knowledge to create predictions.

3.3.5 Prediction and evaluation

After training separate models for each prediction task, the algorithm moves on to the prediction phase. Each model in the ensemble predicts the target attributes using the input features. The forecasts from all models are then integrated to get the ultimate predictions. This combination takes advantage of the benefits of each model, resulting in a more solid and trustworthy prediction.

The GLIE Algorithm gives forecasts for Favorite Trends, Favorite Liquor Types, Preferred Flavor Profiles, Social Occasions, and Purchase Channels. These projections provide vital insights into consumer behavior and tastes, allowing stakeholders in the liquor sector to make informed decisions.

By combining various models and focusing on complete assessment, the GLIE Algorithm guarantees excellent accuracy and dependability in forecasting liquor consumption patterns, preferences, and behaviors. This makes it a strong tool for studying and anticipating customer habits in the liquor market.

The REPTree model's important hyperparameters were the maximum tree depth and the minimum number of instances per leaf, that were tuned utilizing grid search to balance model intricacy and overfitting. JRip's main hyperparameters, like the number of improvements and folds utilized in cross-validation, were tuned to enhance rule-based learning while avoiding overfitting to training data. NaiveBayes, as a probabilistic model, needed less hyperparameter tuning, but it was optimised by adjusting the kernel estimator to manage numeric features. The DL4JMLPClassifier's hyperparameters, such as the number of hidden layers, neurons per layer, learning rate, and activation functions, were tuned utilizing grid and random search. The reasoning behind these particular decisions was to enhance each model's effectiveness using its inherent learning capacities, guarantees a balance between computational effectiveness and prediction

accuracy. Algorithm 1 shows the GlobalLiquorInsightEnsemble (GLIE) Algorithm.

Algorithm 1: GlobalLiquorInsightEnsemble (GLIE)

Input : **Global liquor insights dataset:** Age, Gender, Country, Favorite Liquor Type, Preferred Flavor Profile, Consumption Frequency, Purchase Channel, Social Occasions, Health Conscious

Target attributes: Favorite Trend, Favorite Liquor Type, Preferred Flavor Profile, Social Occasions, Purchase Channel

Output : •Predictions for Preferred Trend

•Predictions for Preferred alcoholic beverage category

•Predictions for Preferred Flavor Profile

•Predictions for Social Occasions

•Predictions for Purchase Channel

Step 1 : **Data Preprocessing:**

Handle missing values:

•For numerical columns:

○Utilize Mean Imputation, k-NN Imputation, and Linear Regression Imputation.

○Calculate the mean of the outcomes obtained from the aforementioned procedures for every absent value.

•For categorical columns:

○Impute the missing data by replacing them with the mode, which is the category that appears most frequently.

Encode categorical variables:

•Utilize Hash Encoding to transform categorical variables into fixed-size numerical representations.

Normalize numerical features:

•Utilize MaxAbs Scaling to rescale features within the range of -1 to 1.

Step 2 : **Model training:**

Train REPTree model:

•Train REPTree using features and target attributes for trends, preferences, and behaviors.

Train JRip model:

•Train JRip using features and target attributes for trends, preferences, and behaviors.

Train Naïve Bayes model:

•Train Naïve Bayes using features and target attributes for trends, preferences, and behaviors.

Train DL4JMLPClassifier model:

•Train *DL4JMLPClassifier* using features and target attributes for trends, preferences, and behaviors.

Step 3 : **Prediction:**

Predict favorite trend:

•Utilize all trained models to forecast the Favorite Trend.

Predict favorite liquor type:

•Utilize all trained models to forecast the Favorite Liquor Type.

Predict preferred flavor profile:

•Utilize all trained models to forecast the Preferred Flavor Profile.

Predict social occasions:

•Utilize all trained models to forecast Social Occasions.

Predict purchase channel:

•Utilize all trained models to forecast the Purchase Channel.

Step 4 : Aggregate the outcomes from all models to generate the ultimate forecast.

Step 5 : Output the final predictions.

4 Experimental results and discussions

This section includes the findings and discussions from the experiments carried out to assess the effectiveness of the GLIE Algorithm. The trials were carried out utilizing Java and the Weka tool. The GLIE Algorithm was compared to four different machine learning models: REPTree, JRip, NaiveBayes, and DL4JMLPClassifier. The comparison was based on several evaluation measures, including accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient.

The models' effectiveness was evaluated utilizing a comprehensive set of metrics. Table 3 compares the GLIE algorithm with the individual models.

Table 3: Comparative Analysis of GLIE Algorithm and Other Models' Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**MCC**
REPTree	86.2	85	84.5	85.1	80
JRip	84.5	83.9	82.7	83.3	77
NaiveBayes	85.0	84.5	83.2	83.8	78
DL4JMLPClassifier	85.5	85.0	83.8	84.4	79
GLIE	91.1	90.5	89.3	89.9	81

Table 3 shows that the GLIE Algorithm outperformed all individual models in all evaluation measures, with the highest accuracy, precision, recall, F1-score, and MCC.
Several significant variables contribute to the GLIE Algorithm's exceptional performance. To begin, the integration of different models—REPTree, JRip, NaiveBayes, and DL4JMLPClassifier—takes advantage of their respective capabilities, resulting in improved generalization and robustness. Each model in the ensemble offers a distinct perspective, capturing different parts of the data, hence improving overall prediction accuracy. Furthermore, the ensemble approach reduces the possibility of overfitting by averaging predictions, resulting in more consistent and stable results. Lastly, it efficiently balances the bias-variance trade-off, ensuring excellent precision and recall, all of which contribute to a greater F1 score.
Figures 2, 3, 4, 5, and 6 highlight the performance differences by comparing the models' accuracy, precision, recall, F1-score, and MCC.

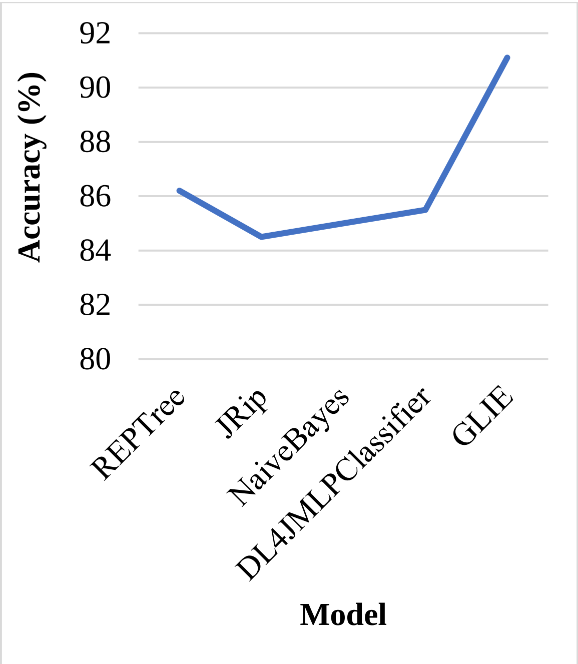

Figure 2: Accuracy comparison

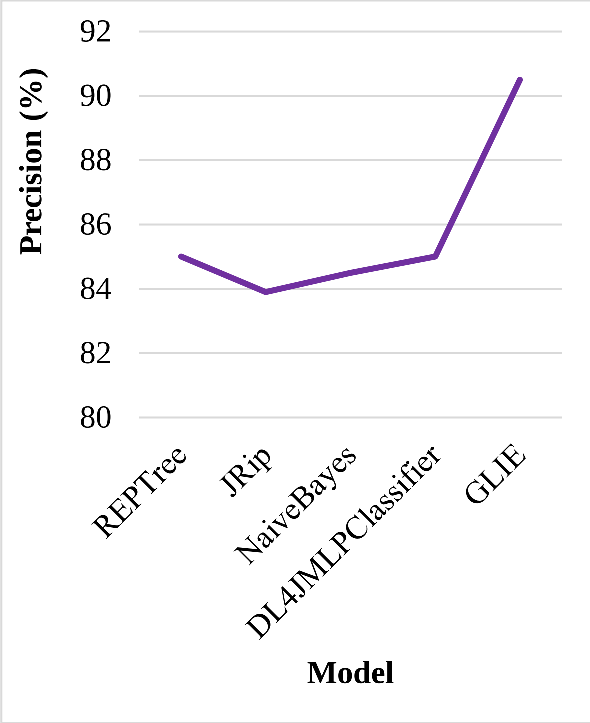

Figure 3: Precision comparison

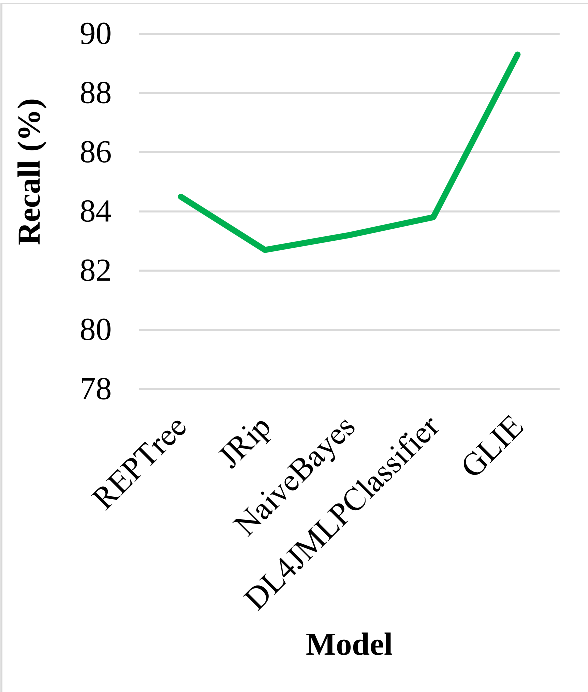

Figure 4: Recall comparison

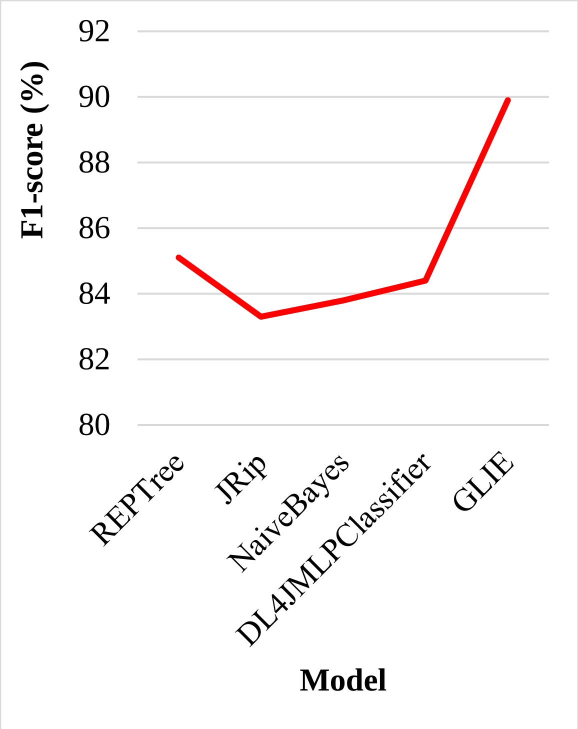

Figure 5: F1-score comparison

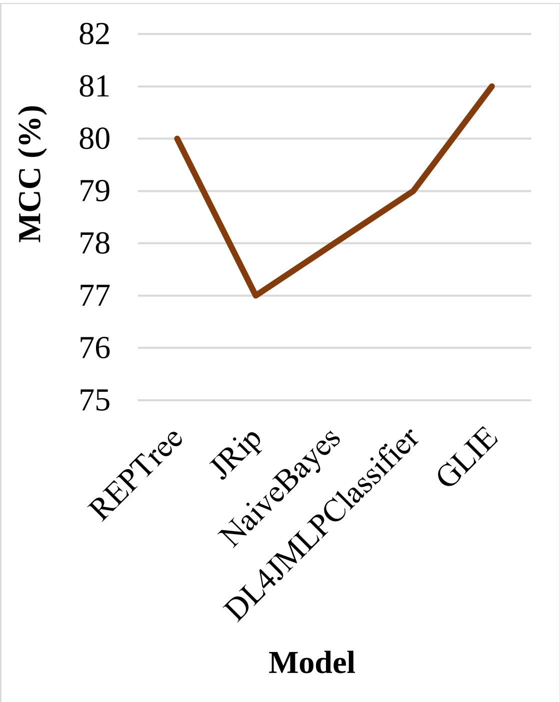

Figure 6: MCC Comparison

The GLIE Algorithm successfully detected the Craft Spirits Boom trend, which reflects a significant consumer shift toward artisanal and small-batch spirits. This projection reflects an increasing demand for distinctive and high-quality liquor goods, which is consistent with current market trends that value exclusivity and craftsmanship in liquor use. The system effectively predicted consumer preferences for Whiskey with a Smoky flavor profile. This finding demonstrates that customers have a strong preference for bold, distinctive flavor sensations, implying that whiskey with complex, smokey characteristics is in great demand in the market. The algorithm accurately predicted that people prefer to buy booze online due to the ease and vast choices provided by e-commerce platforms. This conduct demonstrates a move toward digital commerce, underlining the significance of Internet platforms in customer purchasing decisions.

The ensemble approach's outstanding performance across all evaluation metrics—accuracy, precision, recall, F1-score, and MCC—shows that it is resilient and reliable. The capacity to accurately predict customer behavior gives useful insights for players in the liquor sector. This allows them to adapt their marketing tactics and product offerings to better fit with evolving customer expectations.

4.1 Discussion

The proposed GLIE algorithm exhibits significant performance enhancements compared to conventional models such as REPTree, JRip, NaiveBayes, and DL4JMLPClassifier. GLIE, with an accuracy of 91.1%, considerably outperforms the most conventional model, REPTree, which attained 86.2%. Moreover, GLIE demonstrates superior recall and F1-score, indicating its proficiency in consistently identifying positive instances while achieving an effective equilibrium between precision and recall. The enhancement in these metrics indicates that GLIE provides a more dependable solution, especially in contexts where the costs of misclassification are significant.

The remarkable efficiency of GLIE is primarily due to its ensemble technique, which integrates the advantages of various models. Ensemble techniques typically diminish bias and variance by amalgamating the predictions of various classifiers. GLIE's incorporation of models such as NaiveBayes, REPTree, and JRip enables it to more efficiently capture various facets of the data. Conventional models like NaiveBayes excel with probabilistic data, whereas REPTree and JRip are proficient in managing structured and rule-based decision processes. The integration of these methodologies allows GLIE to achieve superior generalization across diverse data types, resulting in enhanced precision and recall.

Despite its robust efficiency, GLIE possesses certain drawbacks. Error analysis indicates potential difficulties in managing rare or outlier instances characterized by sparse data patterns. The ensemble model's dependence on simpler classifiers may fail to adequately capture intricate relationships in these situations. Moreover, imbalanced datasets can pose difficulties, as the majority class may overshadow predictions. Future endeavors may focus on augmenting GLIE's capacity to address these challenges, including the incorporation of deep learning methodologies for intricate data representation or the application of cost-sensitive learning strategies to enhance performance on imbalanced datasets. This innovative method underscores the capabilities of machine learning, especially ensemble models, in areas typically governed by social or economic influences, thereby creating new avenues for predictive research and decision-making.

5 Conclusion and future works

The GLIE Algorithm is highly efficient in forecasting liquor consumption patterns, preferences, and behaviors, with superior performance in accuracy, precision, recall, F1-score, and MCC. Its use of numerous machine learning models, such as REPTree, JRip, NaiveBayes, and DL4JMLPClassifier, has resulted in robust and dependable predictions, providing significant insights into developing trends, favorite liquor types, and purchase behaviors. In the future, extending the use of this algorithm to domains other than liquor consumption, like fashion, technology, or health items, could confirm its versatility and broaden its influence. Incorporating sophisticated approaches like reinforcement learning, as well as diverse datasets, could further boost the algorithm's prediction skills and present deeper insights into numerous consumer markets.

Funding

This work was supported by Research Center for International Transmission of Sichuan Liquor Culture（CJCB2024-03）

References

[1] Kim, S. Y., & Kim, H. J. (2021). Trends in alcohol consumption for Korean adults from 1998 to 2018: Korea National Health and Nutritional Examination Survey. *Nutrients,* 13(2), 609. https://doi.org/10.3390/nu13020609

[2] Bratt, D. D. M. (2023). A Historical Archaeology of Whisky in the Highlands and Islands of Scotland, C. 1500-1850 (Doctoral dissertation, University of the Highlands and Islands). https://doi.org/10.11141/ia.61.3

[3] Shkolyar, N. A. (2020). Tequila: features of production and consumption. *Latinskaia Amerika*, (3), 33-44. https://doi.org/10.31857/s0044748x0008390-8

[4] Park, E. J., Shin, H. J., Kim, S. S., Kim, K. E., Kim, S. H., Kim, Y. R., ... & Han, K. D. (2022). The effect of alcohol drinking on metabolic syndrome and obesity in Koreans: big data analysis. *International journal of environmental research and public health,* 19(9), 4949. https://doi.org/10.3390/ijerph19094949

[5] Ramos-Vera, C., Serpa Barrientos, A., Calizaya-Milla, Y. E., Carvajal Guillen, C., & Saintila, J. (2022). Consumption of alcoholic beverages associated with physical health status in adults: secondary analysis of the health information national trends survey data. *Journal of primary care & community health*, 13, 21501319211066205. https://doi.org/10.1177/21501319211066205

[6] Barbosa, C., Dowd, W. N., Barnosky, A., & Karriker-Jaffe, K. J. (2023). Alcohol consumption during the first year of the COVID-19 pandemic in the United States: results from a nationally representative longitudinal survey. *Journal of Addiction Medicine*, 17(1), e11-e17. https://doi.org/10.1097/adm.0000000000001018

[7] Ford, J., Nava, C., Tan, J., & Sadler, B. (2020). Automated Machine Learning Framework for Demand Forecasting in Wholesale Beverage Alcohol Distribution. *SMU Data Science Review*, 3(3), 7. https://scholar.smu.edu/datasciencereview/vol3/iss3/7

[8] Cravero, M. C., Laureati, M., Spinelli, S., Bonello, F., Monteleone, E., Proserpio, C., ... & Dinnella, C. (2020). Profiling individual differences in alcoholic beverage preference and consumption: New insights from a large-scale study. *Foods,* 9(8), 1131. https://doi.org/10.3390/foods9081131

[9] Buakate, P., Thirarattanasunthon, P., & Wongrith, P. (2022). Factors influencing alcohol consumption among university students in Southern Thailand. *Roczniki Państwowego Zakładu Higieny*, 73(4). DOI: 10.32394/rpzh.2022.0239

[10] Zhao, J., Stockwell, T., Vallance, K., & Hobin, E. (2020). The effects of alcohol warning labels on population alcohol consumption: an interrupted time series analysis of alcohol sales in Yukon, Canada. *Journal of studies on alcohol and drugs,* 81(2), 225-237. https://doi.org/10.15288/jsad.2020.81.225

[11] Jagadeesan, S., & Patel, P. (2021). Epidemiology, pattern, and prevalence of alcohol consumption in India: need for public health action. *International Journal of Community Medicine and Public Health, 8*(4), 2070-76. https://doi.org/10.18203/2394-6040.ijcmph20211282

[12] Parekh, N., Lin, Y., Chan, M., Juul, F., & Makarem, N. (2021). Longitudinal dimensions of alcohol consumption and dietary intake in the Framingham Heart Study Offspring Cohort (1971–2008). *British journal of nutrition,* 125(6), 685-694. https://doi.org/10.1017/s0007114520002676

[13] Auchincloss, A. H., Niamatullah, S., Adams, M., Melly, S. J., Li, J., & Lazo, M. (2022). Alcohol outlets and alcohol consumption in changing environments: prevalence and changes over time. *Substance abuse treatment, prevention, and policy*, 17(1), 7. https://doi.org/10.1186/s13011-021-00430-6

[14] Rastogi, A., Manthey, J., Wiemker, V., & Probst, C. (2022). Alcohol consumption in India: a systematic review and modeling study for sub-national estimates

of drinking patterns. *Addiction*, 117(7), 1871-1886. https://doi.org/10.1111/add.15777

[15] Dsouza, E. P., Dayanand, M. S., & Borde, N. (2021). The impact of tourist's socio-demographics on the choice of alcohol and choice of drinkscapes. *Revista de turism-studii si cercetari in turism,* (31). https://doi.org/10.9707/2328-0824.1221

[16] Niemelä, O., Aalto, M., Bloigu, A., Bloigu, R., Halkola, A. S., & Laatikainen, T. (2022). Alcohol drinking patterns and laboratory indices of health: does the type of alcohol preferred to make a difference?. *Nutrients,* 14(21), 4529. https://doi.org/10.3390/nu14214529