










<h2>機器學習任務</h2>  <p>預測的輸入和輸出是什麼？ 機器學習任務的類型是什麼？ 可選的演算法模型是什麼？</p> <p>輸入：缺席的事由和時間，缺席員工的相關資料。</p> <p>輸出：缺席的時間長短。</p> <p>Clustering</p>	<h2>決策行動</h2>  <p>模型預測如何變成決策行動？</p> <p>當有新的缺席紀錄產生，根據預測結果做出以下幾種決定之一：</p> <ul style="list-style-type: none"> 把當日到期的工作分配給其他人。 把三個工作日內到期的工作分配給其他人。 把一週內到期的工作分配給其他人。 	<h2>問題定義</h2>  <p>預測系統會為終端使用者帶來什麼價值？我們選擇什麼指標來解決問題？</p> <p>為一個公司來說，員工的缺席時間越長，影響產品的進度越大，此系統會根據員工的缺席的原因，預測員工的時間長短，讓管理者能夠重新分配工作，降低因為員工缺席所帶來的影響。</p>	<h2>數據來源</h2>  <p>我們可以使用哪些原始數據？</p> <p>資料集：Absenteeism at work Data Set。 參考做法： https://www.researchgate.net/publication/349868760_Prediction_of_Absenteeism_at_Work_using_Data_Mining_Techniques</p>	<h2>數據輸入輸出</h2>  <p>來源的哪些資料作為訓練？哪些作為測試？</p> <p>從 Absenteeism at work Data Set 中，660 筆用於 training，80 筆用於 test。</p>
<h2>線上預測</h2>  <p>我們什麼時候會對輸入做出預測？多久做一次預測？</p> <p>當有新的缺席紀錄產生，對此缺席紀錄做出新的預測。</p>	<h2>離線評估</h2>  <p>部署之前，用什麼方法和指標來評估預測系統？</p> <p>依據近一個月內的缺席紀錄，判斷預測的時間長短，和實際的資料比較正確率評估預測系統。</p>		<h2>特徵工程</h2>  <p>從原始數據中提取什麼特徵(feature)？如何處理這些特徵？</p> <p>無關的 ID、日期、教育程度、孩子和寵物數量排除，選取其他資料。</p> <p>對沒有固定選項且分布範圍廣的資料做 Normalize 和 Flitering Missing Value。</p>	<h2>建立模型</h2>  <p>如何建立模型？用什麼演算法來訓練？</p> <p>將資料分成 testing 和 training 資料，利用 training 資料訓練模型。</p> <p>演算法則使用 Agglomerative clustering，參數: (3 clusters，linkage= complete，affinity=euclidan)。</p>
	<h2>即時評估和監測</h2> <p>部署之後，用什麼方法和指標來評估預測系統？</p> <p>如何量化它帶來的價值？</p>		