Zodiac Sign Prediction

Join Us on a Journey Through Stars and Algorithms!

Group 5

- Ahmed Abdulrahim
- Bimsara Geethachapa Siman Meru Pathiranage
- Simranjeet Kaur
- Efemena Theophilus Edoja
- Simran

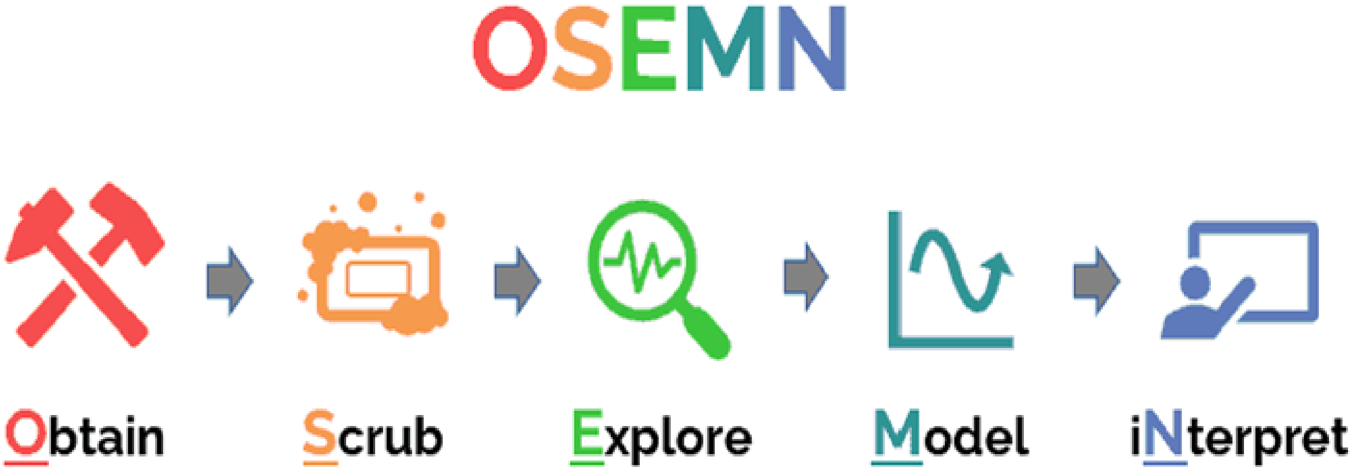
Unveiling the essence

Astrology, a timeless art, has always intrigued humanity with its celestial insights.Our project takes this ancient wisdom into the digital age by harnessing Python programming and machine learning techniques to predict Zodiac signs.

The core objective of this project is to utilize machine learning and natural language processing methods to forecast the astrological signs of bloggers by analyzing their written content.

Methodology

OSEMN offers a structured approach that encompasses five essential stages, ensuring a comprehensive and effective exploration of the data:

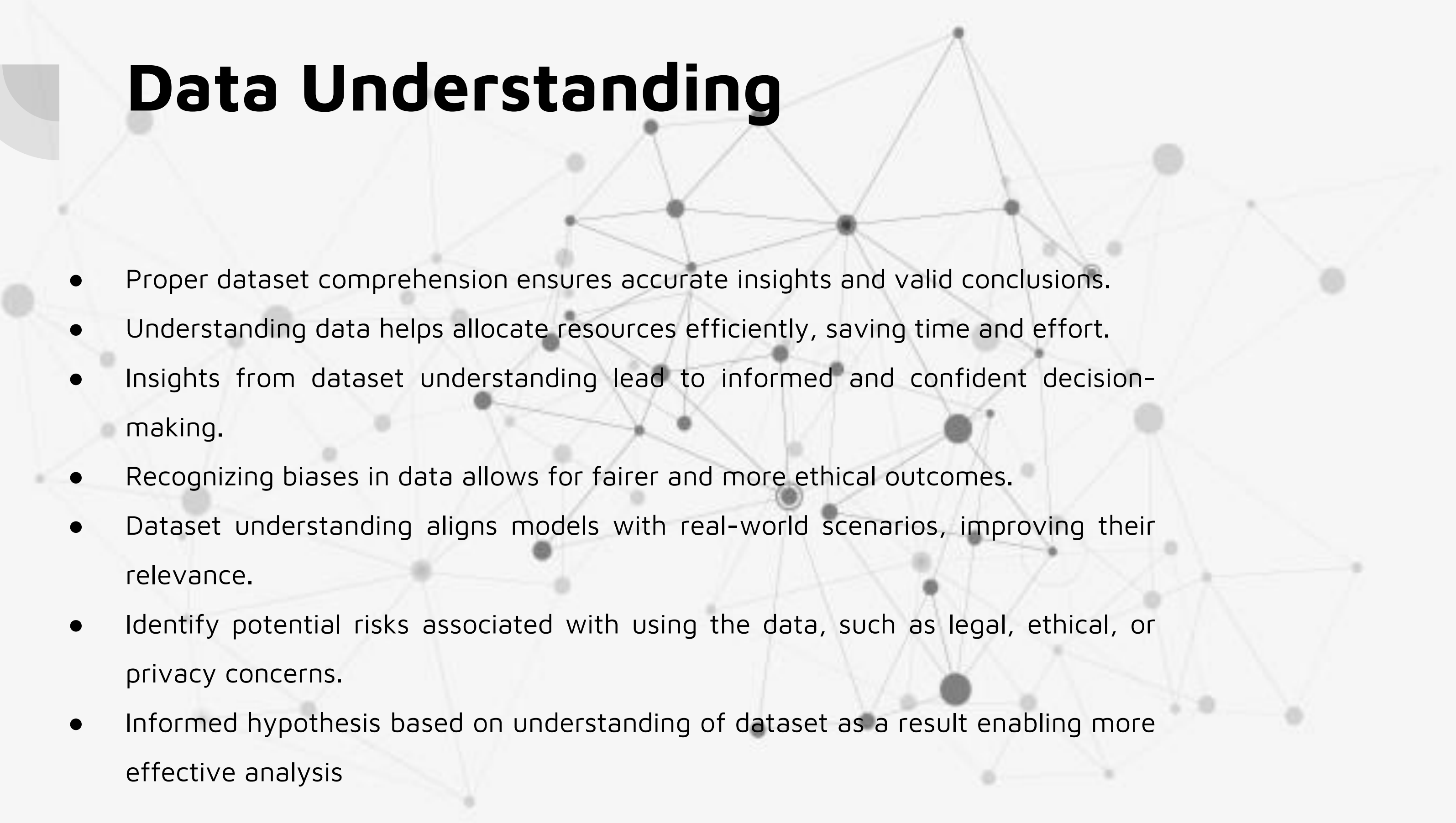


Obtain: Business Understanding

Businesses can unlock a range of valuable insights:

- Personalized Marketing
- Content Recommendation
- Product Development
- Customer Relationship Management
- Market Segmentation
- Competitive Edge
- Content Creation

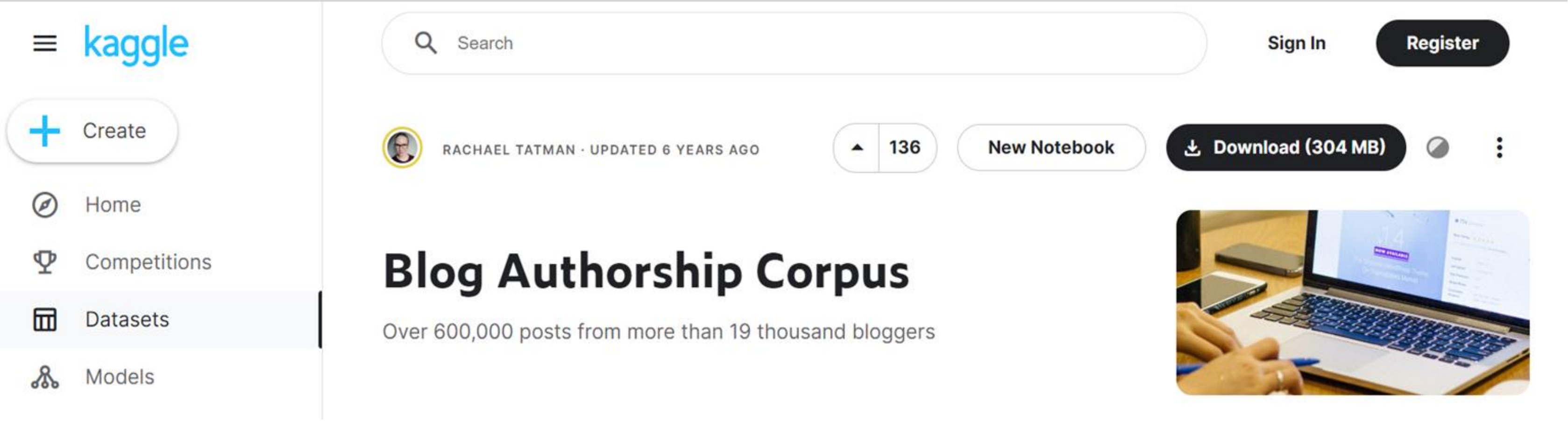
Data Understanding



- Proper dataset comprehension ensures accurate insights and valid conclusions.
- Understanding data helps allocate resources efficiently, saving time and effort.
- Insights from dataset understanding lead to informed and confident decision-making.
- Recognizing biases in data allows for fairer and more ethical outcomes.
- Dataset understanding aligns models with real-world scenarios, improving their relevance.
- Identify potential risks associated with using the data, such as legal, ethical, or privacy concerns.
- Informed hypothesis based on understanding of dataset as a result enabling more effective analysis

- Our dataset originates from Kaggle, a respected platform known for sharing datasets within the data science and machine learning domain.
- Kaggle serves as a reliable hub where data enthusiasts share datasets, making valuable resources accessible to the community.
- For our zodiac sign prediction project, we've specifically chosen a dataset from Kaggle that aligns seamlessly with our objectives.

https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus



RACHAEL TATMAN · UPDATED 6 YEARS AGO

Blog Authorship Corpus

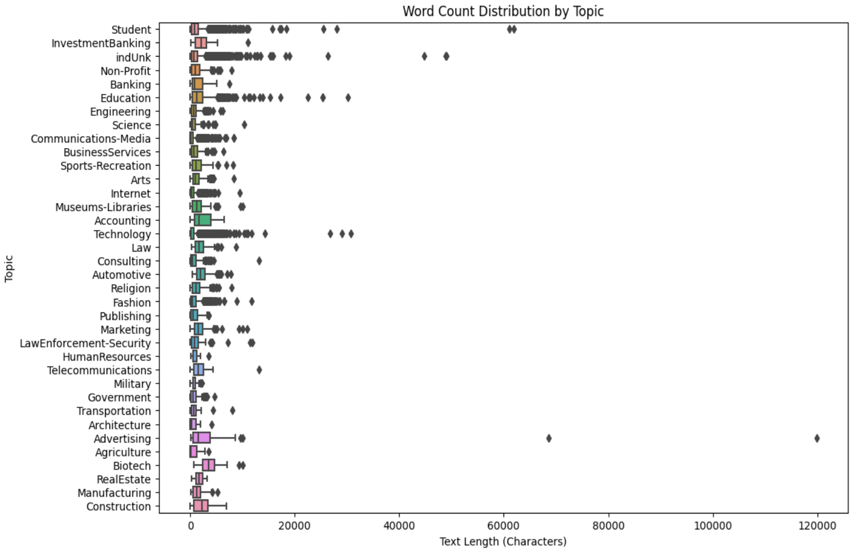
Over 600,000 posts from more than 19 thousand bloggers

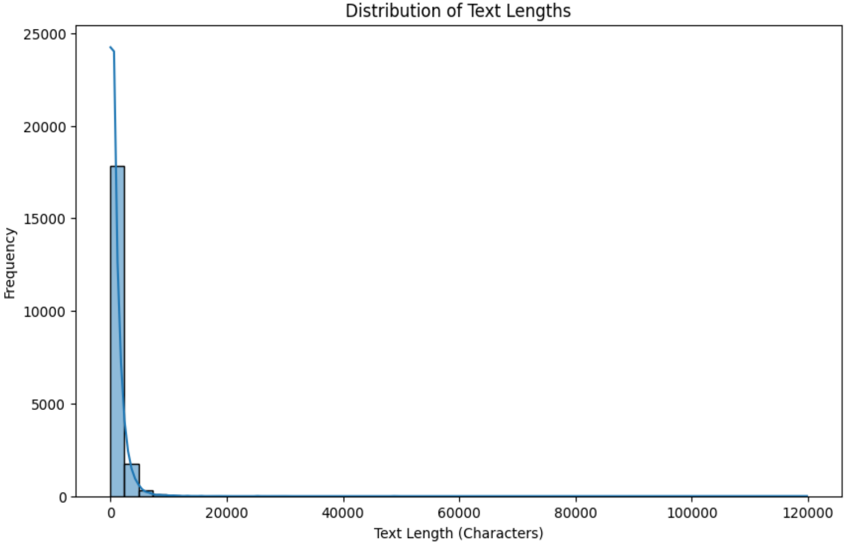
Why this dataset?

- **Data Source Authenticity**: Ensuring the authenticity of the data source is essential to maintain the integrity of the project.
- **Data Size and Diversity**: Provides a substantial amount of textual content for analysis.
- **Zodiac Sign Labels**: Fundamental requirement of our Project
- **Textual Content**: Helps identify writing patterns specific to each zodiac sign.
- **Age and Gender Distribution**: Offers potential insights into how writing styles may vary across demographics.
- **Data Preprocessing Potential**: Provides room for various preprocessing techniques, such as cleaning, tokenization, and lemmatization.
- **Ethical Considerations**: The dataset adheres to ethical guidelines, ensuring the privacy and anonymity of the bloggers.

Explore: Data Preparation

- Data Cleaning
- Text Preprocessing
- Label Encoding
- Feature Extraction





Data Preparation - Explore Phase

Analysis:

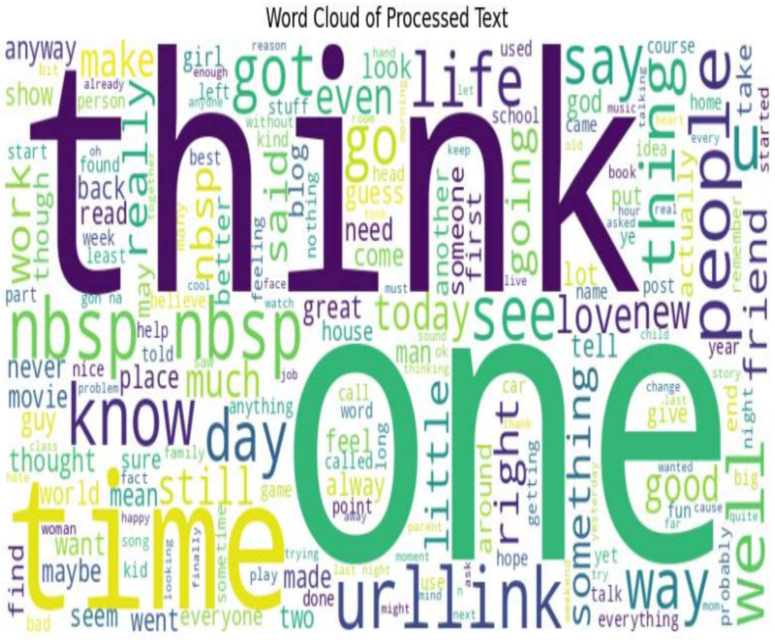
- Explore zodiac sign distribution and attribute summary statistics.
- Visualize data distributions and frequency of zodiac signs.

Feature Selection:

- Focus on significant words/phrases associated with specific signs.

Visualization:

- Create count plots for zodiac sign occurrences in the dataset.



Word Cloud of Processed Text

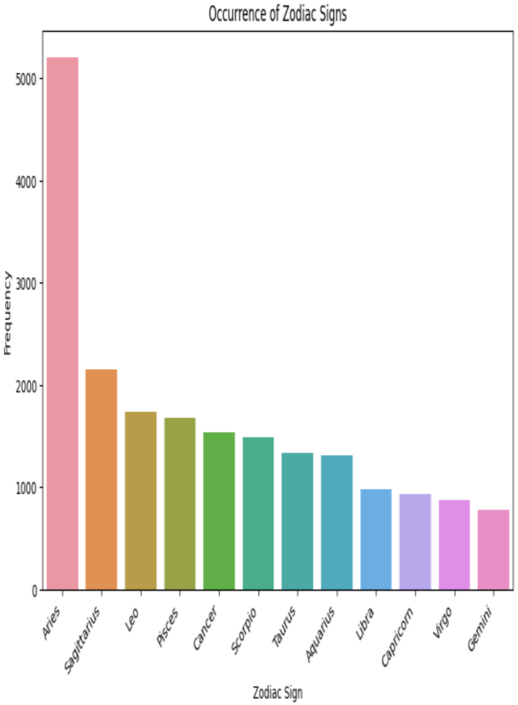
Impact of Data Preparation and Exploration

Significance:

- Ensures accurate, clean, and well-structured data for analysis.
- Identifies meaningful features to enhance model performance.
- Provides insights into zodiac sign distribution and patterns.

Contribution:

- Improves accuracy and effectiveness of Zodiac Sign Prediction models.
- Informs subsequent stages of model development and evaluation.
- Enables data-driven decision making for reliable predictions.



Model

Random Forest Classifier:

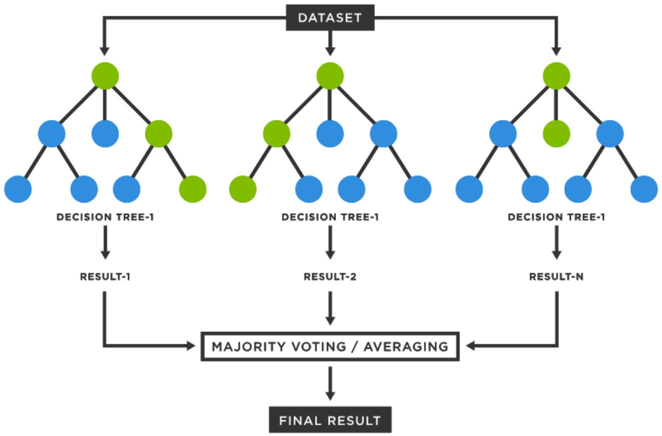
- Ensemble Learning
- Feature Importance
- Robustness

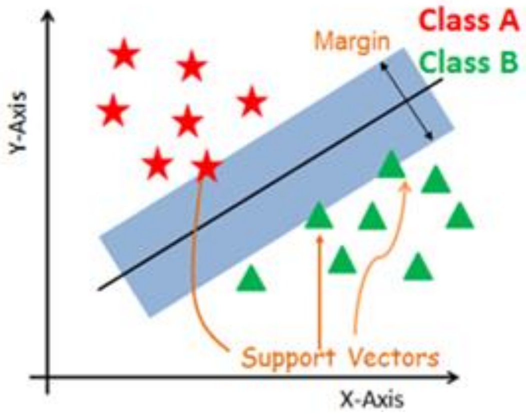
<u>Why We Chose It</u>:

- Ability to handle complex relationships
- Mitigate overfitting
- Provide feature importance analysis

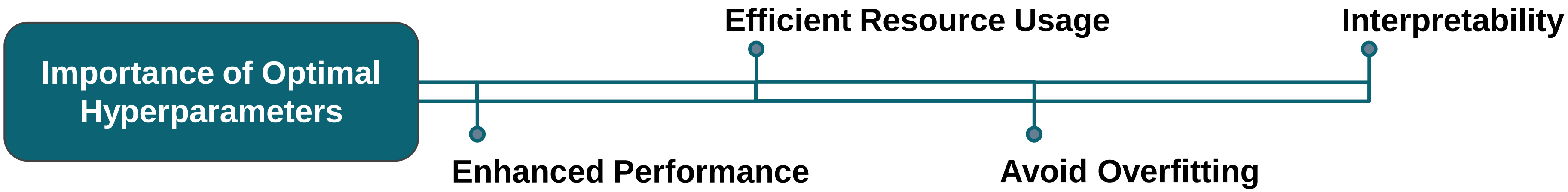
Linear Support Vector Classifier:

- Effective Separation
- Margin Maximization





Hyperparameter Tuning



Importance of Optimal Hyperparameters

Enhanced Performance

Efficient Resource Usage

Avoid Overfitting

Interpretability

Search	Parameters	Best Parameters

Randomized Search CV	`C: 0.01, 0.1, 1, 10 Regularization parameter`
	`Loss: 'hinge','squared_hinge' Loss function`
	`Max_iter: randint(100, 500) Maximum number of iterations`
		`C: 1`
		`loss: hinge`
		`Max_iter: 350`
		`Best Accuracy: 0.5416`

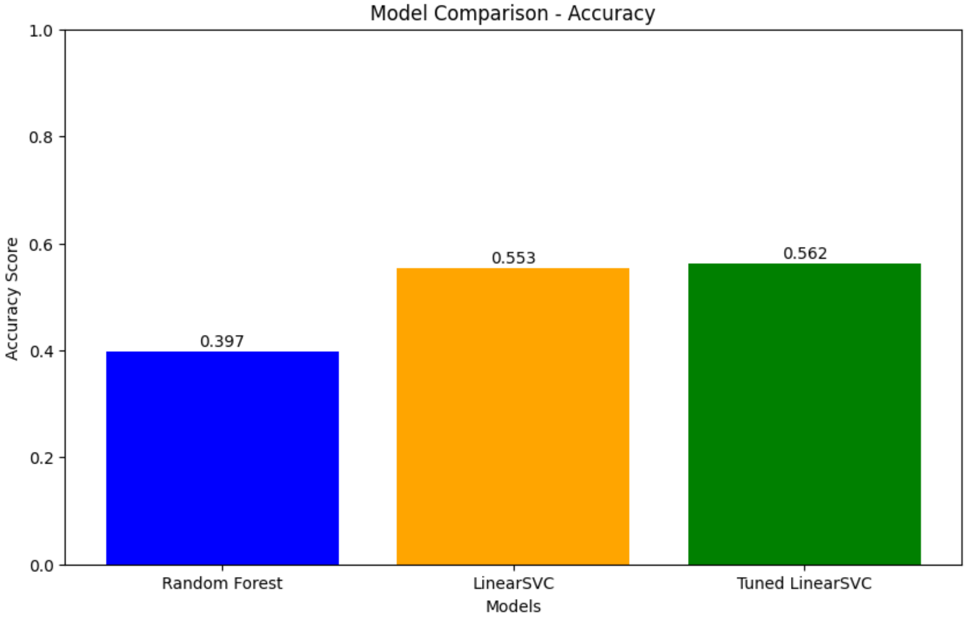
Grid Search	`C: 0.8, 0.9, 1, 1.1, 1.2`
	`loss: 'hinge' (Use the best 'loss' value)`
	`max_iter: 300, 350, 400`
		`C: 1.2`
		`max_iter: 300`
		`Best Accuracy: 0.5433`

Interpretation

To measure performance of each model we used

Accuracy score

Model	Random Forest	Linear SVC	Tuned Linear SVC
Accuracy score	0.397	0.553	0.562

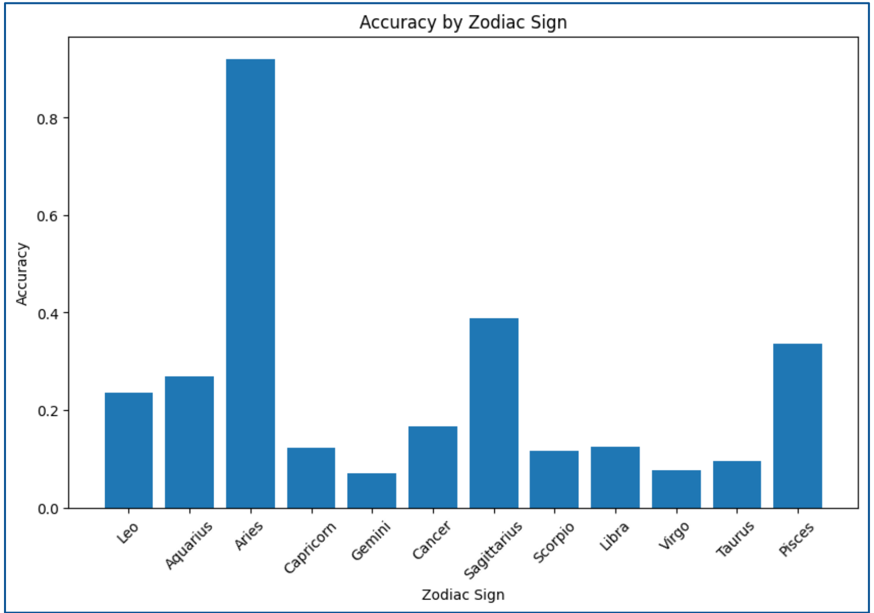


Random Forest Classifier

Top Accuracy: Aries

lowest accuracy: Gemini

Analyzing accuracy by zodiac sign provides valuable insights into the performance of our prediction model and suggests potential areas for improvement



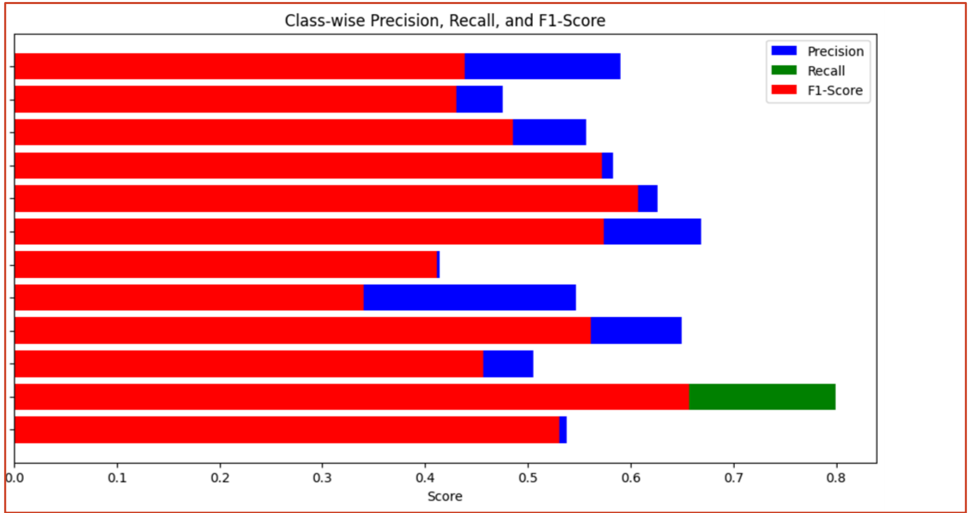
Linear SVC

Highest Precision: Gemini

Highest Recall: Aquarius

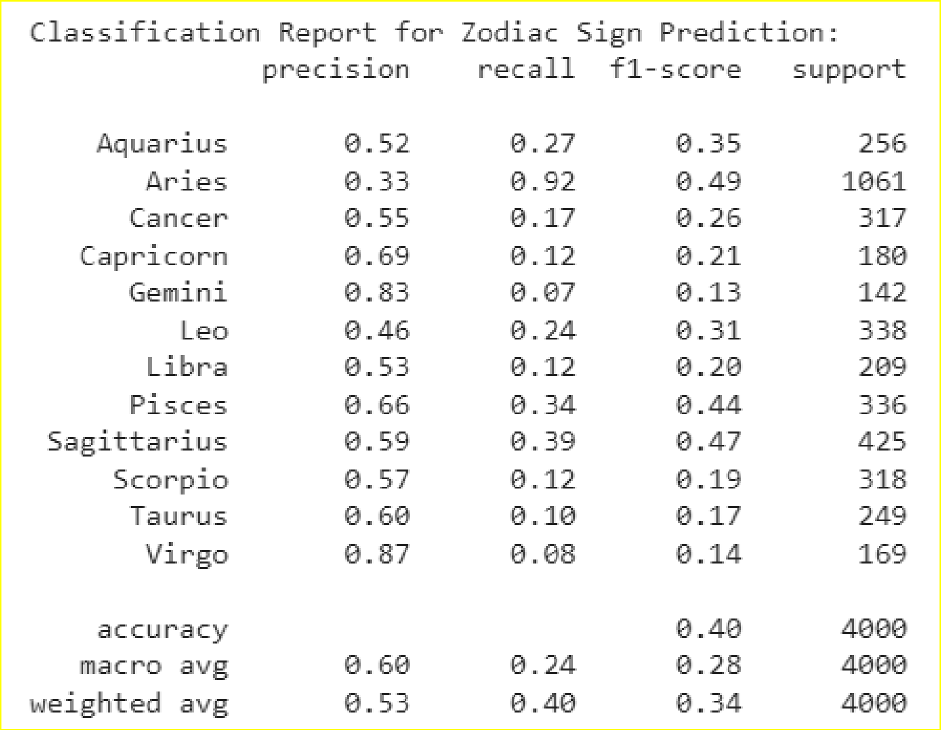
Highest F1-Score: Aquarius

Analyzing class-wise precision, recall, and F1-score provides a comprehensive understanding of our model's performance across different zodiac signs



To measure performance of each model we used

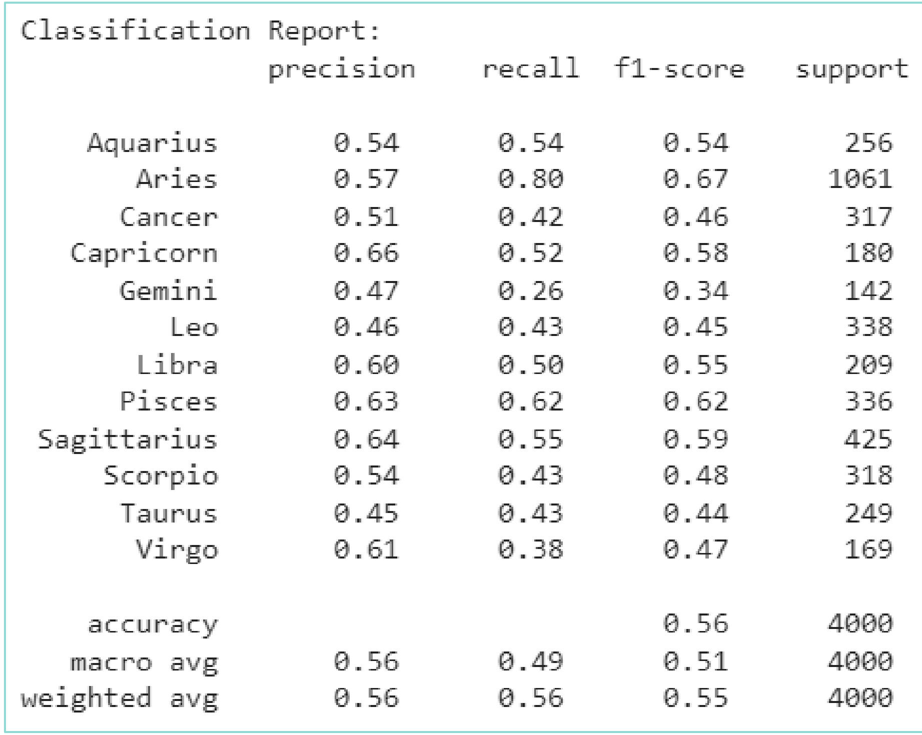
Classification report



```
Classification Report for Zodiac Sign Prediction:
              precision    recall  f1-score   support

    Aquarius       0.52      0.27      0.35       256
       Aries       0.33      0.92      0.49      1061
      Cancer       0.55      0.17      0.26       317
   Capricorn       0.69      0.12      0.21       180
      Gemini       0.83      0.07      0.13       142
         Leo       0.46      0.24      0.31       338
       Libra       0.53      0.12      0.20       209
      Pisces       0.66      0.34      0.44       336
 Sagittarius       0.59      0.39      0.47       425
     Scorpio       0.57      0.12      0.19       318
      Taurus       0.60      0.10      0.17       249
       Virgo       0.87      0.08      0.14       169

    accuracy                           0.40      4000
   macro avg       0.60      0.24      0.28      4000
weighted avg       0.53      0.40      0.34      4000
```



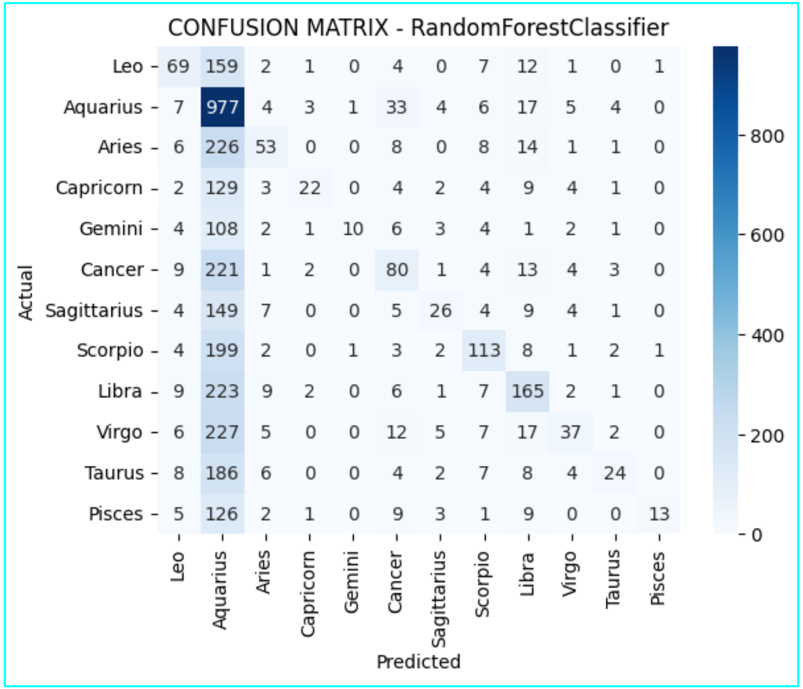
```
Classification Report:
              precision    recall  f1-score   support

    Aquarius       0.54      0.54      0.54       256
       Aries       0.57      0.80      0.67      1061
      Cancer       0.51      0.42      0.46       317
   Capricorn       0.66      0.52      0.58       180
      Gemini       0.47      0.26      0.34       142
         Leo       0.46      0.43      0.45       338
       Libra       0.60      0.50      0.55       209
      Pisces       0.63      0.62      0.62       336
 Sagittarius       0.64      0.55      0.59       425
     Scorpio       0.54      0.43      0.48       318
      Taurus       0.45      0.43      0.44       249
       Virgo       0.61      0.38      0.47       169

    accuracy                           0.56      4000
   macro avg       0.56      0.49      0.51      4000
weighted avg       0.56      0.56      0.55      4000
```

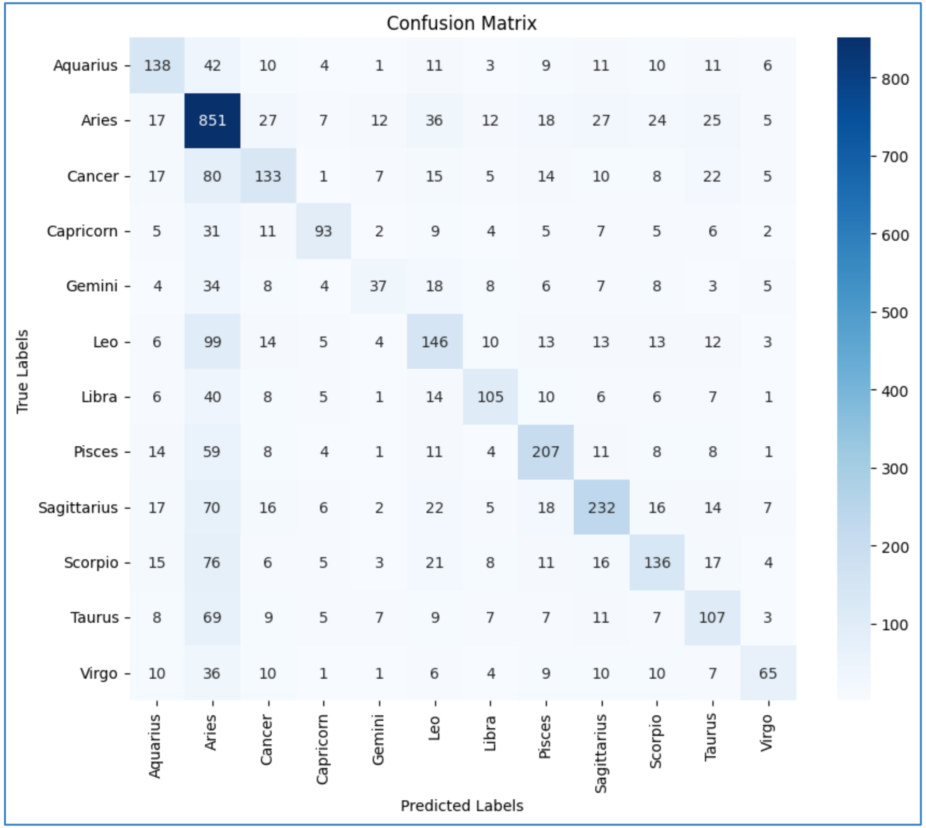
To measure performance of each model we used

Confusion Matrix



CONFUSION MATRIX - RandomForestClassifier

	Leo	Aquarius	Aries	Capricorn	Gemini	Cancer	Sagittarius	Scorpio	Libra	Virgo	Taurus	Pisces
Leo	69	159	2	1	0	4	0	7	12	1	0	1
Aquarius	7	977	4	3	1	33	4	6	17	5	4	0
Aries	6	226	53	0	0	8	0	8	14	1	1	0
Capricorn	2	129	3	22	0	4	2	4	9	4	1	0
Gemini	4	108	2	1	10	6	3	4	1	2	1	0
Cancer	9	221	1	2	0	80	1	4	13	4	3	0
Sagittarius	4	149	7	0	0	5	26	4	9	4	1	0
Scorpio	4	199	2	0	1	3	2	113	8	1	2	1
Libra	9	223	9	2	0	6	1	7	165	2	1	0
Virgo	6	227	5	0	0	12	5	7	17	37	2	0
Taurus	8	186	6	0	0	4	2	7	8	4	24	0
Pisces	5	126	2	1	0	9	3	1	9	0	0	13



Confusion Matrix

	Aquarius	Aries	Cancer	Capricorn	Gemini	Leo	Libra	Pisces	Sagittarius	Scorpio	Taurus	Virgo
Aquarius	138	42	10	4	1	11	3	9	11	10	11	6
Aries	17	851	27	7	12	36	12	18	27	24	25	5
Cancer	17	80	133	1	7	15	5	14	10	8	22	5
Capricorn	5	31	11	93	2	9	4	5	7	5	6	2
Gemini	4	34	8	4	37	18	8	6	7	8	3	5
Leo	6	99	14	5	4	146	10	13	13	13	12	3
Libra	6	40	8	5	1	14	105	10	6	6	7	1
Pisces	14	59	8	4	1	11	4	207	11	8	8	1
Sagittarius	17	70	16	6	2	22	5	18	232	16	14	7
Scorpio	15	76	6	5	3	21	8	11	16	136	17	4
Taurus	8	69	9	5	7	9	7	7	11	7	107	3
Virgo	10	36	10	1	1	6	4	9	10	10	7	65

Comparison and Insights

- Both Random Forest Classifier and LinearSVC models showed strong predictive abilities for Zodiac Sign Prediction.
- After hyperparameter tuning, LinearSVC achieved slightly higher accuracy, showcasing its effectiveness in this task.
- Algorithm choice depends on factors like complexity, interpretability, and efficiency.
- Random Forest excelled in capturing complex relationships, while LinearSVC leveraged text features.

Future Enhancements

- Fine-tuning hyperparameters remains an avenue for improved results."
- Feature engineering can enhance predictive power via techniques like embeddings.
- Model ensemble, combining Random Forest and LinearSVC, could boost accuracy.
- Incorporating external data sources could provide richer context for predictions.
- Advanced NLP techniques like sentiment analysis can further enrich text data.
- Online deployment, multilingual support, and ethical considerations are key for growth.

Thank You !