



# VARIATIONAL AUTOENCODER (VAE) ANALYSIS

AMIRHOSSEIN  
GHANAATIAN

NEW DATASETS  
OLD DATASETS

23 CRS F1 AND HO COMBINED AGRON AND SPECTRAL DATA\_WEATHER TO CARAGEA 9\_16\_24

TO CARAGEA WORKING 2022\_2023 ALL LOC\_KPEGG\_MULTISPECTRAL\_STACKED  
TO CARAGEA WORKING 2023 ALL LOC\_KPEGG\_THERMAL\_STACKED



# DATA PREPROCESSING

- Selected relevant columns
- Sorted data based on 'expt', 'plot', and 'entry'
- Replaced "." with null values
- Converted columns to numeric format
- Aggregated data by grouping and calculating means
- Added 'yield' and 'stage' columns back
- Reordered columns

# AGGREGATED DATA BY GROUPING AND CALCULATING MEANS

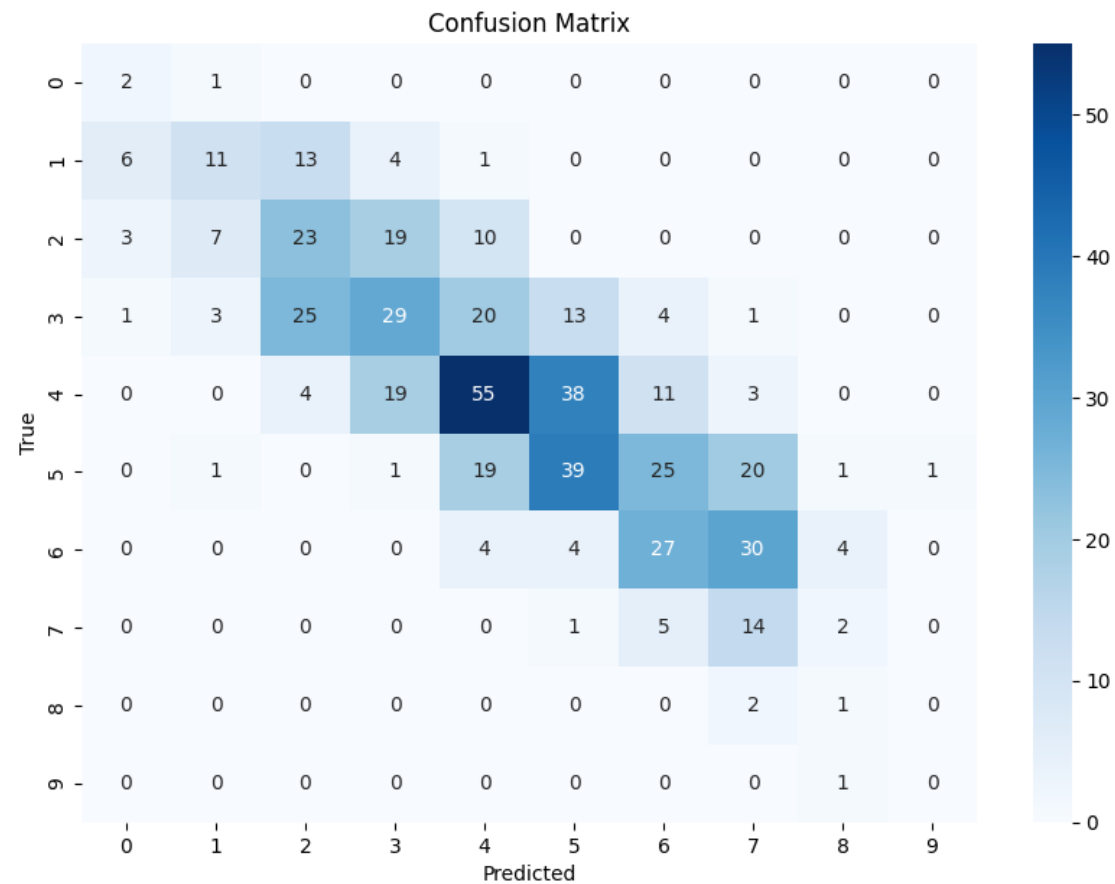
E10	=AVERAGEIF(E2:E9,"<>",E2:E9)										
	A	B	C	D	E	F	G	H	I	J	K
1	expt	plot	entry	yield	CC	ndvi	blu	gre	red	re	nir
2	22 KPE GG F2	1321	621	40.3440678	30388	0.82174715	0.01098792	0.02802799	0.01337262	0.01098792	0.13666843
3	22 KPE GG F2	1321	621	40.3440678		0.86837615	0.0129881	0.03221868	0.01321263	0.13637333	0.18755086
4	22 KPE GG F2	1321	621	40.3440678		0.87546337	0.01012124	0.02541223	0.01187734	0.11676214	0.17886724
5	22 KPE GG F2	1321	621	40.3440678		0.88695479	0.01092556	0.02367702	0.01014531	0.10361819	0.16934593
6	22 KPE GG F2	1321	621	40.3440678		0	0.01058852	0.02319826	0.02319826	0.02319826	0.02319826
7	22 KPE GG F2	1321	621	40.3440678		0.88070039	0.01240885	0.02690455	0.01254708	0.13273332	0.19779861
8	22 KPE GG F2	1321	621	40.3440678		0.77044222	0.01598618	0.04152398	0.02389031	0.14763777	0.18425173
9	22 KPE GG F2	1321	621	40.3440678		0.47161309	0.0161965	0.02680502	0.0365112	0.07100984	0.10168753
10					30388	0.69691215	0.01252536	0.02847096	0.01809434	0.0927901	0.14742107

# VAE IMPLEMENTATION

- Implemented a Variational Autoencoder on the preprocessed data
- Performance metrics:

Metric	Value
Test Mean Squared Error	45.9015
Test R <sup>2</sup>	0.7479
Test Mean Absolute Error	5.1987
Test Mean Percentage Error	13.5262%

# CONFUSION MATRIX



# CONFUSION MATRIX ANALYSIS

- Multi-class classification model performance (classes 0-9)
- Best performance for classes 4, 5, 6, 7
- Class 9: insufficient data, misclassified as class 8
- Classes 0, 1, 2: few instances, class imbalance, similarity
- Misclassifications more common between adjacent classes
- Higher precision for middle classes, better recall for classes 6, 7



# DATA PREPROCESSING FOR STAGE

- Drop the Rows which 'Stage' is null
- Divided the dataset based on the "Stage" columns
- Preprocess every sub-dataset based on number of nulls
- If a column has 100% null, we remove that column in training the models
- If percentage of columns is low, we remove only rows with nulls in that columns

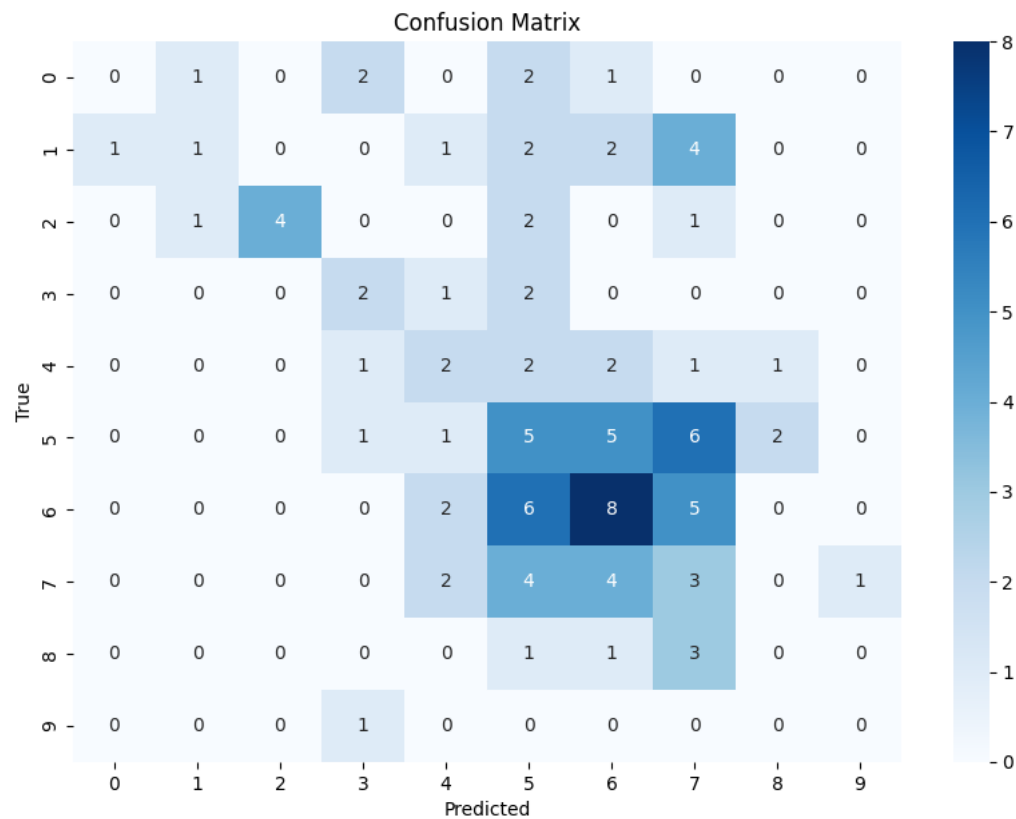
# R1

Percentage of nulls in each column:

yield	0.0
stage	0.0
CC	100.0
ndvi	0.0
blu	0.0
gre	0.0
red	0.0
re	0.0
nir	0.0
expt	0.0
plot	0.0
entry	0.0
loc	0.0
prow	0.0
pcol	0.0

Ignoring CC Column,  
Trained on: 'ndvi', 'blu', 'gre', 'red', 're', 'nir'

Test R<sup>2</sup>: 0.1278





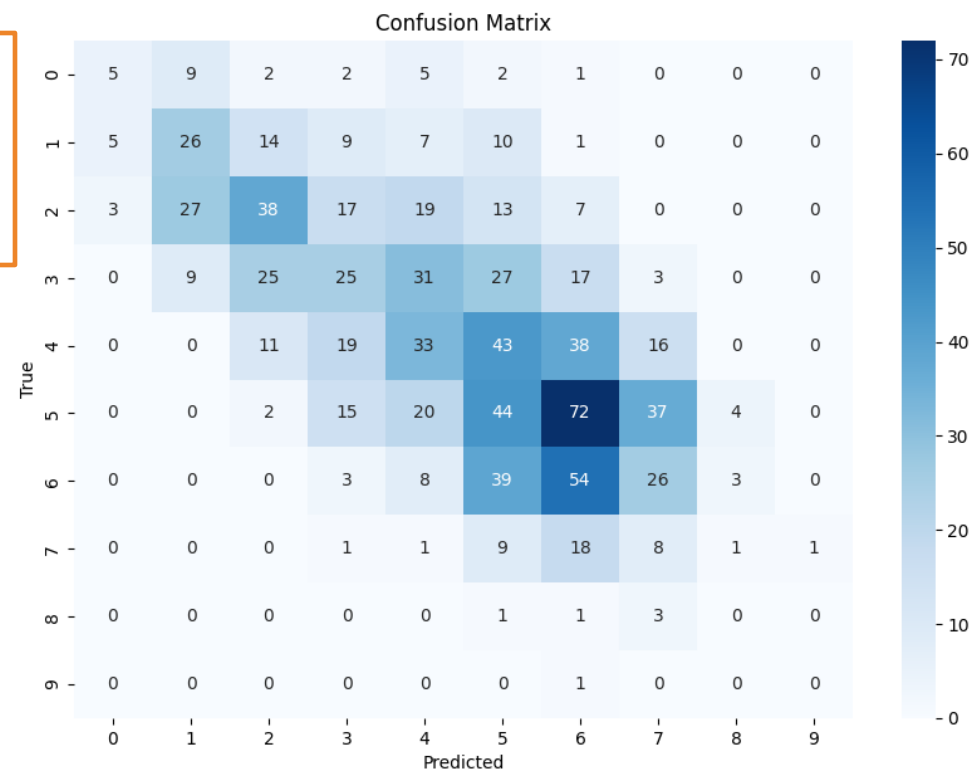
# R5

Percentage of missing values in each column:

yield	1.980198
stage	0.000000
CC	32.673267
ndvi	0.000000
blu	0.000000
gre	0.000000
red	0.000000
re	0.000000
nir	0.000000
expt	0.000000
plot	1.980198
entry	1.980198
loc	0.000000
prow	0.000000
pcol	0.000000

Ignoring CC Column, Removing rows with nulls in Yield  
Trained on: 'ndvi', 'blu', 'gre', 'red', 're', 'nir'

Test R<sup>2</sup>: 0.3878



# R5

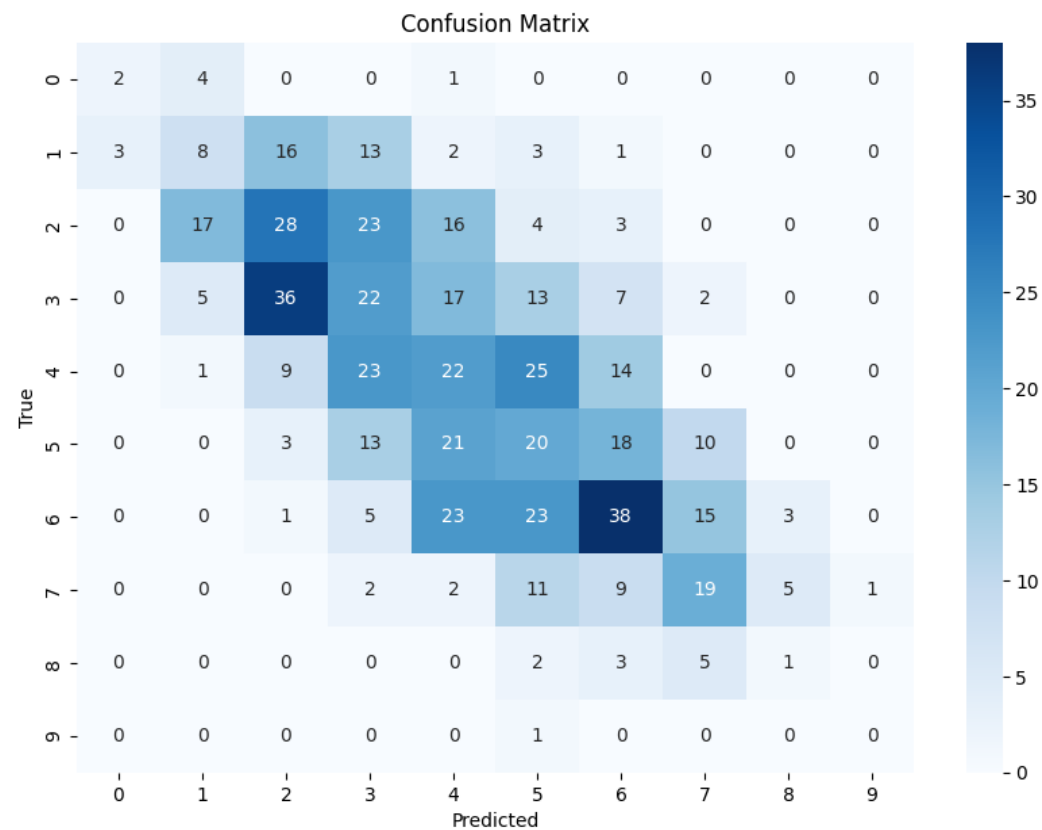
Percentage of missing values in each column:

yield 1.980198  
stage 0.000000  
CC 32.673267  
ndvi 0.000000  
blu 0.000000  
gre 0.000000  
red 0.000000  
re 0.000000  
nir 0.000000  
expt 0.000000  
plot 1.980198  
entry 1.980198  
loc 0.000000  
prow 0.000000  
pcol 0.000000

Removing rows with nulls in Yield and CC

Trained on: 'CC', 'ndvi', 'blu', 'gre', 'red', 're', 'nir'

Test R<sup>2</sup>: 0.5063



# SUMMARY

Stage	Which Column Ignored?	Null Cleaning on?	nu. of rows after preprocessing	R2 (test, 20%)
R1	CC	-	660	0.1278
R2	CC	yield	3300	0.1041
R3		yield	1320	0.7206
R4		yield, CC	1320	0.024
R4	CC	yield	1980	0.1092
R5	CC	yield	5940	0.3878
R5		yield, CC	3960	0.5063
R6	CC	-	1320	0.633
R7	CC	yield	7260	0.6245
R8	CC	-	660	0.1296
V4	-	ndvi', 'blu', 'gre', 'red', 're', 'nir'	0	-
V6.5	CC	-	660	0.3291

# DATA PREPROCESSING

23 CRS F1 AND HO COMBINED AGRON AND SPECTRAL DATA\_WEATHER TO CARAGEA 9\_16\_24.XLSX

- Selected relevant columns
- Sorted data based on 'expt', 'plot', and 'entry'
- Replaced "." with null values
- Converted columns to numeric format
- Aggregated data by grouping and calculating means

# MORE DETAILS

- all P4 data = [cc, blu, gre, red, re, nir, ndvi]
  - 4800 rows
- all Thermal data = [TH]
  - 3072 rows
- Agron = [Yeild]
  - 384 rows
- Merge these data based on the unique match of 'expt', 'plot', and 'entry', remove rows with nulls
  - 381 rows

# AGGREGATED DATA BY GROUPING AND CALCULATING MEANS

E10	fx =AVERAGEIF(E2:E9,"<>",E2:E9)										
	A	B	C	D	E	F	G	H	I	J	K
1	expt	plot	entry	yield	CC	ndvi	blu	gre	red	re	nir
2	22 KPE GG F2	1321	621	40.3440678	30388	0.82174715	0.01098792	0.02802799	0.01337262	0.01098792	0.13666843
3	22 KPE GG F2	1321	621	40.3440678		0.86837615	0.0129881	0.03221868	0.01321263	0.13637333	0.18755086
4	22 KPE GG F2	1321	621	40.3440678		0.87546337	0.01012124	0.02541223	0.01187734	0.11676214	0.17886724
5	22 KPE GG F2	1321	621	40.3440678		0.88695479	0.01092556	0.02367702	0.01014531	0.10361819	0.16934593
6	22 KPE GG F2	1321	621	40.3440678		0	0.01058852	0.02319826	0.02319826	0.02319826	0.02319826
7	22 KPE GG F2	1321	621	40.3440678		0.88070039	0.01240885	0.02690455	0.01254708	0.13273332	0.19779861
8	22 KPE GG F2	1321	621	40.3440678		0.77044222	0.01598618	0.04152398	0.02389031	0.14763777	0.18425173
9	22 KPE GG F2	1321	621	40.3440678		0.47161309	0.0161965	0.02680502	0.0365112	0.07100984	0.10168753
10					30388	0.69691215	0.01252536	0.02847096	0.01809434	0.0927901	0.14742107

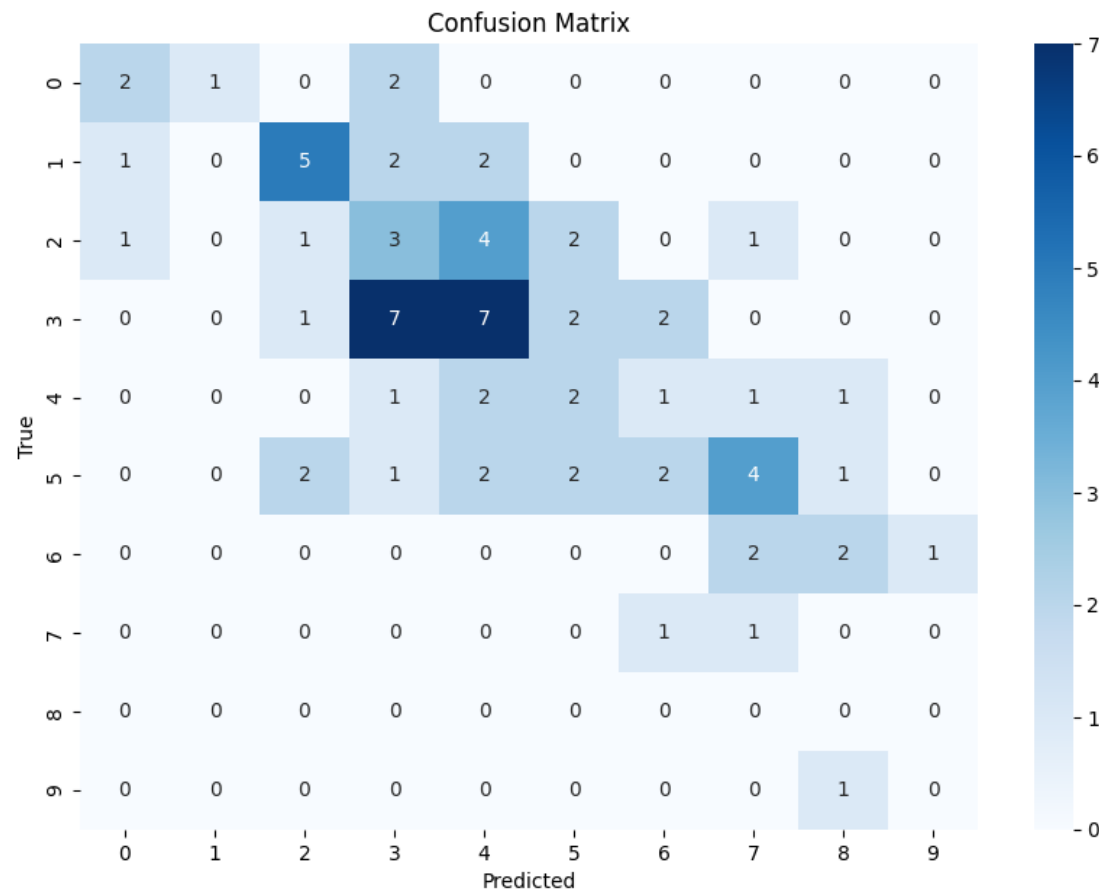
# VAE IMPLEMENTATION

- Implemented a Variational Autoencoder
- K-fold cross-validation results (K=5)
- Average Metrics
  - Mean Squared Error (MSE): 60.1191
  - R<sup>2</sup> Score: 0.5313

Fold	MSE	R <sup>2</sup>	MAE	MPE
1	70.2446	0.4908	6.0127	13.2878%
2	39.6278	0.7020	5.1391	11.2822%
3	61.5302	0.5877	6.1645	12.9856%
4	64.1423	0.3772	5.9125	12.7816%
5	65.0506	0.4986	5.9600	12.5112%

$$\text{MPE} = \frac{1}{n} \sum_{i=1}^n \frac{|actual_y - predicted_y|}{actual_y} \times 100\%$$

# CONFUSION MATRIX





# RESULTS FOR "SAS" DATA

23 CRS F1 AND HO COMBINED AGRON AND SPECTRAL DATA\_WEATHER TO CARAGEA 9\_16\_24.XLSX

## Model Performance Evaluation

K-fold cross-validation results (K=5)

### Average Metrics

- Mean Squared Error (MSE): 53.0747
- $R^2$  Score: 0.5893

### Results by Fold

Fold	MSE	$R^2$	MAE	MPE
1	69.3842	0.4971	6.1160	13.5881%
2	30.8453	0.7681	4.4166	9.1228%
3	55.6282	0.6272	5.6451	12.2313%
4	50.9172	0.5056	5.3736	11.7496%
5	58.5982	0.5483	5.7079	11.4970%

# DATA OVERVIEW

- Old Data
  - Multispectral
  - Thermal
- Data Aggregation
  - Aggregate based on experiment plot entry
  - Calculate averages for:
    - NDVI
    - RED
    - GREEN
    - ...
    - THERMAL
- Data Cleaning
  - Row count mismatch between multispectral and thermal data
  - Remove extra multispectral entries to synchronize
  - Final row count after cleaning: 1980 rows
  - Merge based on experiment plot entry
    - To have Thermal and Multispectral together

# MODEL PERFORMANCE

- Old DATA Performance

- Metrics

- Test Mean Squared Error: 53.6881
    - Test  $R^2$ : 0.7681
    - Test Mean Absolute Error: 5.7176
    - Test Mean Percentage Error: 14.0103%

- New DATAPerformance

- Metrics

- Test Mean Squared Error: 68.9005
    - Test  $R^2$ : 0.5616
    - Test Mean Absolute Error: 6.5699
    - Test Mean Percentage Error: 14.6783%

# MODEL TRANSFER TEST

- Old Data
  - 90% used for training
  - 10% used for testing
  - Model saved after training
- Old Data TEST
  - $MSE = 54$
  - $R^2 = 0.976$
- New Data TEST (ALL 381 ROWS)
  - Saved model tested on new data
  - $R^2 = -2.5$
  - $MSE = 468$

# NEGATIVE R-SQUARED?

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$
$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$
$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

For example, a model with an R-squared value of 0.9 means that approximately 90% of the variance in the dependent variable is explained by the independent variables. This suggests a strong relationship between the variables and indicates that the model provides a good fit to the data.

$R^2$  is negative only when the chosen model does not follow the trend of the data, so fits worse than a horizontal line.

When  $SS_{\text{res}}$  is greater than  $SS_{\text{tot}}$ , that equation could compute a negative value for  $R^2$ , if the value of the coefficient is greater than 1.

It simply means that the chosen model (with its constraints) fits the data really poorly.