در ابتدا کتابخانه‌های مورد نظر را در برنامه خود وارد می‌کنیم.

فایل دیتاست را از گوگل دریافت کرده و در کنار فایل jupyter notebook خود قرار می‌دهیم و با دستور pd.read_csv("./bank.csv", delimiter=";") آن را می‌خوانیم (قرار گیری در RAM). پارامتر delimiter نشان دهنده نوع جداساز ما است که اغلب موارد کاما (,) است اما در اینجا (;) است.

قبل از انجام کاری، لازم است به سایت گفته رجوع کنیم و اطلاعات بیشتری راجب دیتاست کسب کنیم مانند تعداد سطرها/نمونه‌ها و تعداد ستون‌ها/ویژگی‌ها/بعدها و همچنین لازم است بدانیم ستون‌ها از چه نوع ساختاری هستند و مقادیر آنان چیست.

با استفاده از توابع داخلی (builtin function) کتابخانه Pandas می‌توانیم کارهای آماری روی دیتاست خود انجام دهیم نظیر تعداد نمونه‌ها، واریانس داده‌های عددی، کوچک‌ترین و بزرگترین مقدار داده‌ی عددی و ... تا بیشتر راجب داده‌های خود اطلاعات داشته باشیم.

یکی از موارد مهم در مرحله پیش‌پردازش داده‌ها (preprocessing)، حذف داده‌های null است چون الگوریتم یادگیری ماشین قادر به حل این مشکل نیست.
با دستور df.isnull().sum بررسی می‌کنیم که هر ستون چه مقدار داده‌ی null دارد که در این دیتاست، داده‌ی null نداریم.

بعد از این کار لازم است داده‌های طبقه‌بندی (categorical) را به داده‌های عددی (numeric) تبدیل کنیم. چون الگوریتم‌های یادگیری ماشین بر اساس داده‌های عددی یادگیری انجام می‌دهند. اما خوب است بدانیم لازم نیست داده‌های ستونی را که مقادیر آن True یا False است را به عددی تبدیل کنیم چرا که در پایتون True=1 و False=0 تعریف می‌شود.

یکی دیگر از روش‌های پیش‌پردازش داده‌های که باعث فهم بهتر الگوریتم می‌شود، Z-score ،min-max scaler و ... است که در اینجا انجام داده نشده است.

بعد از این مراحل باید دیتاست خود را به بردار ویژگی X و بردار هدف y قسمت‌بندی کنیم و بعد از آن هم دو مجموعه آموزش(train) و آزمایش(test) درست کنیم تا با مجموعه آموزش مدل را آموزش دهیم و با مجموعه آزمایش، مدل را تست و ارزیابی کنیم.

بعد از این مراحل پیش‌پردازش وارد فاز جدید می‌شویم: آموزش مدل

یک نمونه از آن الگوریتمی که قرار است یاد بگیرد، ایجاد کرده و در صورت لزوم ابرپارامترها (hyperparameters) را به آن می‌دهیم. و بعد از آن با دادن مجموعه آموزش به الگوریتم، مدل آن را یاد می‌گیرد.
در اینجا از درخت تصمیم (Decision Tree) استفاده می‌کنیم که زیر مجموعه‌ای از مسائل طبقه‌بندی (classification) است.

بعد از یادگیری مدل، باید مدل را سنجید به همین دلیل ما مجموعه بردارهای تست X را به مدل می‌دهیم تا به‌ما بردار هدف y را بدهد که ما آن را y_pred می‌نامیم.
اکنون باید مقایسه‌ای بین y_pred با مقادیر واقعی بردار هدف y_true انجام دهیم و عملکرد مدل را معیارهای به سنجیم.
چندین معیار ارزیابی عملکرد مدل‌های طبقه‌بندی وجود دارد به نام‌های accuracy و recall و precision و f1-score و ... که پشت تمام این‌ها روابط ریاضی است که تا حد خوبی در کلاس بررسی و یادگرفته شده است و مهم‌تر از روابط لازم است درک کنیم که هر کدام چه چیزی را بیان می‌کند.

در کتابخانه Sklearn، جمعی از این معیارهای ارزیابی وجود دارد :classification_report که با دادن y_true و y_pred به عنوان آرگومان به تابع، برایمان معیارهارا بدست می‌آورد.

هنگامی داریم نمونه‌ای از الگوریتم Decision Tree را ایجاد می‌کنیم یکی از ابرپارامترهای آن max_depth است که مشخص می‌کند عمق درخت تصمیم ما چقدر باشد و یکی از روش‌های جلوگیری از بیش‌برازش (overfit) است.