




# Rainfall Prediction

---

## **Machine Learning I**

Tanmay Vivek Kshirsagar  
Sudhanshu Deshpande  
Shreyas Sunku Padmanabha



# Objective

---

Predict next-day rain by training classification models on the target variable 'RainTomorrow' using the weather observations made on the current day.

# Background

---

Predicting whether it will rain tomorrow is important for multiple reasons such as safety, agriculture, business, and personal planning. It helps make daily decisions and reduce risk of failure due to unexpected rain.

Observations were drawn from numerous weather stations. The daily observations are available from Climate Data Online - Map search.

An example of latest weather observations in Canberra: Canberra, Australian Capital Territory March 2023 Daily Weather Observations

Definitions adapted from <http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>

Data source: <http://www.bom.gov.au/climate/dwo/> and <http://www.bom.gov.au/climate/data>.

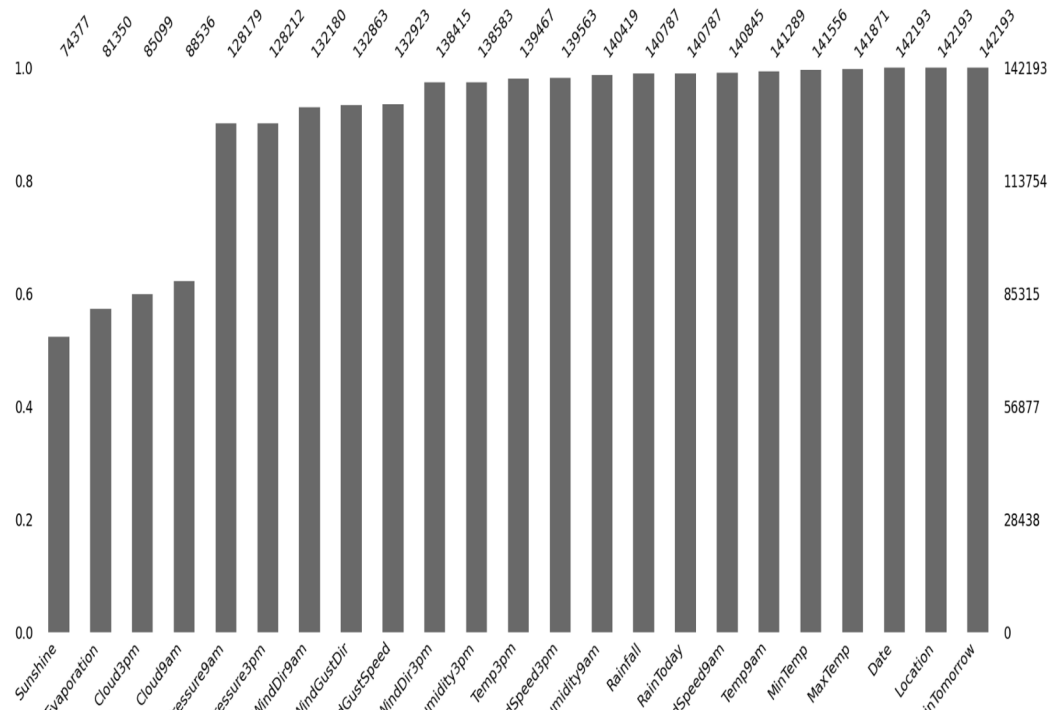
# Dataset Description

---

- This dataset contains about 10 years of daily weather observations from many locations across Australia.
- 145,460 observations and 23 columns.
- Source: [Kaggle](#).

```
'Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation',  
'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm',  
'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',  
'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am',  
'Temp3pm', 'RainToday', 'RainTomorrow'],
```

# Missing Data



## Data Imputation

- Numerical:  
mean
- Categorical:  
mode

# Feature Importance

## Chi-Square - Categorical Features

```
Chi-square test    results
0 Pearson Chi-square ( 15.0) = 1882.7416
1 p-value = 0.0000
2 Cramer's V = 0.1151
```

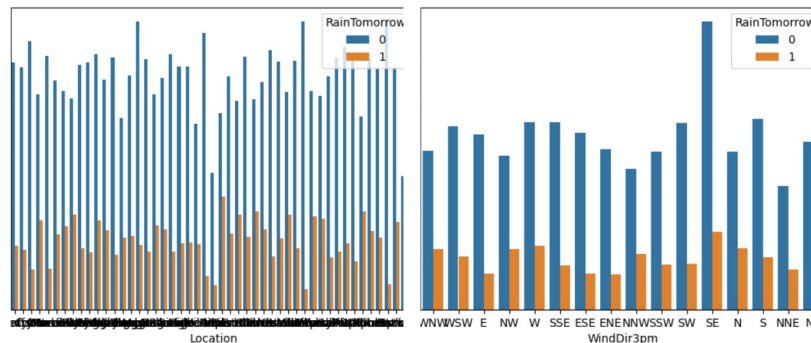
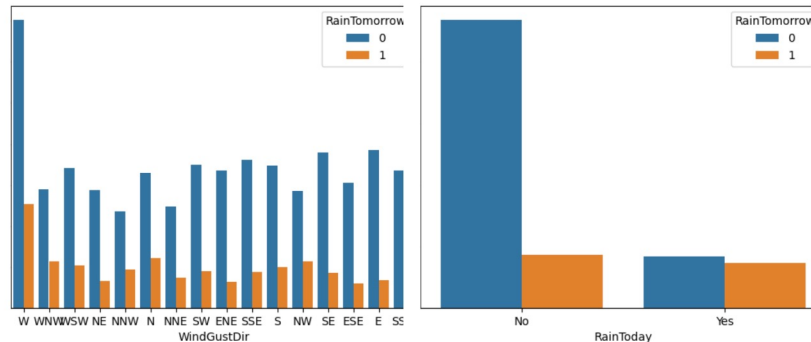
```
Chi-square test    results
0 Pearson Chi-square ( 15.0) = 1216.6671
1 p-value = 0.0000
2 Cramer's V = 0.0925
```

```
Chi-square test    results
0 Pearson Chi-square ( 1.0) = 13362.7100
1 p-value = 0.0000
2 Cramer's phi = 0.3066
```

```
Chi-square test    results
0 Pearson Chi-square ( 48.0) = 3544.7902
1 p-value = 0.0000
2 Cramer's V = 0.1579
```

## SelectKBest - Numerical Features

```
['MaxTemp', 'Rainfall', 'Sunshine', 'WindGustSpeed', 'Humidity9am',  
'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm',  
'Temp3pm', 'WindSpeed9am', 'Evaporation', 'WindSpeed3pm', 'MinTemp', 'Temp9am']
```



# Feature Handling

---

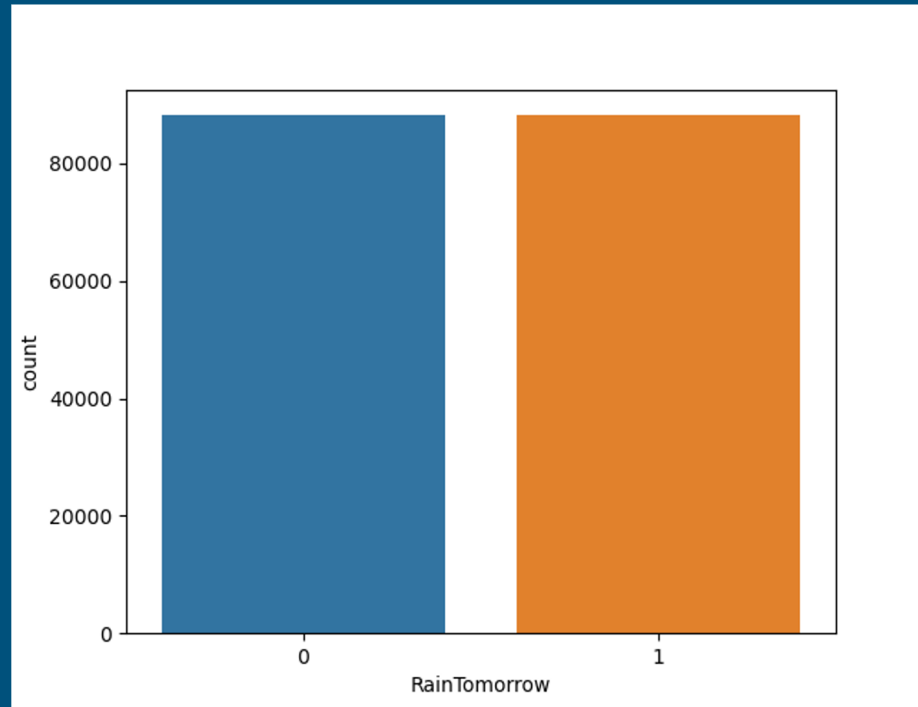
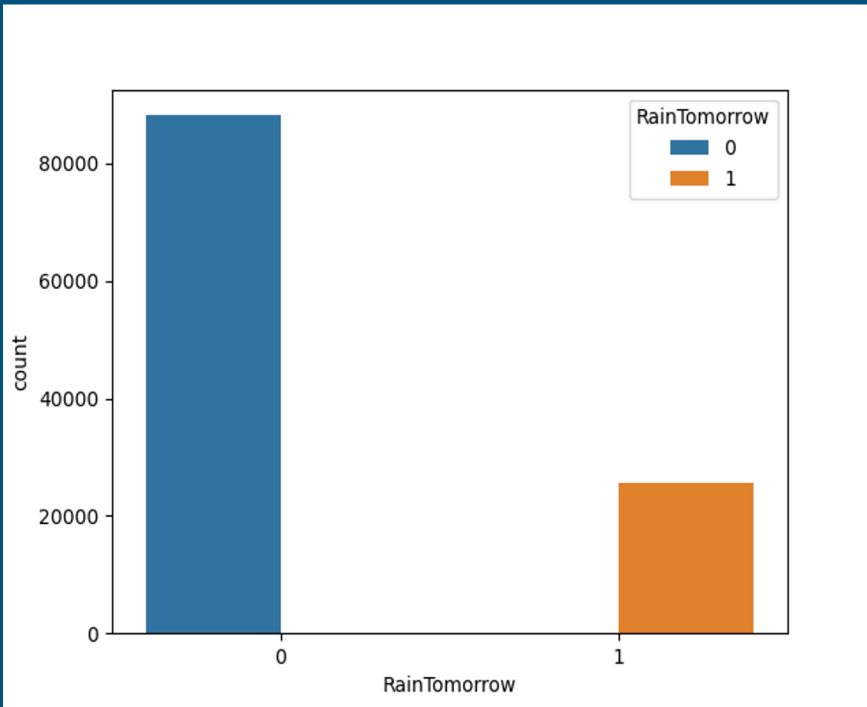
- Standard Scaling for numerical data
- One-hot Encoding for categorical data

MinTemp	MaxTemp	Rainfall	...	WindDir3pm_WSW	RainToday_No	RainToday_Yes
0.458008	0.095246	-0.278744	...	0.0	1.0	0.0
0.387859	-1.269075	0.623608	...	0.0	0.0	1.0
0.160388	0.306223	-0.278744	...	0.0	1.0	0.0
0.716808	-1.184684	-0.278744	...	0.0	1.0	0.0
0.113396	-0.143862	-0.278744	...	0.0	1.0	0.0



# Data Imbalance - SMOTE

---

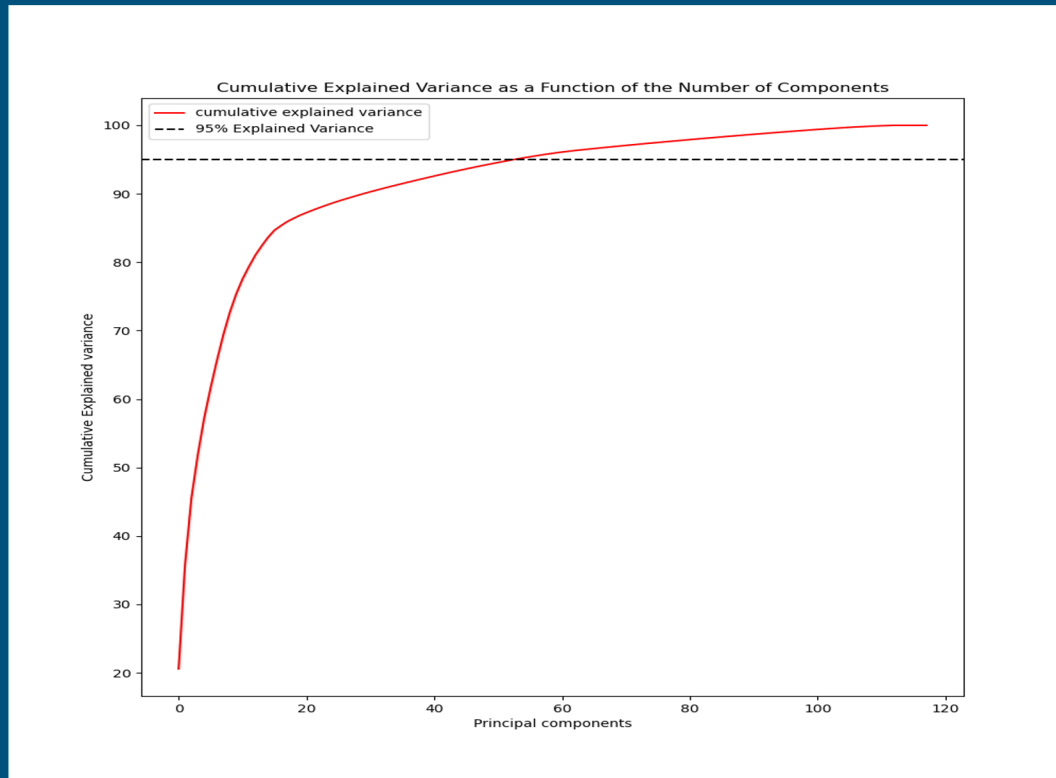


# Model Performance

---

	Model	Accuracy	ROC AUC	Precision	Recall	F1-score
	Logistic Regression	0.791484	0.866708	0.833882	0.791484	0.803847
	KNN	0.735504	0.827634	0.820171	0.735504	0.756074
	Decision Tree Classifier	0.782236	0.709302	0.792941	0.782236	0.786875
	Random Forest Classifier	0.838496	0.853481	0.830538	0.838496	0.832982
	Naive Bayes	0.624635	0.726397	0.761511	0.624635	0.656077
	MLP Classifier	0.797743	0.872896	0.836366	0.797743	0.809208
	XGB Classifier	0.856465	0.887170	0.849258	0.856465	0.850591

# Principal Component Analysis



# Model Performance after PCA

---

Model	Accuracy	ROC AUC	Precision	Recall	F1-score
Logistic Regression	0.781462	0.858268	0.828295	0.781462	0.794994
KNN	0.734133	0.824595	0.816726	0.734133	0.754639
Decision Tree Classifier	0.745490	0.692742	0.777388	0.745490	0.757363
Random Forest Classifier	0.812265	0.835040	0.815916	0.812265	0.813964
Naive Bayes	0.718274	0.758036	0.772822	0.718274	0.736181
MLP Classifier	0.808221	0.861895	0.829649	0.808221	0.815804
XGB Classifier	0.808960	0.859951	0.830101	0.808960	0.816454

# K-Fold Cross Validation

---

```
k-Fold Cross Validation:
```

```
[0.85541827 0.85071412 0.85626842 0.87043754 0.86732034 0.87134437  
0.87020348 0.86685938 0.86997676 0.87196055]
```

```
average accuracy: 0.8650503230569602
```

# Conclusion

---

- We successfully build training classification models on the target variable 'RainTomorrow' to predict next-day rain using the weather observations made on the current day.
- XGBoost Classifier model gave the best F-1 score followed by Random Forest Classifier and MLP Classifier.

Questions?

Thank You!