US Drought Predictions using Meteorological Data

1st Akshaya Srinivasan
Department of Applied Data Science
San Jose State University
San Jose,United States
Akshaya.Srinivasan@sjsu.edu

2nd Harshitha Ramesh
Department of Applied Data Science
San Jose State University
San Jose,United States
harshitha.ramesh@sjsu.edu

3rd Praphul Kenkere Omkarmurthy
Department of Applied Data Science
San Jose State University
San Jose,United States
Praphul.kenkereomkarmurthy@sjsu.edu

4th Manisha Paliwal
Department of Applied Data Science
San Jose State University
San Jose,United States
manisha.paliwal@sjsu.edu

5th Sonal Sharma
Department of Applied Data Science
San Jose State University
San Jose,United States
Sonal.Sharma@sjsu.edu

***Abstract*—Agriculture is one of the pillars of the country's economy. Over years climate change has been impacted agriculture negatively, drought has been one of them. The effect of drought can be very drastic to the economy of a country. Global warming and climate change are few factors that cause drought. Machine learning as a technology has been marking marks in various sectors including agriculture.Implementing most appropriate prediction model can help the farmers take the necessary precautions and actions to avoid large-scale crop damage and financial loss.In this work, we performed a comparative study of performance of machine learning models- KNN, Naïve Bayes, Decision Tree, and Random Forest. Performance of each model was evaluated using performance metrics like accuracy, Precision, Recall, F1 Score, Cohen Kappa score.This project concludes that Random Forest gives the highest performance at 80.8% closely followed by KNN at 79.8%.**

***Index Terms*—Drought, Prediction, Decision Tree, Naïve Bayes, KNN, Random Forest**

I. Introduction

Knowing the relation between soil moisture and drought indices, drought prediction and forecasting might be targeted at a certain index over a specific time period. Hydrological models and data-driven methodologies have been the most widely utilized tools for assessing and forecasting drought up to now. The former might aid in the comprehension of physical developments. Although data-driven approaches may not be appropriate for long-term forecasting of nonlinear dynamical systems, physical-based models' limits on spatial-temporal data make data-driven models complementary to physical-based models. Data-driven strategies may aim to extract relationships that might help to better inform and supplement present physical knowledge [2]. In the domain of hydrology, the Support Vector Machine (SVM) is one of the data-driven algorithms that has been effectively employed in classification, regression, and forecasting. The need of merging physics knowledge with data mining has been underlined, not just to prevent creating false insights, but also to generate new results[3].

Decision Tree and SVM are two of the most extensively utilized machine learning algorithms for constructing drought prediction models. SVM inputs are mapped onto a higher-dimensional space, where the nonlinear relationship between predictors and predictands is transformed into a linear relationship. [4]. SVM can learn from a smaller amount of data and can handle a vast number of variables. The selection and identification of predictors in the creation of drought prediction models are critical steps in ensuring that droughts are accurately predicted. Although non-linear models are so flexible, they may be prone to over fitting. Nonetheless, ensemble models like random forests and gradient boosted trees, which are among the most successful machine learning techniques in use today, have been developed to avoid decision trees from over fitting.

The primary goal of this study is to implement the well know non-linear classifiers such as Decision Tree, Support Vector machine and Naïve Bayes and compare the performance of the models using various evaluation metrics such as accuarcy,precision,recall and F1 score.

II. Significance to the Real World

Drought related concerns are a worldwide problem which needs immediate attention. NASA's GISS examined past precipitation and tree ring statistics between 1900 and 2005, discovering that carbon emissions - has had a major influence on worldwide droughts. According to the research, this human effect is expected to increase, potentially leading to "catastrophic" effects for society, such as increasingly severe droughts, food and water scarcity, deadly wildfires, and tensions combating for resources. (Written by Douglas Broom, 5 droughts that changed human history n.d.). Drought prediction can help mitigate extreme conditions by predicting possibilities so that measures can be taken proactively to mitigate extremities.

III. Literature Survey

The authors perform diagnoses of droughts are based on the Soil Moisture Index, comparisons are done using various

models with and without sequential lead bias based on ERA5-Land atmospheric input data from MODIS satellite data. The model makes use of a time series of variables drawn from the ERA5-Land reanalysis. To anticipate drought labels, the model takes climatic input data for the present site, such as geographical and periodic features. For classification, SVMs with linear kernels and an MLP are employed. The CNN (M3) and LSTM (M4) models are used as sequence encoders. The drought classification is generated via a completely connected layer on top of this representation for both sequence encoders. There is no distinctive classifying model except linear SVM performed comparable across iterations. In addition, the authors provide ablation studies to convert to coarser input data resolutions and show that model functions can be converted to lower resolutions when trained at higher resolutions. This study looks at a performance by changing the resolution of the data from 0. 1 degree to a coarser spatial resolution. This promising result shows that it is possible to predict drought events under various future climatic situations using a model trained with a fine-grained drought label. Overall, deliberately limiting the variables available in the climate model paves the way for an application to simulated data, thereby facilitating the study of agricultural drought in changing climates[8]

In another research,the authors work improves the predictor selection and builds a novel model to forecast droughts in Shaanxi province, China, utilizing previous drought indexes, meteorological indicators, and climatic signals from 32 stations from 1961 to 2016. To choose the best predictors and determine their lag duration, the authors used and contrasted 2 techniques: a cross-correlation function and a distributed lag nonlinear model (DLNM). The validations of a DLNM, an artificial neural network model, and an XGBoost model for forecasting the Standardized Precipitation Evapotranspiration Index (SPEI) were compared. In predictor selection and lag effect determination, DLNM was shown to outperform the cross-correlation function. The XGBoost model predicted SPEI more effectively than the DLNM and the artificial neural network. Furthermore, the XGBoost model exhibited the greatest forecast accuracy for overall droughts (89 percent – 97 percent) as well as three drought classifications (mild, serious, and catastrophic) (76 percent – 94 percent). This study proposed a new estimation method for forecasting SPEI and droughts. Incorporating the nonlinear and hysteretic effects of variables into the XGBoost algorithm can improve SPEI and drought accuracy[9]

One of the research work done for drought in Pakistan, it was seen that KNN-based drought models display limited performance in comparison to that of SVM and ANN- based drought models in validation.SVM Based models performed better for temporal and spatial characteristics of droughts due to the superior generalization skills of the SVM algorithm which guarantees a global optimum solution unlike ANN which can get trapped at a local optimum.However, the models based on all ML techniques displayed limited ability to capture extreme droughts seen during the Rabi season.[10]

Classification of droughts will help in strategizing better response for droughts. Authors have developed drought forecast model and a comparison was made between climatological data and long-range climate forecast data. User and producer's drought accuracy were used as two performance measures. Randomized trees predicted producer's accuracy of 64

IV. Project Development Methodology

The project followed agile methodology by implementing iterative and incremental development. Daily scrum meetings were conducted for 15 minutes over zoom to discuss progress of the task assigned. Epic was created for the project. Each phase of project development was divided into sprints and further task assigned to team members. The development of project code was done using pair programming by synching through google documents, git, and other methods.

V. Data Engineering

A. Dataset

The dataset for this analysis and prediction was sourced from the US Drought monitor. The data consists of 18 meteorological indicators that could be predictors of drought. The extracted data consists of 10 Million records observed at a specific location in the US over the years. The purpose of this dataset is to aid in the analysis of the possibilities of drought prediction only using meteorological data, leading to the generalization of US forecasts to other parts of the world. The dataset is labeled with drought scores ranging from over category D0 – Abnormally Dry to D4-Exceptional Drought

The dataset contains below features:

- WS10M-MIN: Minimum Wind Speed at 10 Meters (m/s)
- QV2M: Humidity at 2 Meters (g/kg)
- T2M-RANGE: Temperature Range at 2 Meters (C)
- WS10M: Wind Speed at 10 Meters (m/s)
- T2M: Temperature at 2 Meters (C)
- WS50M-MIN: Minimum Wind Speed at 50 Meters (m/s)
- T2M-MAX: Maximum Temperature at 2 Meters (C)
- WS50M: Wind Speed at 50 Meters (m/s)
- TS: Earth Skin Temperature (C)
- WS50M-RANGE: Wind Speed Range at 50 Meters (m/s)
- WS50-MAX: Maximum Wind Speed at 50 Meters (m/s)
- WS10-MAX: Maximum Wind Speed at 10 Meters (m/s)
- WS10-RANGE: Wind Speed Range at 10 Meters (m/s)
- PS: Surface Pressure (kPa)
- T2MDEW: Dew/Frost Point at 2 Meters (C)
- T2M-MIN: Minimum Temperature at 2 Meters (C)
- T2MWET: Wet Bulb Temperature at 2 Meters (C)
- PRECTOT: Precipitation (mm day-1)

B. Preprocessing and cleaning dataset

The dataset contained records without labels as drought scores are updated weekly while the meteorological statistics are updated daily. Therefore, records without labels were removed as the scope of this project is to perform supervised learning models. The count of data was reduced to $2,756,796$ records. There were no null values in any other column. Data columns were reformatted to the required datatype. Date

column was further split into Day, Month and Year columns for analysis. The outlier for each feature was detected. Values that do not fall under three sigma curves of normalization were removed.$Values \leq Mean(x) + 3(Std(x)) - Values \geq Mean(x) - 3(Std(x)))$

C. Exploratory Data Analytics

Data analysis and visualizations were performed to identify data patterns and anomalies.

1) Identifying Data Imbalance: US Drought dataset is a labelled dataset. Distribution of scores (referring to labels) is analyzed to identify if data is biased or not. As per Fig 1, dataset is imbalanced containing high volume of data with score 0 compared to other scores.

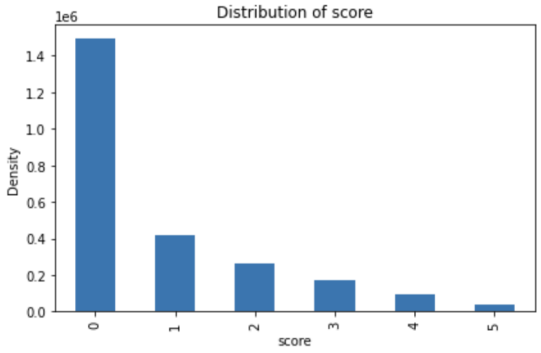

Fig. 1. Data Imbalance

2) Univariate Data Analysis of features : Univariate data analysis helps in identifying skewness of data for each feature. Fig 2. and Fig. 3 represents distribution of each feature. We can clearly see that features PRICTOT ,WS10M-MIN,WS50M-MIN,WS10M-RANGE are skewed to the left and PS,T2M,T2M-MAX are skewed to the right while remaining features are fairly well distributed across all ranges.

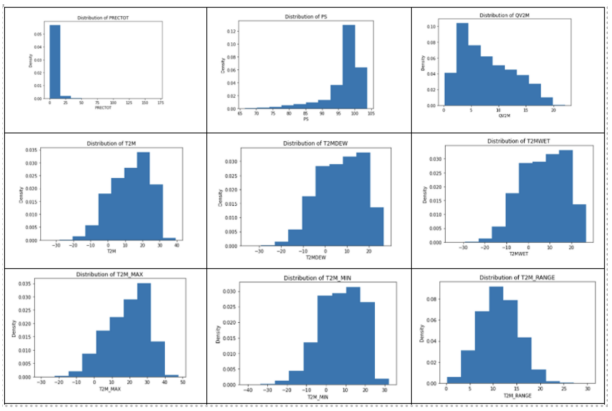

Fig. 2. Histogram for univariate analysis

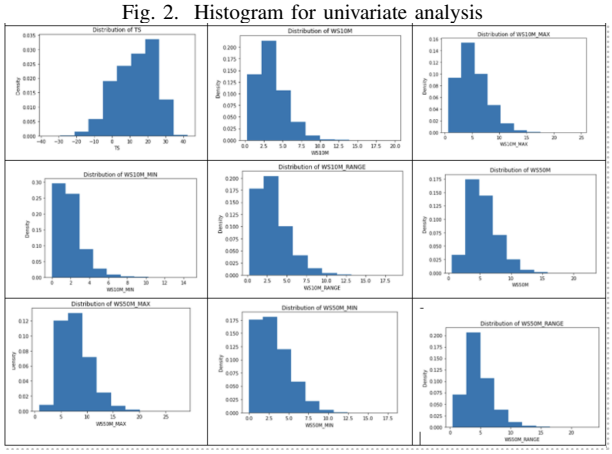

Fig. 3. Histogram for univariate analysis

3) Outlier Analysis : Box plot is the suitable representation of spread of values. Fig.4 shows clearly that features PRECTOT, PS,WS10M, WS10M-RANGE,WS50M-RANGE,WS50M-MIN. There are no outliers in QV2M.

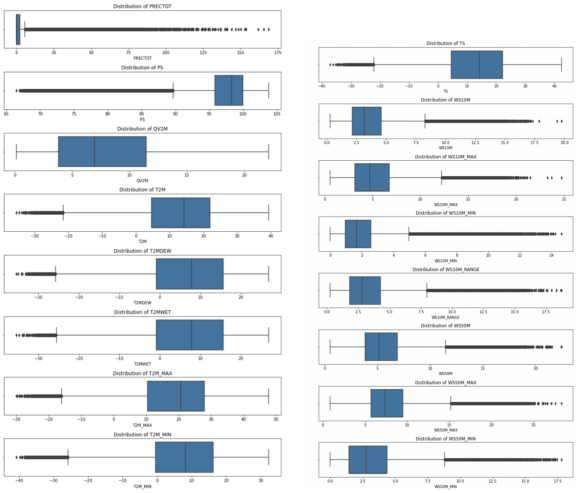

Fig. 4. Box Plot for Outlier analysis

4) Correlation Matrix : Correlation between features is represented in the form of heatmap. Fig. 5 shows attributes QV2M, T2M, T2MDEW, T2MWET, T2M-MAX, T2M-MIN and TS have shown strong positive correlation.Similarly, WS10M, WS10M-MAX and WS10M-MIN have shown a strong positive correlation. Also, WS50M, WS50M-MAX and WS50M-MIN show strong positive correlation.

	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE
PRECTOT	1.000000	0.068775	0.245081	0.093258	0.231035	0.230975	0.026773	0.144929	-0.304171
PS	0.068775	1.000000	0.282412	0.164160	0.341234	0.341252	0.111979	0.208285	-0.225935
QV2M	0.245081	0.282412	1.000000	0.870242	0.959385	0.960434	0.804338	0.906144	-0.071547
T2M	0.093258	0.164160	0.870242	1.000000	0.913530	0.914218	0.983356	0.981629	0.244357
T2MDEW	0.231035	0.341234	0.959385	0.913530	1.000000	0.999970	0.854716	0.939934	-0.015643
T2MWET	0.230975	0.341252	0.960434	0.914218	0.999970	1.000000	0.855401	0.940629	-0.015500
T2M_MAX	0.026773	0.111979	0.804338	0.983356	0.854716	0.855401	1.000000	0.937762	0.407534
T2M_MIN	0.144929	0.208285	0.906144	0.981629	0.939934	0.940629	0.937762	1.000000	0.065037
T2M_RANGE	-0.304171	-0.225935	-0.071547	0.244357	-0.015643	-0.015500	0.407534	0.065037	1.000000
TS	0.089598	0.163830	0.862559	0.997515	0.905184	0.905911	0.980101	0.979134	0.241564
WS10M	0.049730	-0.080747	-0.225449	-0.207874	-0.238299	-0.237971	-0.216764	-0.206382	-0.080163
WS10M_MAX	0.060981	-0.135905	-0.256452	-0.220192	-0.268686	-0.268292	-0.221671	-0.225829	-0.043127
WS10M_MIN	0.023346	0.022932	-0.108789	-0.125407	-0.115920	-0.115882	-0.141911	-0.112878	-0.110952
WS10M_RANGE	0.065755	-0.198332	-0.269203	-0.209030	-0.280702	-0.280199	-0.199614	-0.225256	0.018746
WS50M	0.069057	-0.043315	-0.205971	-0.193196	-0.204238	-0.204143	-0.195727	-0.197991	-0.041778
WS50M_MAX	0.079508	-0.091821	-0.249961	-0.206444	-0.245323	-0.245147	-0.196236	-0.225744	0.029737
WS50M_MIN	0.057816	0.036238	-0.081554	-0.112579	-0.082416	-0.082497	-0.133234	-0.096593	-0.128844
WS50M_RANGE	0.047477	-0.154479	-0.246203	-0.159589	-0.239335	-0.239029	-0.126331	-0.200157	0.163320

Fig. 5. Correlation Matrix

5) Bivariate Analysis : Scatter plots were drawn to understand correlations between attributes having strong correlation. As shown in Fig. 6- Fig. 8, between the independent variables that have shown strong positive correlation, the relationship is one-to-one for pairs WS10M - WS50M. However, for pairs T2M – T2MDEW and QV2M - T2M the relationship is not so linear, but the overall correlation is strong.

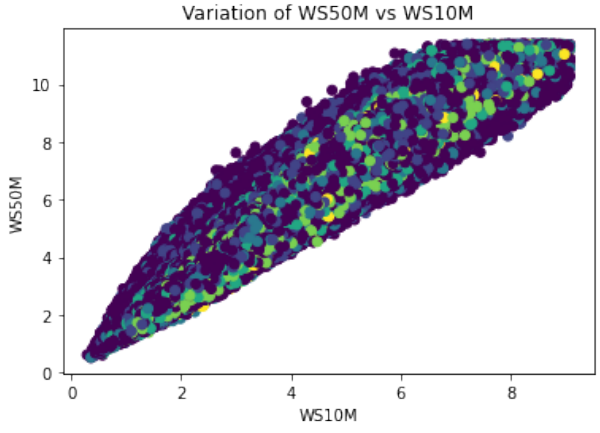

Fig. 6. Distribution of WS50M vs WS10M

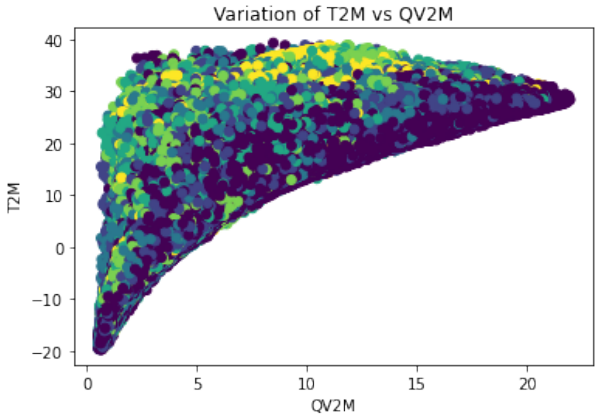

Fig. 7. Distribution of T2M vs QV2M

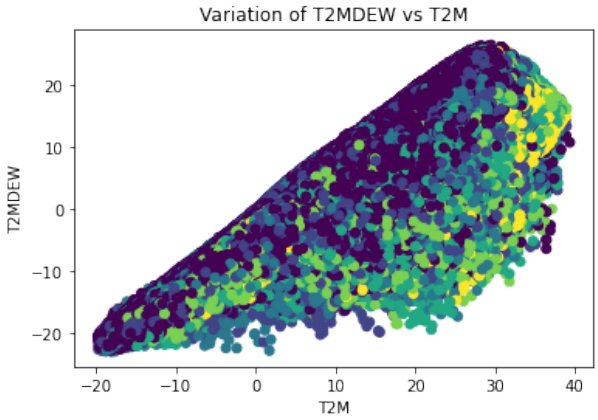

Fig. 8. Distribution of T2MDEW vs T2M

D. Feature selection

The Recursive Feature Elimination technique using Random Forests has been used for feature selection. The reason for this is that random forests' tree-based methods are naturally ranked by how effectively they increase node purity. This is the average reduction in impurity across all trees (called gini impurity). The number of features to be chosen is set to 8. The resultant features are fed back into the model and fed into the standard scalar function. Initially there are 23 features present in the raw dataset which are fed into the RF model and using the RFE function the best 15 features are selected for further modelling.

E. Dimensionality Reduction

Dimensionality reduction is a technique for reducing model complexity and avoiding overfitting. The imbalance class is fixed by performing two different operations: Upsampling using SMOTE (SMOTE()) and Down sampling using Near Miss(NearMiss()) The data obtained from under sampling of data using NearMiss function is used to sit into the PCA to perform dimensionality reduction. The principal component analysis is a technique widely used for dimensionality reduc-

tion it is an approach for compressing a dataset into a lower-dimensional feature subspace while retaining the majority of the relevant data. The down sampled training data is used to find the eigenvector with the help of the covariance. The resultant eigenvector array is fit as the new training dataset which is, in turn, fit into a data frame find the reduced principal features of the drought data.

F. Train and split data

The drought data is split into (80-20) training and testing datasets. The training dataset is then fit into the Standard scalar function to remove the main and scale the features unit variance

VI. Model Development

A. Modelling with Decision Tree

DTs are one of the most often utilized approaches for classification and prediction. These algorithms give a method that proceeds from top to bottom or from broad to specific throughout the training phase. In this approach, which is a flowchart-like tree structure, the attribute value of each node is examined, and branches are formed utilizing the findings. The root node of a DT is followed by separation criteria such as information gain. Check for clustering at the root node of the tree, which contains all training cases. The solution is found when all cases in the root node correspond to a single cluster. Otherwise, the root node is split into branches and repeated until the branch is simple enough to decide on its own. Depending on the amount of the dataset, finding tree branches might be tricky. Pruning methods are used to prevent overfitting by deleting leaf nodes holding a limited number of objects from the decision tree. Figure 9 represents the decision tree structure.

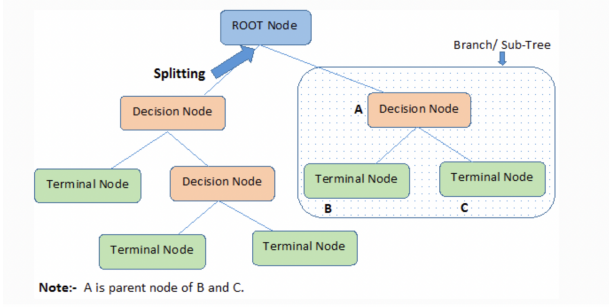

Fig. 9. Source:https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

The purpose of machine learning is to reduce uncertainty and disorder in a dataset, and we employ decision trees to do it. The uncertainty in our dataset or measure of disorder is called entropy. Given below is the formula for entropy:

Where, p+ is probability of positive classes , p– probability of negative class, S - is subset of the training example

The upsampled training data with SMOTE is fed into the decision tree classifier and the performance of the model is

$$E(S) = -p_{(+)}\log p_{(+)} - p_{(-)}\log p_{(-)}$$

Fig. 10. Source:https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

evaluated using accuracy,, precision, recall, and F1score. The models shows in accuracy of 76.30%. To improve the accuracy of the model, hyper-parameter tuning was performed by setting the maximum depth parameter to 70 and data fed back into the model and the model's accuracy came to be 76.33%.

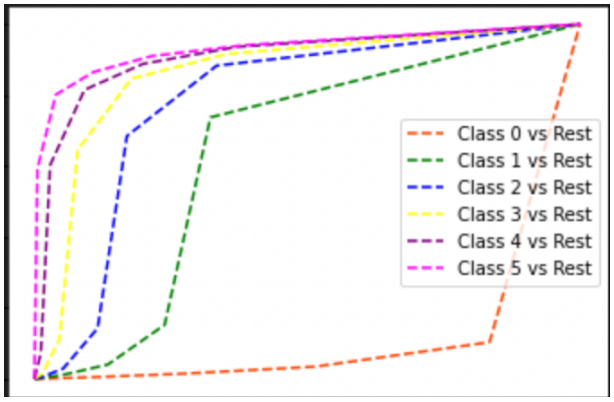

Fig. 11. Mutliclass ROC curve for Decision Tree without resampling

B. Modelling with Naïve Bayes

Naive Bayes classifiers are a type of statistical Classification Algorithm centered on the Bayes theorem that assumes feature independence. Below equation expresses the Bayes theorem: $P(X \mid Y) = P(X).P(Y \mid X)/P(Y)$ This equation calculates the probability of event X occurring if Y is true. Since it is presumed that features are independent, NB classifier does not correlate features while making predictions. The benefits of this classifier are in its easy design and its ability to scale effectively for huge databases.

In the project, we are using Gaussian Naïve Bayes, with an assumption that the continuous values are normally distributed across the associated classes. Fig. 10 depicts the operation

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Fig. 12. Source:https://iq.opengenus.org/gaussian-naive-bayes/

of a GNB classifier. At each data point, the z-score distance between that point and the class mean is determined, which is the distance from the mean divided by standard deviation of a class. The results of Naïve bayes classifier algorithm applied without resampling the data resulted in an accuracy of 0.58

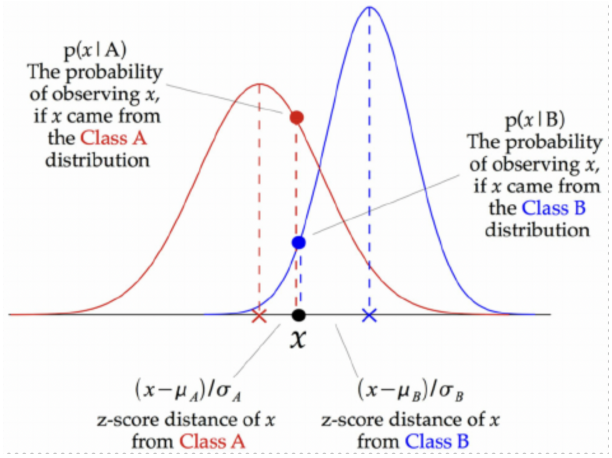

Fig. 13. Operation of Gaussian Naïve Bayes

which is low compared to other models used in the study. Fig. 11 shows the ROC curve.

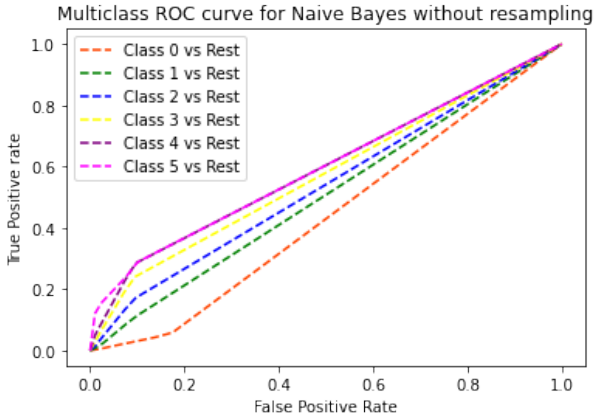

Fig. 14. Multiclass ROC curve for Naive Bayes without resampling

C. Modelling with KNN

Most of the real-world scenarios require decision making that means in most cases we must make decisions and hence we encounter classification problems frequently. KNN is widely used for classification problems, and it is extremely easy to interpret. A classification problem has a distinct set of fixed outcomes. KNN works on the assumption that similar data points lie close to each other. KNN works on the concept of similarity (closeness, distance). KNN uses various distance metrics like Euclidean distance, Manhattan distance etc., we are using Minkowski distance, it can be measured in a space where distances are represented by a vector having a particular length. For computing Minkowski consider two points, P1: (X1,X2,...XN) P2: (Y1,Y2,...YN)

Minkowski distance between these two points will be calculated using below formula,

Refer below image to visualize the above formula, For

$$\sqrt[p]{(x1-y1)^p + (x2-y2)^p + \ldots + (xN-yN)^p}$$

Fig. 15. Source:https://iq.opengenus.org/minkowski-distance/

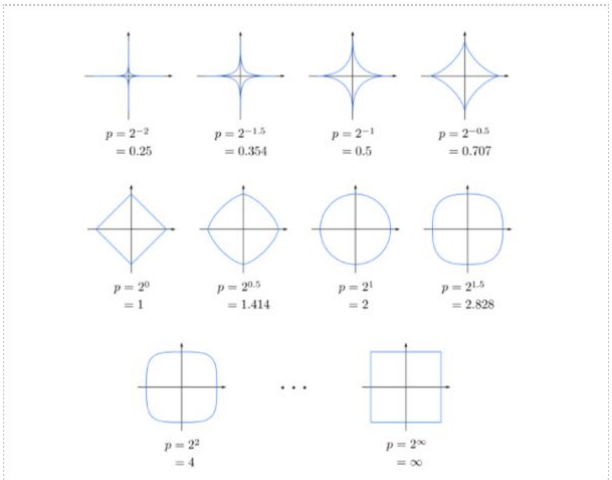

Fig. 16. Source:https://iq.opengenus.org/minkowski-distance/

implementing KNN, we will first select the number k, where k is nearest neighbors then calculate the distance for those k number of neighbors. Take those k neighbors according to calculated distance and then count data points in every category. We will then assign these new data points to that category. For hyperparameter tuning for KNN we considered K value from 1 to 10, cross-validation batch size as 3 and the scoring metrics was set to accuracy. Accuracy of KNN Algorithm without resampling after hyperparameter tuning came out to be 0.798

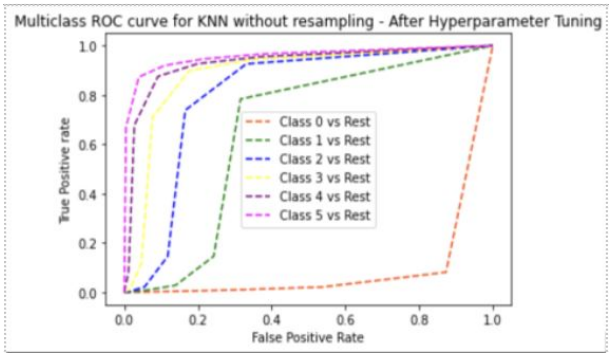

Fig. 17. Multiclass ROC Curve for KNN without resampling

D. Modelling with Random Forest

Random forest is one of the most popular bagging techniques that is used. The name random forest comes from 2 parts. One, we will be doing a random bootstrapped sampling with replacement and hence the name random. Two, we will be using many decision trees and hence the name forest. It

is taking decision trees as your base learners and applying bagging on top of it. We want to have our base learners to have low bias and high variance. Because of aggregation, it reduces the variance. We train decision trees to full depth and keep growing the depth as long as a reasonable model is achieved and there are enough points at each stage. The out of bag samples can be used for cross-validation. The aggregation here can be both classification (majority vote) and regression(mean/median). The base learners are decision trees with good depth.

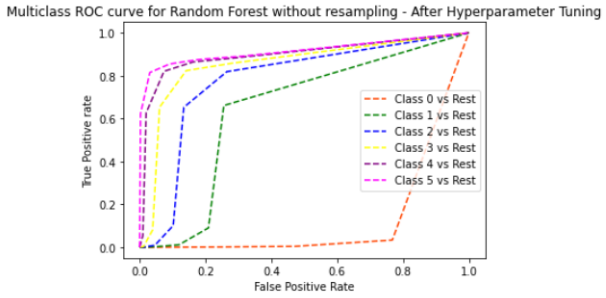

Fig. 18. Multiclass ROC curve for random forest without re-sampling- After hyperparameter tuning

E. Evaluation

Accuracy: It is used to assess how well machine learning models perform. The accuracy ratio is the number of correctly classified points divided by the total number of points. We can attain excellent accuracy with a dumb model if we have an imbalanced dataset. As a result, accuracy cannot be applied to datasets that are unbalanced. Accuracy does not provide a good indication of whether a model is weak or powerful.

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

Fig. 19. Accuracy formula

Precision: This performance metric is often used in information retrieval problems. It is the number of true positives divided by the total number of true positives and false positives. It means, of all the points, the model is predicted to be positive. What percentage of them are actually positive? In information retrieval, we only care about the positive class.

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

Fig. 20. Precision formula

Recall: This performance metric measure is also used in information retrieval problems. Of all the points which actually belong to class 1, how many did the model detect it to be class

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Fig. 21. Recall formula

1? It is nothing but the true positive rate. It is the number of true positives divided by the total number of positives.

VII. Results

Out of all the models Random Forest without resampling has the highest accuracy 80.896% followed by KNN without resampling which has accuracy of 79.87% . Fig. 12 shows the result comparison of all models evaluated using performance metrics - Accuracy, Precision,Recall,F1 Score,Cohen Kappa score

Algorithm	Accuracy	Precision	Recall	F1 Score	Cohen Kappa Score
Random Forest without resampling	0.808959	0.796925	0.808959	0.798690	0.654981
KNN without resampling	0.798651	0.798294	0.798651	0.798471	0.657498
KNN with SMOTE Upsampling	0.795267	0.801758	0.795267	0.798198	0.657827
Decision Tree with SMOTE Upsampling	0.764228	0.772588	0.764228	0.767987	0.607222
Decision Tree without resampling	0.763337	0.762305	0.763337	0.762809	0.596681
Decision Tree with SMOTE Upsampling and PCA	0.691158	0.721815	0.691158	0.703299	0.504504
Decsion Tree with SMOTE Upsampling and LDA	0.602809	0.674628	0.602809	0.628327	0.394723
Naive Bayes without resampling	0.585144	0.449910	0.585144	0.480441	0.080746
SVM with Near Miss Downsampling	0.299534	0.512324	0.299534	0.362867	0.078111
KNN with Near Miss Upsampling	0.232508	0.566489	0.232508	0.268879	0.093552
Decision Tree with Near Miss Downsampling	0.224805	0.543185	0.224805	0.262600	0.078760
Decision Tree with Near Miss Downsampling and LDA	0.204748	0.514249	0.204748	0.249713	0.057772
Decision Tree with Near Miss Downsampling and PCA	0.189016	0.520894	0.189016	0.224072	0.059722

Fig. 22. Results Comparison

VIII. Conclusions and Future Scope

The proposed system predicts drought through the climatic data for USA.The intent is to understand scope of climatic changes alone in causing drought. This might lead to the extension of US projections to other places around the world.There is a need to develop such system because drought have become major cause of concern world wide.Our system predicts drought severity across five levels.The comparative study will assist in implementing appropriate models in future works.As a continuation to our current work, we would like to experiment further with non-linear dimensionality reduction methods such as TSNE, Isomap, etc. for feature extraction. We can also try creating a separate binary classifier for predicting whether there is a drought, and if yes, build a multi-class classifier to further predict the level of drought in order to achieve better accuracies for each class.

References

[1] Bourdin, D. R., Fleming, S. W., & Stull, R. B. (2012). Streamflow modelling: A primer on applications, approaches and challenges. Atmosphere-Ocean, 50(4), 507-536. https://doi.org/10.1080/07055900.2012.734276

[2] Labudová, L., Labuda, M., & Takáč, J. (2016). Comparison of SPI and SPEI applicability for drought impact assessment on crop production in the Danubian lowland and the east slovakian lowland. Theoretical and Applied Climatology, 128(1-2), 491-506. https://doi.org/10.1007/s00704-016-1870-2

[3] Ganguli, P., Reddy, M. J. (2013). Ensemble prediction of regional droughts using climate inputs and the SVM-copula approach. Hydrological Processes, 28(19), 4989-5009. https://doi.org/10.1002/hyp.9966
[4] Ganguly, A. R., Kodra, E. A., Agrawal, A., Banerjee, A., Boriah, S., Chatterjee, S., Chatterjee, S., Choudhary, A., Das, D., Faghmous, J., Ganguli, P., Ghosh, S., Hayhoe, K., Hays, C., Hendrix, W., Fu, Q., Kawale, J., Kumar, D., Kumar, V., … Wuebbles, D. (2014). Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. Nonlinear Processes in Geophysics, 21(4), 777-795. https://doi.org/10.5194/npg-21-777-2014
[5] Lantz, B. (2019). Machine learning with R: Expert techniques for predictive modeling (3rd ed.). Packt Publishing.
[6] Tabari, H., Kisi, O., Ezani, A., Hosseinzadeh Talaee, P. (2012). SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid Highland environment. Journal of Hydrology, 444-445, 78-89. https://doi.org/10.1016/j.jhydrol.2012.04.007
[7] Wang, H., Rogers, J. C., & Munroe, D. K. (2015). Commonly used drought indices as indicators of soil moisture in China. Journal of Hydrometeorology, 16(3), 1397-1408. https://doi.org/10.1175/jhm-
[8] Sundararajan, K., Garg, L., Srinivasan, K., Bashir, A. K. (2021, July). A Contemporary Review on Drought Modeling Using Machine Learning Approaches. Retrieved March 12, 2022, from https://www.researchgate.net/profile/Jayakumar-Kaliappan/publication/353447055_A_Contemporary_Review_on_Drought_Modeling_Using_Machine_Learning_Approaches/links/611f41c11ca20f6f86345f47/A-Contemporary-Review-on-Drought-Modeling-Using-Machine-Learning-Approaches.pdf
[9] hang, R., Chen, Z.Y., Xu, L.J., & Ou, C.Q. (2019, February 10). Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi Province, China. Science of The Total Environment. Retrieved March 12, 2022, from url-https://www.sciencedirect.com/science/article/pii/S0048969719302281casa_token=B6P2W6Dt45niAAAAA%3ADQdch2f1TR9S6Yn1npPNkSLTtozRVCCJF_QZduJ2EJt xHH6u2yGq8t88TceXxC_a9hw
[10] Jiang, W., Luo, J. (2021, July 6). An evaluation of machine learning and deep learning models for drought prediction using weather data. arXiv.org. Retrieved March 13, 2022, from https://arxiv.org/abs/2107.02517
[11] Rhee, J., Im, J. (2017, February 10). Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and Remote Sensing Data. Agricultural and Forest Meteorology. Retrieved March 13, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0168192317300448
[12] Khan, N., Pour, S.H., Shahid, S., Ismail, T., Ahmed, K., Chung, E.S., Wang, X., 2019a. Spatial distribution of secular trends in rainfall indices of Peninsular Malaysia in the presence of long-term persistence. Meteorol. Appl. 655–670

IX. Appendix

A. Report

Team members have done combined work to complete the report. The content of the report is our original work and references are cited appropriately wherever required. Equal contribution is provided by each team member

B. Language correctness Evaluation

As recommended, Grammarly tool was used to ensure correctness of language in all documents specifically IEEE report

C. Used creative presentation techniques

Microsoft PowerPoint is an amazing presentation software used in the corporate world. We have used Microsoft PowerPoint to create our presentation.

D. Project Management and methodology

Agile methodology is used, and Trello Agile Sprint board is used for implementation of project life cycle. Regular meetings as part of agile were organized over zoom.

E. Teamwork

Equal contributions have been provided by each member of the team and all deliverable accomplished with proper coordination and due respect was given to each member's ideas during project implementation

F. Version Control

Git is a free and accessible medium of version control. Project source code and related documents are uploaded in GitHub with repository being Public so that everyone can have access to our project implementations. Each team member responsibly committed their changes timely to ensure proper version control and coordination.
Link: https://github.com/pbpablo/Drought_Data_Analysis/

G. Saving the model for demo

Pickle python package is used for serializing python code. We have used it to save model simulations. .P file is part of code deliverable

H. Used latex

Overleaf web application is used to create IEEE report. Latex file .tex is part of deliverables of this project

I. Pair Programming

A well-coordinated teamwork model was incorporated to enable productive pair programming details of which are mentioned under section 'Project Development Methodology'.

J. Prospect of winning competition

The project work aims at predicting U.S. drought with use of meteorological data. This approach can be generalized to predict drought throughout the world with basic information unlike in real world scenarios where complex attributes are used for prediction. Alongside performance evaluation of implementing the same models with different sampling methodologies, feature selections are unique to our project. Further, each model is compared with others to perform In depth performance evaluation. Project is uploaded in Kaggle for reference by people as baseline for their implementations

K. Velocity

US Drought monitoring is a real time data which keeps adding incrementally on daily and weekly basis. The project implementation can be easily extended to incorporate real time predictions and handle data velocity by performing parallel computing and avoiding overfitting by using regularization methods

L. Innovation

- Using a combination of methods like EDA, Correlation, Recursive Feature Extraction and Dimensionality Reduction using PCA and LDA for selecting and extracting the best features
- Use of Upsampling and Downsampling methods to handle class imbalance, to ensure the models developed are not biased towards a particular class

- Using insights from EDA to determine the best choice of models and feature selection methods

M. Technical difficulty

- Data Volume - The volume of our input dataset is 2 million rows with 21 attributes, which made processing multiple Machine Learning algorithms too complex in terms of time and storage
- Class Imbalance - The data was heavily imbalanced with over 1.7M records for class 0 alone and only 0.3M records for all other class together - We used resampling techniques such as SMOTE up sampling, Neighborhood Cleaning and Near Miss down sampling methods to handle this
- Non-Linearity - The data was not linearly distributed, due to which traditional dimensionality reduction methods such as PCA and LDA were inefficient
- Kernel Computations - Due to the huge data volume, kernel methods such as Kernel PCA, Kernel SVM with any type of kernel required more computational resources than what is available in free versions of coding consoles
- Hyperparameter Tuning - Experimenting with a diverse range of hyper parameters for the Machine Learning Models had a lot of time and resource complexity, so we had to settle for experimenting with extremely limited options

N. Visualizations

We have performed extensive visualization for EDA including univariate analysis plotting histograms, bivariate analysis using scatter plots, box plots for outlier analysis, heap map for correlation analysis. Matplotlib and pyplot libraries are used for implementation.

O. Evaluation of performance

- Confusion matrix was used to get an overview of misclassifications on a class-by-class level Metrics such as Accuracy, Precision, Recall, F1 Score and Kappa Score were used for overall performance evaluation of models
- Along with the statistical metrics, attributes such as time complexity, resource utilization such as GPU and Application memory were taken into account to pick the best model for this dataset
- Additionally, the RUC curve was also used to evaluate model performances on a class-by-class basis

P. Lessons learned

- Importance of analyzing the data distribution and identifying whether the data is linear or non-linear majorly helps in picking the right Dimensionality Reduction methods and Machine Learning Algorithms
- Performance of a Machine Learning model is based on a variety of parameters such as Overall Accuracy, Precision, Kappa Score, etc., and class-wise misclassification rate, and time and resource complexity
- Ensemble methods like Random Forest fare much better when compared to discriminative approaches like Naïve Bayes
- Implementing agile methodology for project work was new to all team members and helped us improve our management skills
- Referring signification research paper and other existing research in our field of study for the project broadened our understanding to come up with machine learning models implementation and evaluation strategies.
- Various methods of feature selecting were explored and random forest was finalised.Importance of using right feature selection technique and it's implications on modelling was an important learning