Predict Number of Rented Bikes each hour

- Team Members:

name	ID
Norhan Mahmoud Mohamed	2018170833
Amira Mohmed Gomaa	2018170721
Ganna Ayman Esmail	2018170729
Esraa Mohamed Ali	2018170713
Yomna Abdelsamed Abdelaal	2018170844

1. Introduction

A company needs to determine the number of rented bikes for each hour, depending in some features

Like: Temperature(آ°C) , Humidity(%),Wind speed (m/s), Visibility (10m), Functioning Day,….etc and We will predict it.

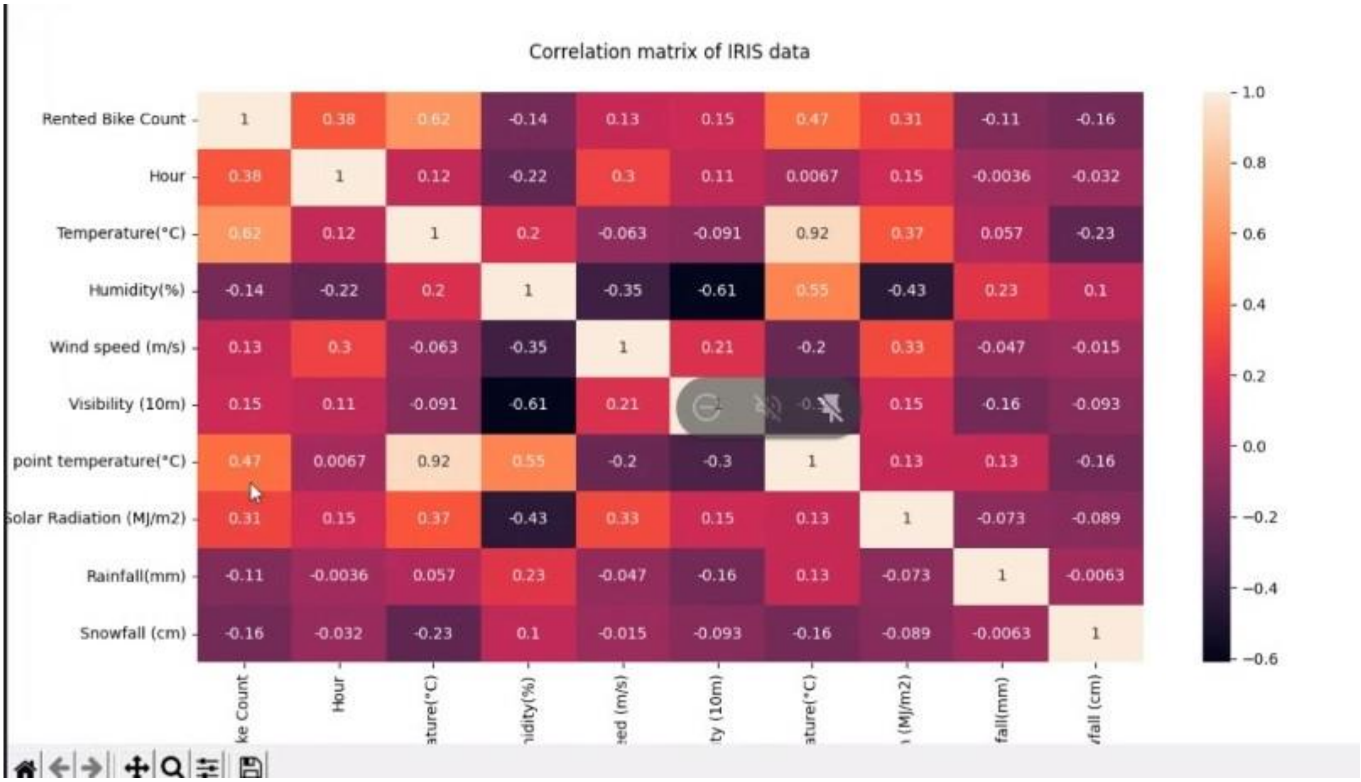
.

2- Data Analysis:

2.1- Columns description

Columns	Description
Date	Unique date
Hour	Describe 24 hours of the day from 0-23
Temperature(آ°C)	Describe day temperature
Humidity(%)	Describe if the degree of humidity is high or not ?
Wind speed (m/s)	Describe wind speed of the day
Visibility (10m)	Contain degrees of the day visibility
Dew point temperature(آ°C)	Describe if the degree of Dew point temperature is high or not ?
Solar Radiation (MJ/m2)	Describe solar radiation of the day
Rainfall(mm)	Describe the strength of the rainfall
Snowfall (cm)	Describe the strength of the snowfall
Seasons	Describe which season we are in
Holiday	Describe if a holiday or not?
Functioning Day	Describe if a Functioning Day or not?
Rented Bike Count	predicted column, contain how many bikes will be rented

2.1- Correlation matrix:



Correlation matrix of IRIS data

	ke Count	Hour	ature(°C)	idity(%)	ed (m/s)	ty (10m)	ature(°C)	(MJ/m2)	fall(mm)	fall (cm)
Rented Bike Count	1	0.38	0.62	-0.14	0.13	0.15	0.47	0.31	-0.11	-0.16
Hour	0.38	1	0.12	-0.22	0.3	0.11	0.0067	0.15	-0.0036	-0.032
Temperature(°C)	0.62	0.12	1	0.2	-0.063	-0.091	0.92	0.37	0.057	-0.23
Humidity(%)	-0.14	-0.22	0.2	1	-0.35	-0.61	0.55	-0.43	0.23	0.1
Wind speed (m/s)	0.13	0.3	-0.063	-0.35	1	0.21	-0.2	0.33	-0.047	-0.015
Visibility (10m)	0.15	0.11	-0.091	-0.61	0.21	1	-0.3	0.15	-0.16	-0.093
point temperature(°C)	0.47	0.0067	0.92	0.55	-0.2	-0.3	1	0.13	0.13	-0.16
olar Radiation (MJ/m2)	0.31	0.15	0.37	-0.43	0.33	0.15	0.13	1	-0.073	-0.089
Rainfall(mm)	-0.11	-0.0036	0.057	0.23	-0.047	-0.16	0.13	-0.073	1	-0.0063
Snowfall (cm)	-0.16	-0.032	-0.23	0.1	-0.015	-0.093	-0.16	-0.089	-0.0063	1

Correlation analysis:

1-rented bikes-> moderate positive correlation with temperature, hour, Dew point temperature

(It should be moderate negative correlation) and Solar Radiation.

-Weak positive correlation with wind speed and visibility.

-weak negative correlation with humidity, snowfall and rainfall.

2- Hour-> weak positive correlation with temperature, visibility, Dew point temperature and Solar Radiation.

-moderate positive correlation with wind speed.

-weak negative correlation with humidity, snowfall and rainfall.

3- temperature-> strong positive correlation with Dew point temperature.

-moderate positive correlation with Solar Radiation.

-weak positive correlation with humidity and rainfall.

-weak negative correlation with snowfall, wind speed and visibility.

4-humidity -> moderate positive correlation with Dew point temperature.

-weak positive correlation with snowfall and rainfall.

-moderate negative correlation with solar radiation, wind speed and visibility.

5-wind speed -> moderate positive correlation with solar radiation.

-weak positive correlation with visibility.

-weak negative correlation with Dew point temperature ,snowfall and rainfall.

6- visibility -> weak positive correlation with solar radiation.

-moderate negative correlation with Dew point temperature.

-weak negative correlation with snowfall and rainfall.

7- Dew point temperature -> weak positive correlation with solar radiation and rainfall.

-weak negative correlation with snowfall.

8-solar radiation -> weak negative correlation with snowfall and rainfall.

9-rainfall -> weak negative correlation with snowfall.

2.3-Effect of features:

- The higher the dew point temperature, the greater the amount of water vapor is present (source for clouds).

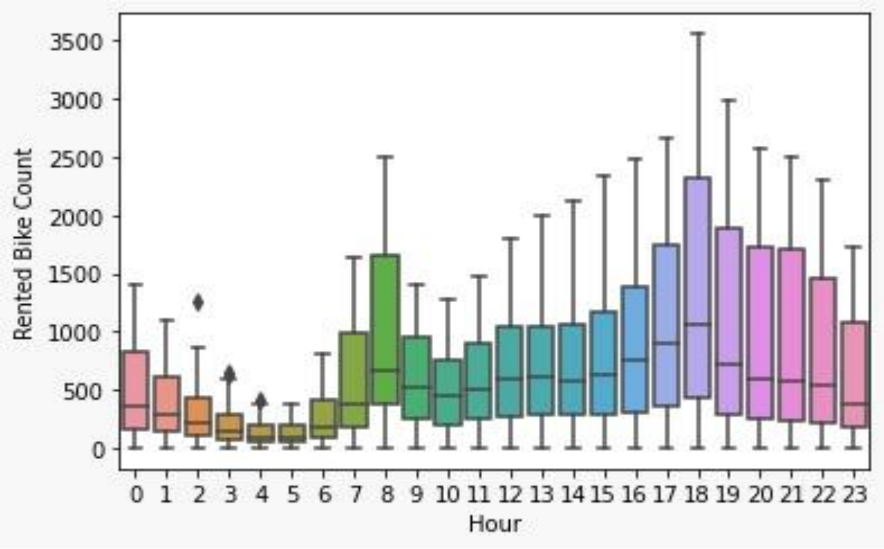
-The smaller the difference between the temperature and the dew point temperature, the higher the relative humidity (the closer the atmosphere is to a state in which water vapor would condense).

-It has been found that the solar radiation is directly proportional to the atmospheric temperature while it is inversely proportional to the relative humidity. It has also been found that wind speed has little influence on solar radiation

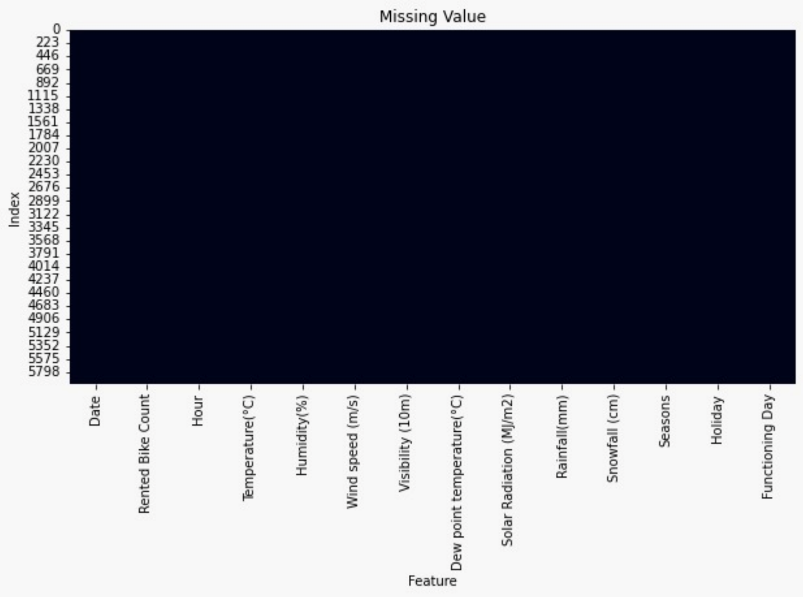
-When it rains, it will increase the relative humidity because of the evaporation. The air where the rain is falling may not be completely saturated with water vapor. However, the longer it rains, the more the humidity will increase because of the air constantly drawing the water.

-In relation to snow, when the humidity is high pressure is high and temperature is low snow is formed. When the humidity is high, pressure is constant and temperature is high the snow melts.

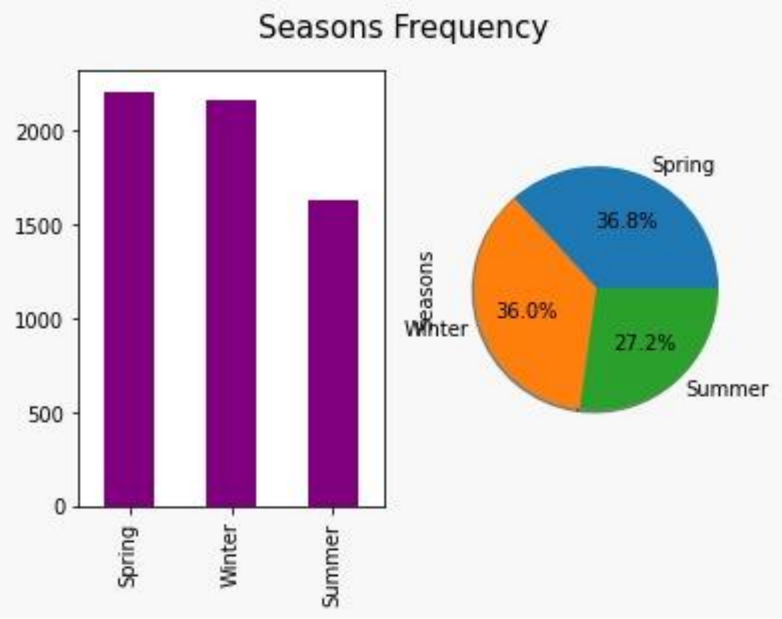
2.4- Graphs:



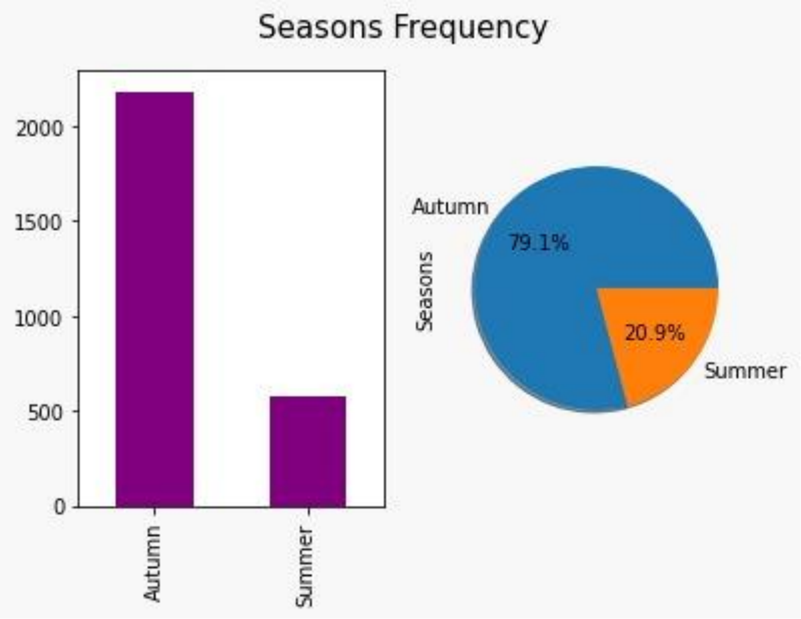
Box-blot between hour and rented bikes



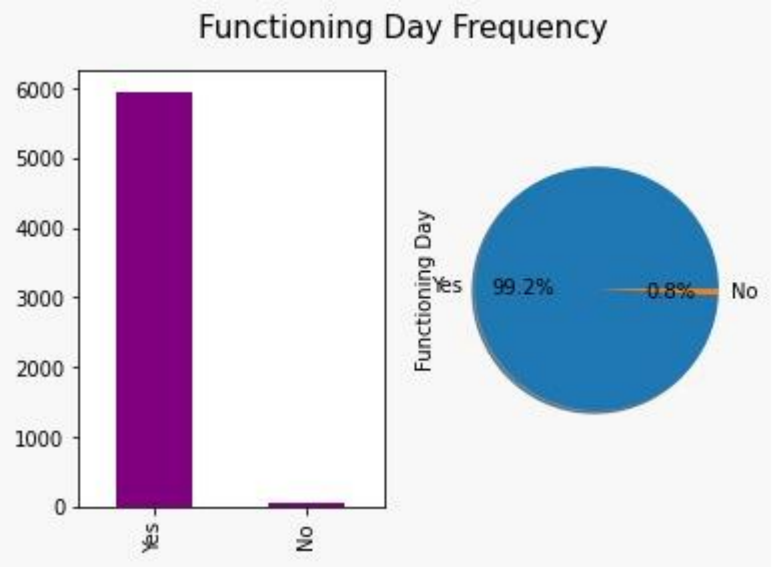
This graph shows if there is a missing value or not (no missing values)



Frequency of seasons in training dataset



Frequency of seasons in testing dataset



Frequency of functioning in training dataset

3- Pre-processing on Dataset

1- label encoding to date (day-month-year) then drop year

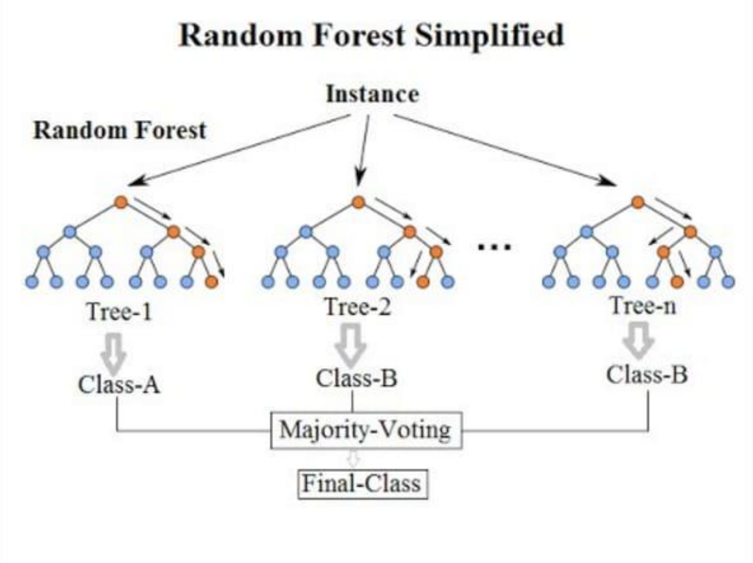
2-label encoding to seasons (spring-winter-summer-autumn)

3-hot encoding to holiday and functioning day

4-normalization OR standardization

4- Algorithms

1-



Name	Random forest
description	Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Bootstrap sampling is used in random forest ensemble algorithm called bootstrap aggregating (also called bagging). It helps in avoiding overfitting and improves the stability of machine learning algorithms. In bagging, a certain number of equally sized subsets of a dataset are extracted with random with replacement. Then, a machine learning algorithm is applied to each of these subsets and the outputs are ensembled as I have illustrated below:

hyperparameters	-hyperparameters-> 1-N_Estimators (num of trees) -N=640 2-max_feature (num of features in every tree) -Max features=13 3-max_depth (num of levels) -Max depth=21 4-min_sample_split (min number of samples in node) -min sample split=2 5--min_sample_leaf (min number of samples in leaf node) -min sample leaf=1
Mean score	46406.68701930383
error	0.89548

Name	Decision tree
description	The decision tree models can be applied to all those data which contains numerical features and categorical features. Decision trees are good at capturing non-linear interaction between the features and the target variable. Decision trees somewhat match human-level thinking so it's very intuitive to understand the data. It tends to overfit. A small change in the data tends to cause a big difference in the tree structure, which causes instability.
hyperparameters	Max_depth-> Number of Layer Max_depth=13
Mean score	93772.95143676517
error	0.78879

Name	Linear algorithm
Description	Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables Outliers affect this algorithm badly. It over-simplifies real-world problems by assuming a linear relationship among the variables, hence not recommended for practical use-cases.
Hyperparameters	No
Mean score	209744.743143571
error	0.53

Name	SVR algorithm
Description	Support Vector Regression uses the same principle as the SVMs to predict discrete values. the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and boundary line. Support Vector Regression depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target.
Hyperparameters	1-epsilon->determines the width of the tube around the estimated function (hyperplane). Points that fall inside this tube are considered as correct predictions and are not penalized by the algorithm. -Epsilon=0.5 2-kernel='linear' Specifies the kernel type to be used in the algorithm. If none is given, ‘rbf’ will be used. If a callable is given it is used to precompute the kernel matrix. 3- gamma'scale' (default) is passed then it uses 1 / (n_features * X.var()) as value of gamma,if ‘auto’, uses 1 / n_features. 4- C->default=1.0 Regularization parameter.

	-c=2000
Mean score	129585.57227207467
error	0.71

Name	Polynomial regression
Description	Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.
Hyperparameters	degree->A higher-degree polynomial allows the .fit() process more freedom to fit the data better. -Degree=4 include bies=true->create column with 1s to avoid the need to intercept -include bias=true
Hint	used cross validation with K-Fold =10
Mean score	91048254.5653148
error	-204.07136

Name	ridge regression
Description	Regularization terms can be used to reduce overfitting. The value of *w* cannot be too large or too small in the sample space. Linear least squares with l2 regularization. Minimizes the objective function:

	eq: $J(w)=\frac{1}{2m}\sum\left(h_w(x)-y\right)^2+\lambda\sum\lVert w\rVert_1$
Hyperparameters	1-Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. --default=1.0 2-copy_X, default=True If True, X will be copied; else, it may be overwritten. --copy_X=true 3-normalize, default=False the regressors X will be normalized before regression. If you wish to standardize, normalize=False. 4-fit_intercept: default=True fit the intercept for this model. 5-max_iter(int), default=None Maximum number of iterations for conjugate gradient solver. 6-tol(float), default=1e-3 ,we are made it tol=0.001 Precision of the solution. 7-solver:'auto' chooses the solver automatically based on the type of data. 8-random_state(int),we are made it random_state=0, default=None Used when solver == 'sag' or 'saga' to shuffle the data.
Mean score	209745.46235490753
error	0.52758

4.1- submitted Algorithms

-first submission

- Pre-processing ->split date, drop year, hot encoding to all categorical data, hot encoding manually to column season to include disappeared seasons in training data
- Algorithm (random forest):

Hyperparameters	random_state=0,
	max_features=13,
	min_samples_split=2,
	min_samples_leaf=1,
	n_estimators=640,
	max_depth=21

score	Mean score: 49656.78037919681 error = 0.88816

-second submission

- Pre-processing ->split date, drop year, rainfall and snowfall, hot encoding to all categorical data, hot encoding manually to column season to include disappeared seasons in training data
- Algorithm (random forest):

Hyperparameters	random_state=0,
	max_features=13,
	min_samples_split=2,
	min_samples_leaf=1,
	n_estimators=640,
	max_depth=21

score	Mean score: 46406.68701930383 error = 0.89548

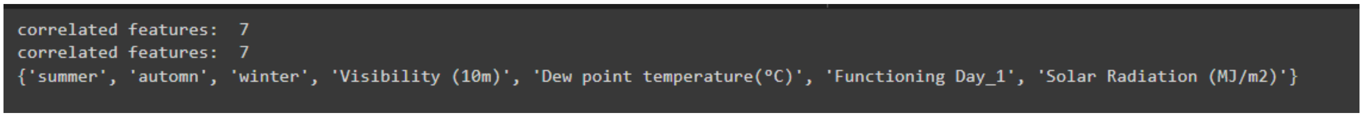
5- data divided into 80% training 20 % testing .

-used cross validation in polynomial regression with K-Fold =10

--used cross validation in ridge regression with K-Fold =100

6- further techniques that were used to improve the results:

-We used decision tree to extract correlated features



```
correlated features:  7
correlated features:  7
{'summer', 'automn', 'winter', 'Visibility (10m)', 'Dew point temperature(°C)', 'Functioning Day_1', 'Solar Radiation (MJ/m2)'}
```

-we used grid search to get the best hyper-parameters with SVR , Random forest and decision tree

7- conclusion:

ℹ We applied different models. The above output show that the best pre- processing are label encoding to date (day-month-year) and drop year from it ,label encoding to seasons (spring-winter-summer-autumn)and hot encoding to holiday and functioning day and when we dropped least correlated columns like :snowfall and rainfall, MSE is increased . All the models have slight over fitting but to consider and select model among all the above mentioned one based on MSE, we will go with random forest as it has relatively lesser over fitting and also lower MSE for both training and testing data. We could that in relative terms it could perform equally well if it’s applied on unseen data.