Wine Weather: Exploring Meteorology and Grape Harvests in California

Purdue University — The Data Mine

Uttam Reddy Lingareddygari, Olukayode Ifeloluwa Rebekah, Aidan Dibble, Suhani Gupta.

BASF — We create chemistry

INTRODUCTION

The project aims to develop a 2024 weather forecast in California to assist farmers in optimizing grape production by adjusting irrigation schedules and selecting grape varieties.

Objectives:

- Create a model capable of predicting wine grape yields from weather data.
- Test for accuracy of our results.
- Ultimately provide this data to farmers to ensure crop security and success through the coming growing seasons.

CENTRAL QUESTIONS

- What weather variables influence wine grape growth and yields the most?
- How does each weather variable influence growth and yields?
  - Is the variable beneficial or detrimental?
- Where is the best place to find data on past weather and forecasts for two years ahead?

REFERENCES

- *NASA Power View Weather Data* ***https://power.larc.nasa.gov/data-access-viewer/***
- Kaggle "California Wine Grape Yields 1980-2020" ***https://www.kaggle.com/datasets/jarredpriester/california-wine-production-19802020***
- "Climate analysis with satellite versus weather station data" by Robert Mendelsohn, Pradeep Kurukulasuriya, Alan Basist, Felix Kogan, Claude Williams *10.1007/s10584-006-9139-x*
- Stack overflow – for syntax and methodology

EXPLORATORY DATA ANALYSIS – NAPA COUNTY

Table 1: The dataset has 15 different variables. We needed to find out the percentage of this data which is missing. These are those percents.

Figure 1: We wanted to visualize the spread of each variable in our data set, to determine the skew and frequency of outliers for each variable. Some are relatively normal, such as temperature, while others are heavily skewed with lots of outliers like precipitation.

Figure 2: This is the process for transforming our data from numerical to categorical. Since our yield data is organized by year, we had to organize our weather data to reflect this. First, we took each variable and assigned a range of values which would be classified as "low", "medium", and "high" before we summed the number of each of these categories for each year.

Table 1

	Null, %
temp	0.019557
temp_dew	0.019557
temp_max	0.019557
temp_min	0.019557
sol_irr	8.924381
spec_hum	0.019557
rel_hum	0.019557
precip	0.019557
sur_pres	0.019557
wind_speed	0.019557
wind_speed_max	0.019557
wind_speed_min	0.019557
wind_dir	0.019557
sur_soil_wet	1.786180
root_soil_wet	1.786180

Figure 1

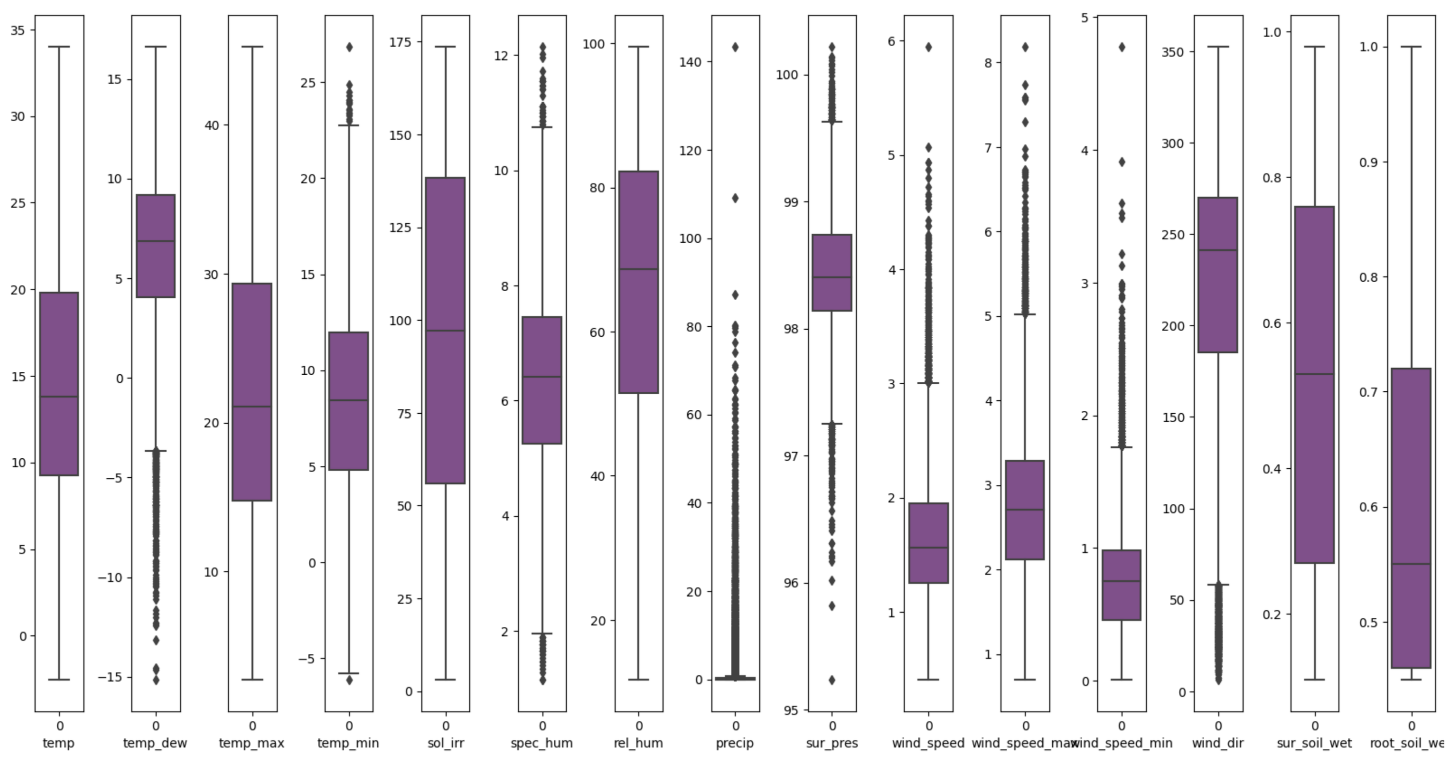


Figure 2

Date	temp	temp_dew	temp_max	temp_min
1981-01-01	7.83	1.66	16.84	3.16
1981-01-02	6.49	2.70	10.82	2.47
1981-01-03	9.48	8.44	12.90	6.66
1981-01-04	9.49	6.90	16.92	5.98
1981-01-05	8.30	3.20	17.37	2.98

→

date	temp	temp_dew	temp_max	temp_min
1981-01-01	low	low	medium	low
1981-01-02	low	low	low	low
1981-01-03	low	high	low	medium
1981-01-04	low	medium	medium	low
1981-01-05	low	low	medium	low

→

date	temp_high_count	temp_medium_count	temp_low_count
1981-12-31	133	111	121
1982-12-31	95	124	146
1983-12-31	110	110	145
1984-12-31	129	80	157
1985-12-31	110	114	141

FEATURE SELECTION

To begin making our ML model, we had to select a list of features – or variables – which we would analyze.

- These features were found by running a Chi-squared analysis.
- We selected these features primarily due to the results of the Chi-squared analysis.
- We did not just choose the top 8 results because we believe that some variables outside will have a higher impact on the yield.

For our model, we selected these features to analyze:

1. Rel-humidity medium count
2. Temp low count
3. Temp-max low count
4. Sur-pressure high count
5. Wind-speed medium count
6. Rain yes count
7. Temp-min high count
8. Sur-soil-wetness medium count

ML MODEL CREATION

We tested decision tree, logistic regression, support vector machines, random forest classifier, and k-nearest neighbors' methods to create our model. To apply our data to these models we used a label encoder. Our results from all these model kinds are displayed in Table 2.

RESULTS

Our model is trained on the first 2 years of weather data, and it attempts to predict the final 20 years of the data set we fed it.

- Accuracy is a measure of how close the model was to predicting the last 20 years.
- We believe the low accuracies from our initial models were from the choice of variables or from lurking variables.

We implemented clustering in Figure 3 to visualize the similarities between the years. We expect the years in one cluster to have similar weather and yields. We can use years within the cluster as analog years for forecasts with similar characteristics.

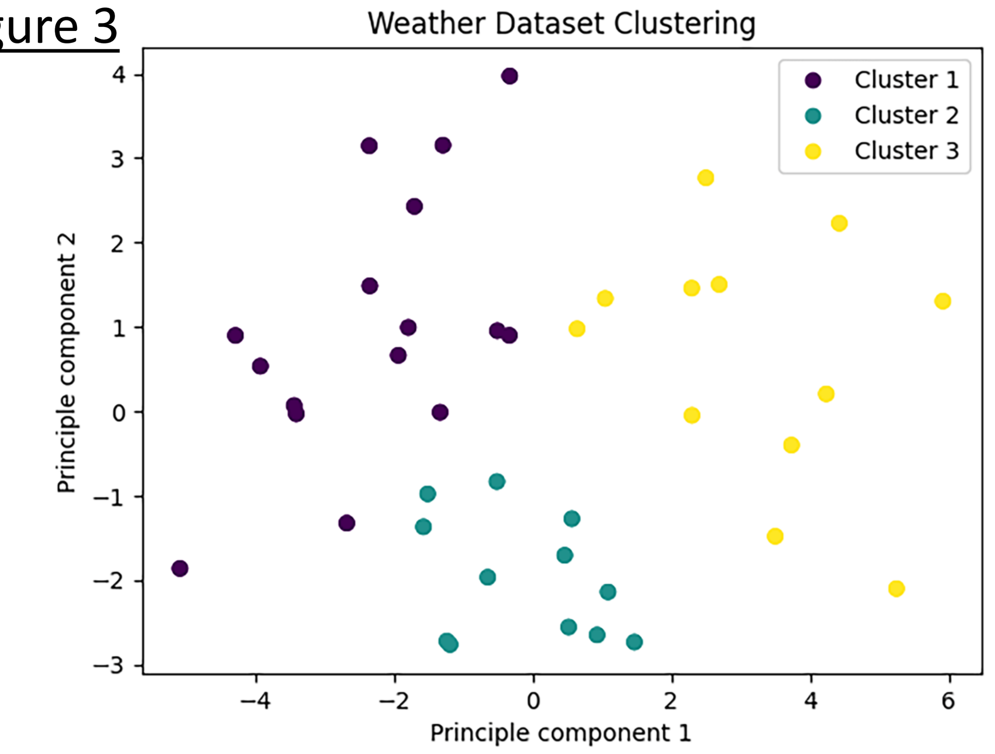
Table 2

ML model	Accuracy (%)
Logistic Regression	50%
Support Vector Machine	33.3%
Decision Tree Classifier	40%
Random Forest Classifier	43.3%
K Nearest Neighbors	66%

Table 3

```
Cluster 1: 1984, 1985, 1987, 1988, 1989, 1990, 1991, 1994, 1999, 2002, 2007, 2008, 2009, 2013, 2018, 2020
Cluster 2: 1981, 1992, 1996, 1997, 2001, 2004, 2012, 2014, 2015, 2016, 2017, 2019
Cluster 3: 1982, 1983, 1986, 1993, 1995, 1998, 2000, 2003, 2005, 2006, 2010, 2011
```

Figure 3



PLANS FOR FUTURE

Our model is not complete, and we have some plans to optimize it with the time remaining in the semester.

- We plan on making growing degree days a continuous numerical variable in all our models.
- We plan to test more combinations of weather variables, not just one set.
- Expand on the clusters and use this to help predict analog years.

We also have some suggestions for any work which may be done next semester, if this project continues.

- Investigate data on fertilizer or extreme weather.
- Reduce the size of the area a model covers, possibly down to all vineyards in a city or even down to one vineyard.

ACKNOWLEDGMENTS

- Divya Chauhan – Teaching Assistant
- Ryan Mills – BASF Marketing Lead
- Brady Spangenberg – BASF Marketing Director
- David Glass – TDM Managing Director
- Emily Hoeing – TDM Corporate Partners Advisor
- Miranda Purcell – Purdue Wine Grape Team Specialist