

Importer les packages

```
In [16]:

import pyforest
from sklearn import metrics
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV
# Hide warnings
import warnings
warnings.filterwarnings('ignore')
```

Importer les données

```
In [3]:

Data_Urbanization_Forest = pd.read_csv('data/concat_data/Data_Urbanization_Forest.csv')
```

```
In [4]:

Data_Urbanization_Forest.tail()
```

Out[4]:

	region	Date	Effectif de la population	Population rurale	Population urbaine	Taux d'urbanisation	Nombre de ménages ruraux	Nombre de ménages urbains	Taille moyenne des ménages ruraux	Taille moyenne des ménages urbains	Superficie Perdue
507	Ziguinchor	2014	565940.00	303344.00	262596.00	46.40	50814.08	40063.58	7.91	9.19	2988.99
508	Ziguinchor	2015	583528.00	309854.00	273674.00	46.90	50814.08	40063.58	7.91	9.19	1926.69
509	Ziguinchor	2016	601929.00	317216.00	284713.00	47.30	50814.08	40063.58	7.91	9.19	1988.32
510	Ziguinchor	2017	621168.00	324250.00	296918.00	47.80	50814.08	40063.58	7.91	9.19	3788.39
511	Ziguinchor	2018	641254.00	332170.00	309084.00	48.20	50814.08	40063.58	7.91	9.19	1706.27

Exploration des données

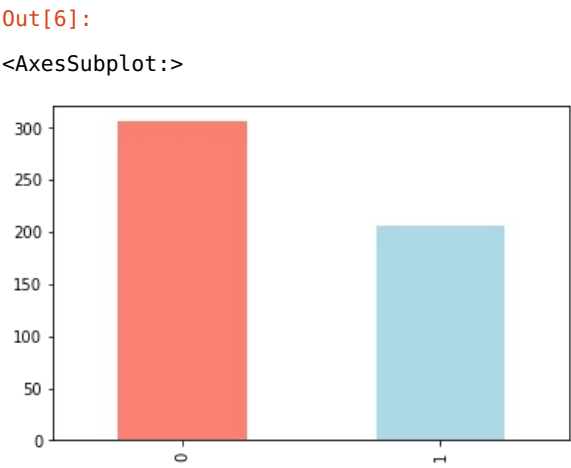
```
In [5]:

pd.set_option("display.float", "{:.2f}".format)
Data_Urbanization_Forest.describe()
```

Out[5]:

	Date	Effectif de la population	Population rurale	Population urbaine	Taux d'urbanisation	Nombre de ménages ruraux	Nombre de ménages urbains	Taille moyenne des ménages ruraux	Taille moyenne des ménages urbains	Superficie Perdue
count	512.00	512.00	512.00	512.00	512.00	512.00	512.00	512.00	512.00	512.00
mean	1994.80	786769.36	463799.61	322712.98	30.05	50814.08	40063.58	7.91	9.19	555.18
std	14.41	521011.50	234370.01	550205.13	23.56	17966.71	55553.61	0.17	0.73	1160.85
min	1970.00	122333.00	27614.74	20054.00	7.51	2785.35	3867.33	7.42	7.00	0.07
25%	1982.00	480522.11	338683.24	82088.50	15.54	49939.38	12643.48	7.91	8.80	1.69
50%	1995.00	642156.40	456375.94	137861.49	21.83	50814.08	25930.51	7.91	9.19	9.36
75%	2007.00	885796.16	573673.31	259445.00	35.69	56703.03	40063.58	7.91	9.19	744.32
max	2018.00	3630324.00	1463564.00	3499631.00	97.22	100729.97	321110.09	9.20	11.43	12172.03

```
In [6]:
Data_Urbanization_Forest['risque déforestation'].value_counts().plot(kind="bar", color=["salmon", "lightblue"])
```



```
In [7]:
Data_Urbanization_Forest.isna().sum()
```

Out[7]:

```
region                0
Date                  0
Effectif de la population  0
Population rurale       0
Population urbaine      0
Taux d'urbanisation     0
Nombre de ménages ruraux 0
Nombre de ménages urbains 0
Taille moyenne des ménages 0
Taille moyenne des ménages ruraux 0
Taille moyenne des ménages urbains 0
Superficie Perdue      0
risque déforestation    0
dtype: int64
```

Sélection des variables continues

```
In [8]:
Data_Urbanization_Forest = Data_Urbanization_Forest.drop(['region', 'Date'], axis= 1 )
Data_Urbanization_Forest.head()
```

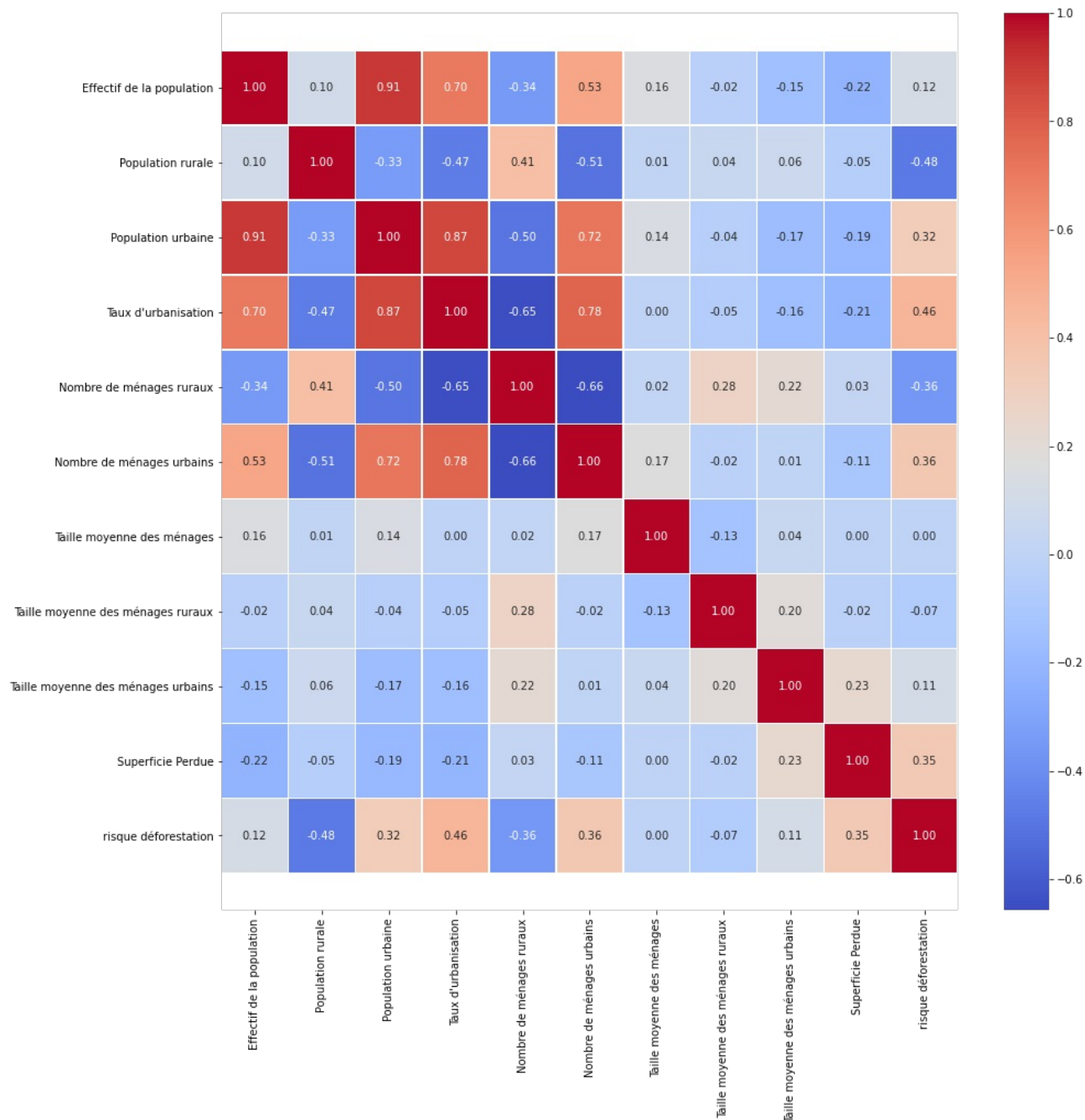
Out[8]:

	Effectif de la population	Population rurale	Population urbaine	Taux d'urbanisation	Nombre de ménages ruraux	Nombre de ménages urbains	Taille moyenne des ménages	Taille moyenne des ménages ruraux	Taille moyenne des ménages urbains	Superficie Perdue	risque déforestation
0	724461.69	27614.74	696846.95	96.19	2785.35	97871.76	7.55	9.91	7.12	4.44	1
1	759203.25	28805.87	730397.38	96.21	2928.70	101632.29	7.52	9.84	7.19	4.44	1
2	795610.84	30048.38	765562.46	96.22	3079.63	105546.30	7.48	9.76	7.25	4.44	1
3	833764.36	31344.49	802419.87	96.24	3238.55	109620.20	7.45	9.68	7.32	4.44	1
4	873747.53	32696.50	841051.03	96.26	3405.89	114397.58	7.42	9.60	7.35	4.44	1

Matrice de corrélation

```
In [9]:  
  
# Let's make our correlation matrix a little prettier  
corr_matrix = Data_Urbanization_Forest.corr()  
fig, ax = plt.subplots(figsize=(15, 15))  
ax = sns.heatmap(corr_matrix,  
                  annot=True,  
                  linewidths=0.5,  
                  fmt=".2f",  
                  cmap="coolwarm");  
  
bottom, top = ax.get_ylim()  
ax.set_ylim(bottom + 0.5, top - 0.5)
```

Out[9]:  
  
(11.5, -0.5)

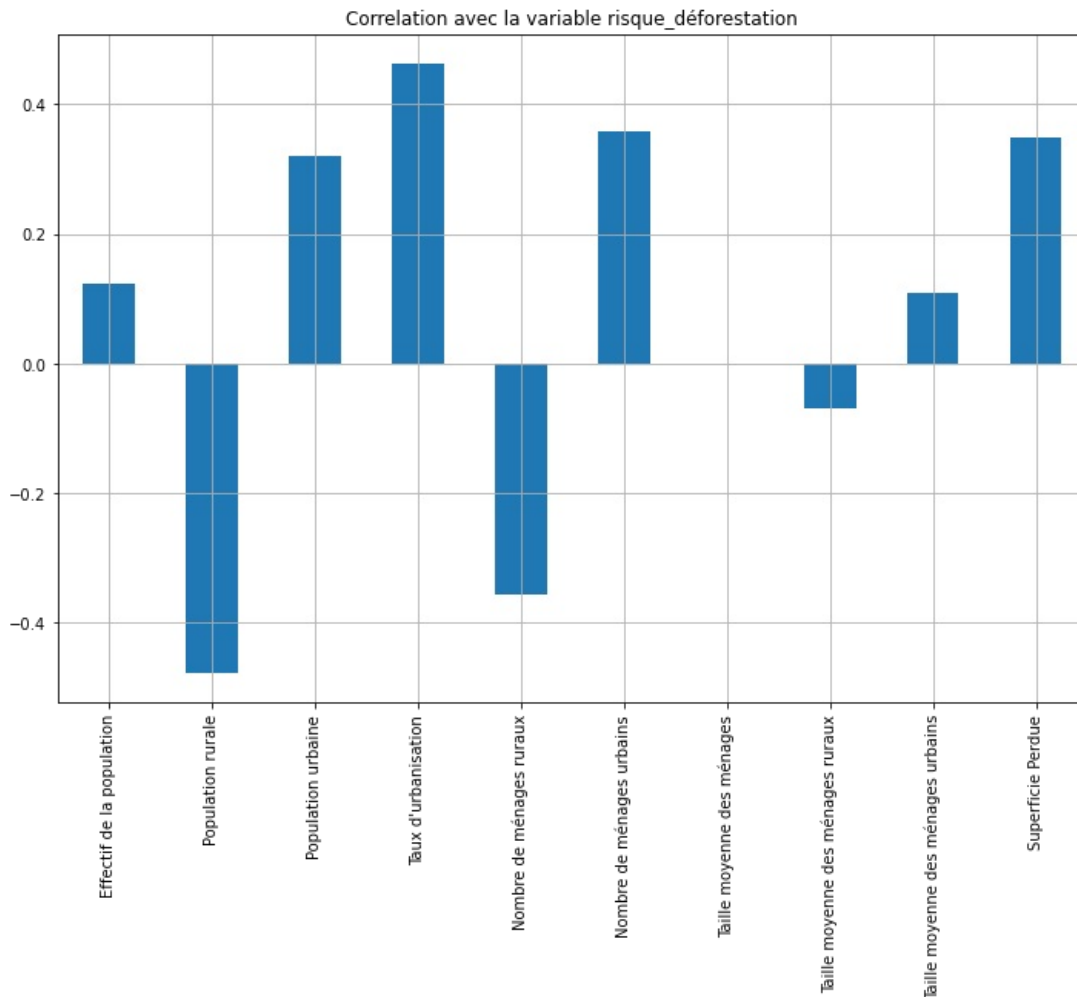


In [10]:

```
Data_Urbanization_Forest.drop('risque déforestation', axis=1).corrwith(Data_Urbanization_Forest['risque déforestation']).plot(kind='bar', grid=True, figsize=(12, 8), title="Correlation avec la variable risque_déforestation")
```

Out[10]:

```
<AxesSubplot:title={'center':'Correlation avec la variable risque_déforestation'}>
```



### Commentaires

- On voit les variables 'Population rurale', 'Population urbaine', 'Taux d'urbanisation', 'Nombre de ménages urbains' et 'Superficie Perdue' sont plus corrélées avec la variable 'risque\_déforestation' contrairement aux autres.
- Donc les variables citées ci-dessus sont considérées comme de bons prédicteurs.

### Sélection des variables les plus corrélées avec la variable cible

In [11]:

```
Selected_columns = ['Population urbaine',  
                    "Taux d'urbanisation", 'Nombre de ménages urbains', 'Superficie Perdue', 'risque déforestation']
```

In [12]:

```
Data_Urbanization_Forest = Data_Urbanization_Forest[Selected_columns]
```

### Transformation des données

In [13]:

```
X = Data_Urbanization_Forest.drop(['risque déforestation'], axis = 1 )  
y = Data_Urbanization_Forest['risque déforestation']
```

In [14]:

```
# from sklearn.preprocessing import StandardScaler
# s_sc = StandardScaler()
# X = s_sc.fit_transform(X)
```

## Fractionner les données

In [15]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

## Application de la méthode RandomForest sur les données

In [16]:

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

def print_score(clf, X_train, y_train, X_test, y_test, train=True):
    if train:
        pred = clf.predict(X_train)
        # clf_report = pd.DataFrame(classification_report(y_train, pred, output_dict=True))
        print("Train Result:\n=====")
        print(f"Accuracy Score: {accuracy_score(y_train, pred) * 100:.2f}%")
        # print(" ")
        # print(f"CLASSIFICATION REPORT:\n{clf_report}")
        print(" ")
        print(f"Confusion Matrix: \n {confusion_matrix(y_train, pred)}\n")

    elif train==False:
        pred = clf.predict(X_test)
        # clf_report = pd.DataFrame(classification_report(y_test, pred, output_dict=True))
        print("Test Result:\n=====")
        print(f"Accuracy Score: {accuracy_score(y_test, pred) * 100:.2f}%")
        # print(" ")
        # print(f"CLASSIFICATION REPORT:\n{clf_report}")
        print(" ")
        print(f"Confusion Matrix: \n {confusion_matrix(y_test, pred)}\n")
```

## Trouver les paramètres optimaux

In [17]:

```
rfc =RandomForestClassifier(random_state=42)
```

In [18]:

```
param_grid = {
    'n_estimators': [200, 500],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [4,5,6,7,8],
    'criterion' :['gini', 'entropy']
}
```

In [19]:

```
CV_rfc = GridSearchCV(estimator=rfc, param_grid=param_grid, cv= 5)
CV_rfc.fit(X_train, y_train)
```

Out[19]:

```
GridSearchCV(cv=5, error_score=nan,
             estimator=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,
                                              class_weight=None,
                                              criterion='gini', max_depth=None,
                                              max_features='auto',
                                              max_leaf_nodes=None,
                                              max_samples=None,
                                              min_impurity_decrease=0.0,
                                              min_impurity_split=None,
                                              min_samples_leaf=1,
                                              min_samples_split=2,
                                              min_weight_fraction_leaf=0.0,
                                              n_estimators=100, n_jobs=None,
                                              oob_score=False, random_state=42,
                                              verbose=0, warm_start=False),
             iid='deprecated', n_jobs=None,
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [4, 5, 6, 7, 8],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'n_estimators': [200, 500]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

In [25]:

```
CV_rfc.best_params_
```

Out[25]:

```
{'criterion': 'gini',
 'max_depth': 5,
 'max_features': 'auto',
 'n_estimators': 200}
```

#### **Application du modèle avec les paramètres optimaux obtenus**

In [26]:

```
rfc1=RandomForestClassifier(random_state=42, max_features='auto', n_estimators= 200,
                           max_depth=5, criterion='gini')
rfc1.fit(X_train, y_train)
```

Out[26]:

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=5, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=200,
                       n_jobs=None, oob_score=False, random_state=42, verbose=0,
                       warm_start=False)
```

In [27]:

```
print_score(rfc1, X_train, y_train, X_test, y_test, train=True)
print_score(rfc1, X_train, y_train, X_test, y_test, train=False)
```

Train Result:

```
=====
Accuracy Score: 99.44%
```

Confusion Matrix:

```
[[216  0]
 [ 2 140]]
```

Test Result:

```
=====
Accuracy Score: 93.51%
```

Confusion Matrix:

```
[[87  3]
 [ 7 57]]
```

## Prédiction

In [28]:

```
def Prédiction(Value):  
    if Value == [1]:  
        print('PREDICTION : \n La région est fortement menacée par la déforestation')  
    else:  
        print('PREDICTION : \n La région est faiblement menacée par la déforestation')
```

In [29]:

```
Value = rfcl.predict([[213312.93, 24.78, 4983.11, 8.87]])  
Prédiction(Value)
```

```
PREDICTION :  
La région est faiblement menacée par la déforestation
```

In [ ]: