LA POSTE DE 21 SIÈCLE

RAPPORT

PRÉPARÉ PAR

TOM FRIEDRICH MARCHAL
&
RODRIGUE ULUA LONGOMO
&
ALEKSANDRINA STOYANOVA-CHRISTEN

LA POSTE SE MODERNISE

LA DISTRIBUTION DU COURRIER EN FRANCE

STATISTIQUES

Chaque seconde La Poste française distribue 453 colis et courriers (lettres, plis, presse) en France, soit plus de 39 millions d'objets acheminés par jour et 14,3 milliards distribués chaque année en parcourant 1,4 milliard de kilomètres.*

l'acheminement du courrier reste le cœur de métier de La Poste. En 2017, 86,4 % des plis envoyés sous timbre rouge sont arrivés dans les délais.

LA POSTE DE 21 SIÈCLE

La Poste investit largement dans le développement des outils basés sur l'Intelligence Artificielle (IA) afin d'optimiser certaines tâches, comme, par exemple, la lecture d'une adresse sur une enveloppe. L'objectif principal est le développement d'une IA qui est capable de lire les caractères sur l'enveloppe, comme un humain.

La Poste a fait appel aux services de notre société afin de développer une application IA qui leur permettra de reconnaître automatiquement les codes postaux rédigés sur les enveloppes même si la graphologie est médiocre.

Le but est d'automatiser et donc accélérer la distribution du courrier en fonction des départements.

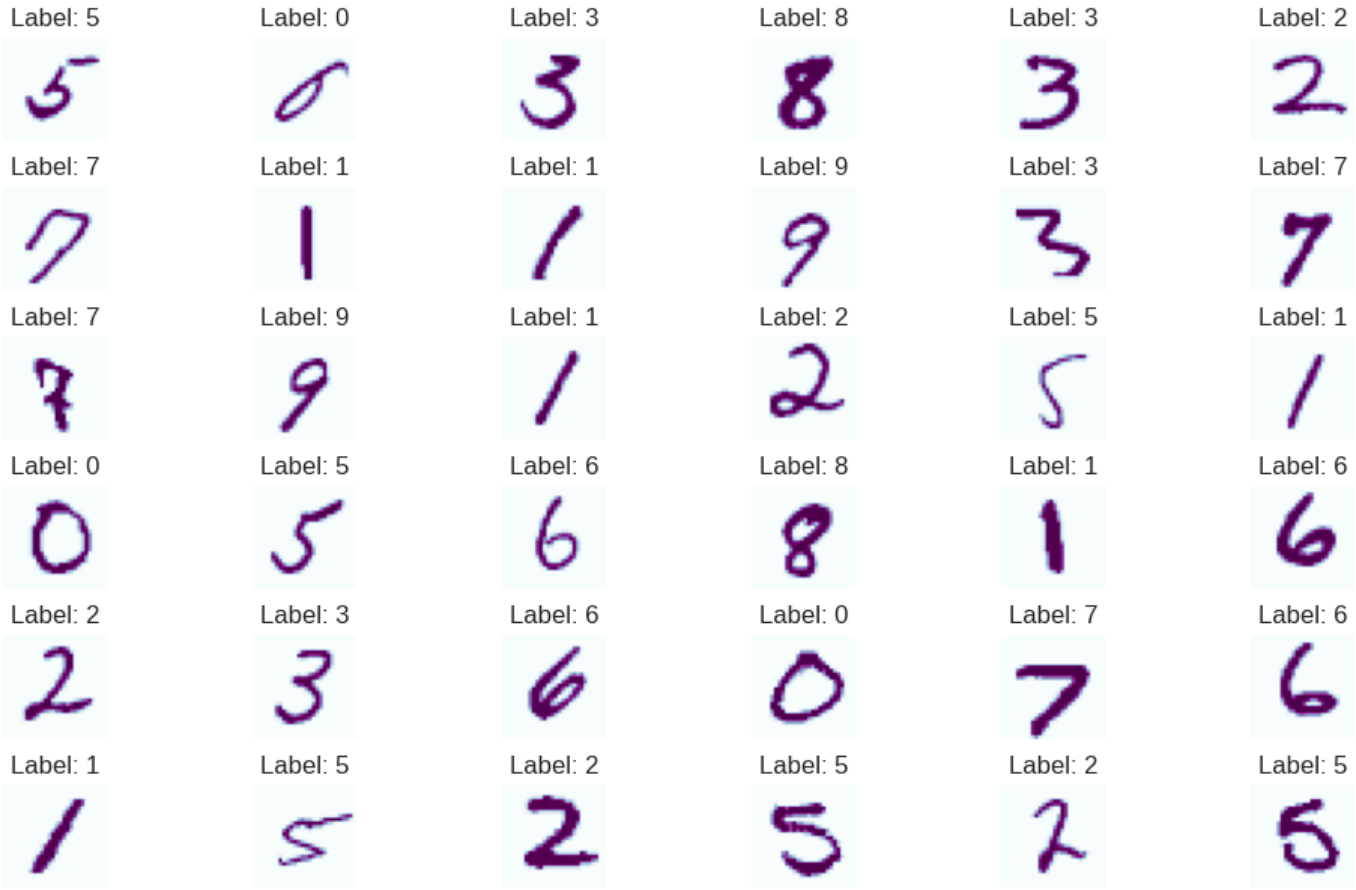
https://www.planetoscope.com/Mobilite/1318-nombre-de-lettres-et-colis-distribues-par-la-poste-en-france.html* - données 2014

Données MNIST

EXPLORATION DU JEU DE DONNÉES

MNIST ("Modified National Institute of Standards and Technology") est une base de données de chiffres écrits à la main. Le jeu de données MNIST comprend un lot de 42 000 images dédiées à l'apprentissage et de 28 000 images pour le test set, le dernier est utilisé afin de tester le modèle développé. Chaque image contenant un chiffre est de 28 x 28 pixels et chaque pixel représente une caractéristique, il y a donc 784 caractéristiques pour chaque image. La valeur d'un pixel est un entière (ici int64) entre 0 et 255. La variable 'label' (int64) donne la chiffre (de 0 à 9) décrite par les pixels.

Nous utilisons le jeu de données contenant 42 000 images afin de construire un modèle de Support Vector machine (SVM) qui doit être capable de reconnaître les chiffres manuscrits contenus dans les codes postaux. Plus précisément, nous utilisons un sous-échantillon de 30% de ces données en vu de faciliter l'optimisation des hyper-paramètres du modèle SVM. Nous avons vérifié que le sous-échantillon est statistiquement signifiant en utilisant le test de Kolmogorov-Smirnov. Le jeu de données ne contient pas des valeurs manquantes (NaN). Des chiffres sélectionnées d'une manière aléatoire sont montrées ci-dessous. La distribution des valeurs du 'label' montre un count pour la chiffre 1 de ~30% plus élevé que celui de la chiffre 5.

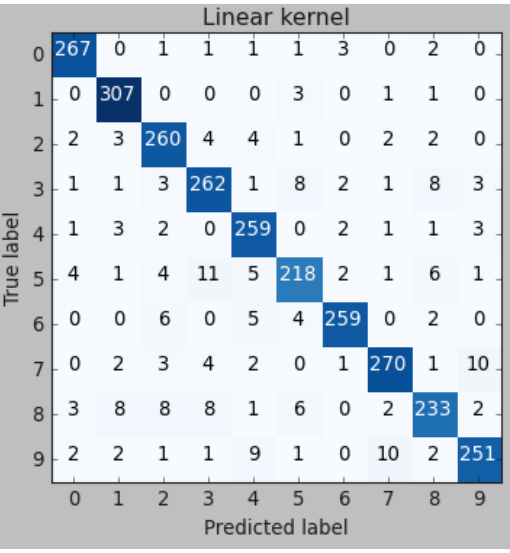


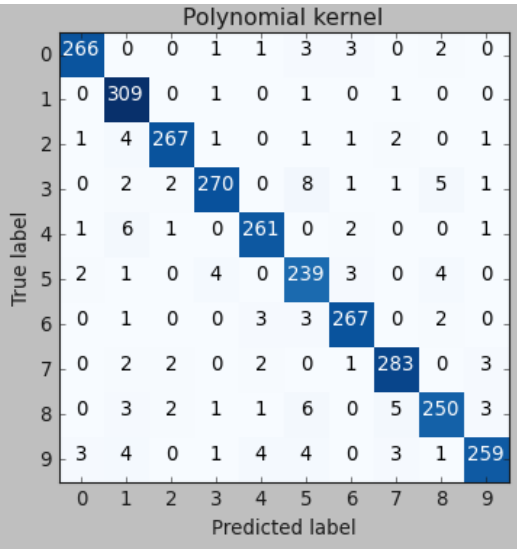
MODÈLE SVM

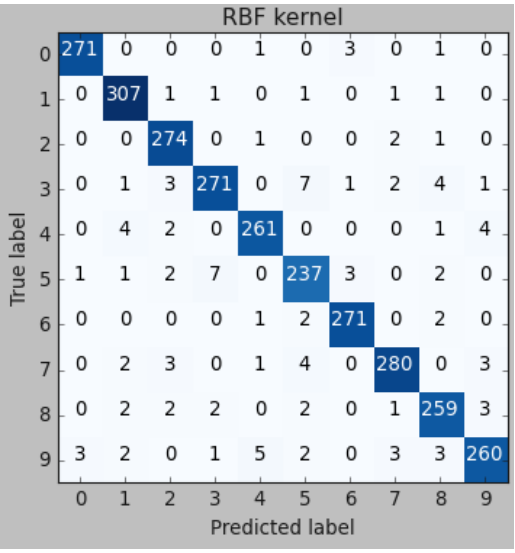
Les données sont séparées en training et testing sets (test_size=20 %) en préservant la distribution des labels dans les deux sets. Les valeurs des pixels varient entre 0 et 255. Nous appliquons une normalisation de type MinMaxScaler(). En inspectant la matrice de corrélation nous avons constaté la présence des nombreuses valeurs NaN due aux valeurs 0 des écarts types des variables-pixels. Une standardisation des variables est à éviter dans ce cas.

Nous avons construit trois modèles SVM de classification en utilisant trois noyaux RBF, linear et polynomial.

kernel	scores	accuracy
linear	0.995893	0.923571
poly	0.98625	0.953929
rbf	0.985268	0.961071







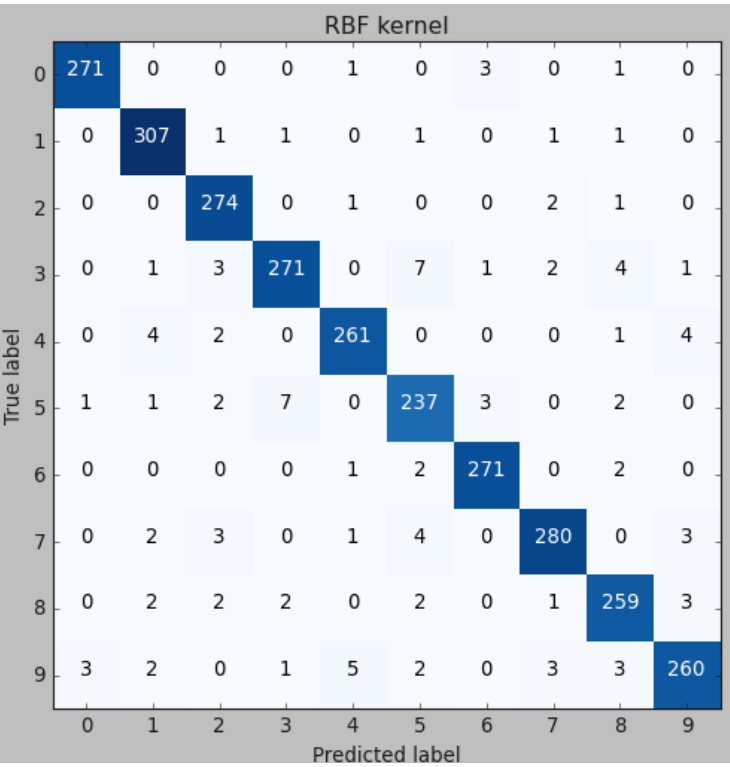
Les scores sur les trainings sets montrent une tendance de over-fitting pour le modèle avec le noyau linaire. Le modèle RBF semble être le meilleur (~96%) en ce qui concerne le score d'accuracy ainsi que les mesures de précision et de recall (voir notebook). Néanmoins, le modèle avec le noyau polynomial se distingue aussi avec un bon score d'accuracy.

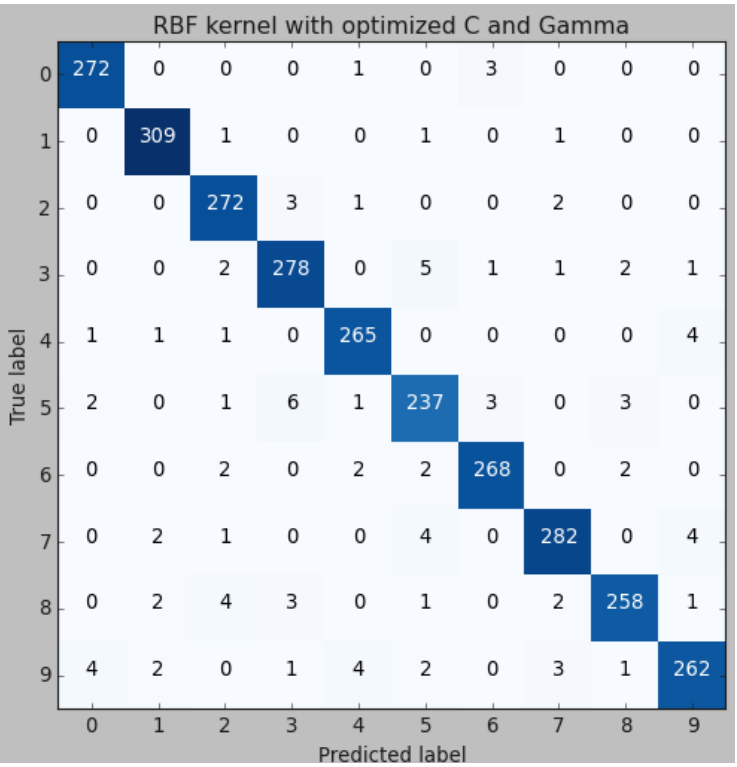
Résultats modèle optimisé. Prédictions.

Afin d'optimiser les hyper-paramètres du modèle RBF, nous avons utilisé GridSearchCV avec 10 cross-validations. Les paramètres optimaux qui étaient obtenus sont les suivants : C = 10, gamma = 0.01. Nous notons toutefois que avec ces valeurs pour C et gamma, le score sur le training set montre une tendence de over-fitting (score=0.99973). Une recherche des paramètres optimaux doit être effectuée dans un range des valeurs plus large.

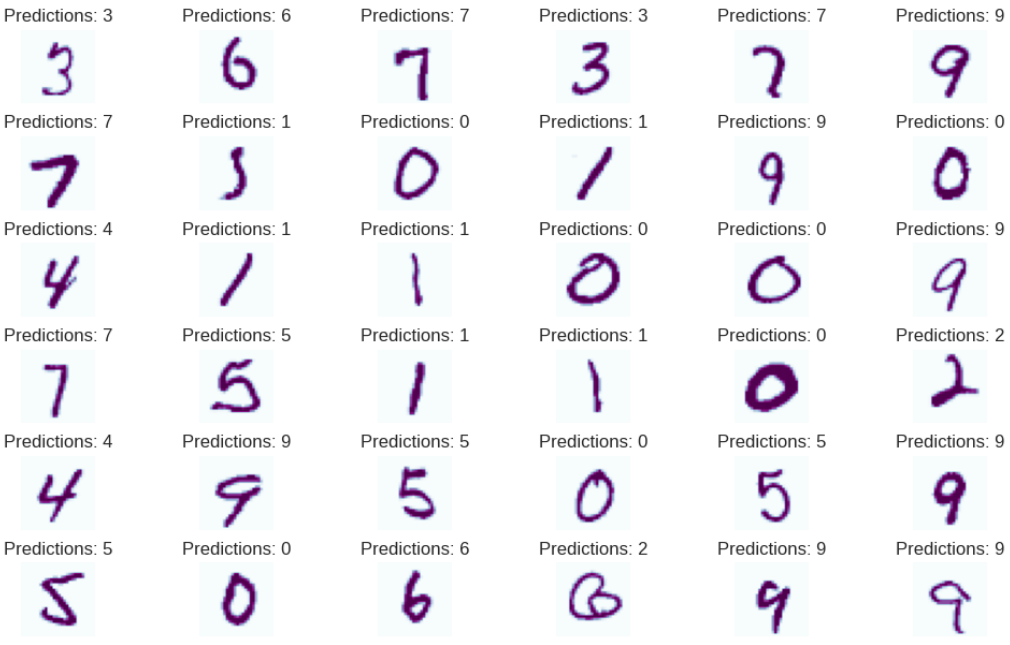
KERNEL	SCORES	ACCURACY
RBF	0.98526	0.96107
RBF OPTIM	0.99973	0.96535

MATRICES DE CONFUSION





PRÉDICTIONS



CONCLUSIONS

Suite à la demande de la Poste de développer une application IA qui sera capable de reconnaître les chiffres manuscrits dans les codes postaux, nous avons construit un modèle SVM de classification optimisé qui a une accuracy de 96%. Les scores du recall et de la précision varient entre 0.94 et 0.99 et 0.94 et 0.97, respectivement. La matrice de confusion montre la structure d'une matrice diagonalement dominante qui résonne avec la bonne valeur de l'accuracy et suppose une faible erreur dans la classification des chiffres.

Nous proposons ce modèle à La Poste afin de leur aider d'automatiser et d'améliorer la distribution des colis et des courriers.

PERSPECTIVES

Tenons compte du grand nombre de dimensions (784 pixels) dans lesquelles évoluent nos 42000 observations (images), nous avons entamé des analyses de Principal Component Analyse afin d'étudier la possibilité à réduire notre espace initiale de variables explicatives. Nous avons trouvé que plus que 85% de la variance dans le jeu de données est expliquée par les premières 50 composantes principalles.