1.1: I klassificeringsproblem försöker man lära sin modell att kunna dela upp datan i kategorier. Exempelvis kan man ha data som är mätningar av olika blodvärden, och man vill med hjälp av dessa kunna prediktera om ett nytt blodprov kommer från en patient med en viss sjukdom. Detta är ett binärt klassificeringsproblem; antingen har patienten sjukdomen, eller så har hen inte det. Klassificeringsproblem kan också ha fler kategorier än 2.

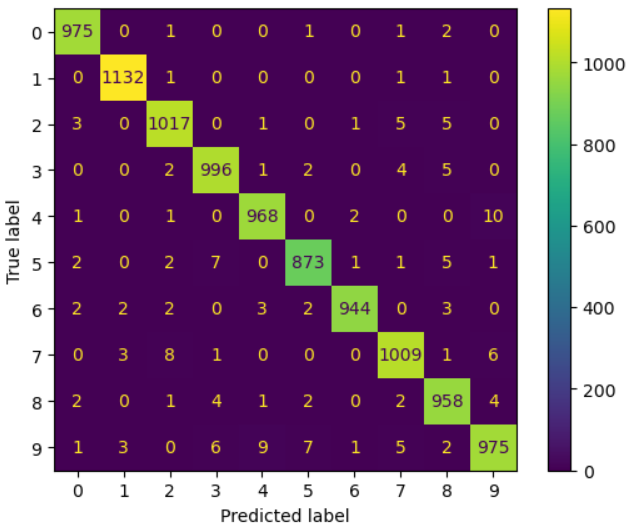
I ett regressionsproblem å andra sidan, så ber man modellen att uppskatta en kontinuerligt värde. Man kan säga att man ber modellen titta på datan, och sedan peka ut var på ett spektrum denna datapunkt borde befinna sig. Ett exempel är att titta på en persons utbildning, erfarenhet, hemort, ålder, etc. för att försöka avgöra vad den personen borde ha i lön.

1.2: I **supervised learning** har man ett facit som modellen kan använda sig av. Vill man till exempel använda sig av modellen för att förutse lön som nämndes ovan, så ger man modellen all data, plus hur hög lön personerna har. Då lär sig modellen vad som leder fram till vilken lön, och kan applicera det på nya fall, där man inte anger lönen. Vanliga modeller är Random Forrest, Support Vector Machines, K-Nearest Neighbors och Logistic Regression. Supervised learning är praktiskt när man redan har stora mängder data, och vill lära sig förutse ny data av samma sort. Exempelvis kan man tänka sig att ett försäkringsbolag vill bygga statistiska modeller rörande bilolyckor för sin bilförsäkring. Då bilolyckor sker ganska kontinuerligt så kan man räkna med att likartad data kommer fortsätta strömma in, och bolaget kan då ta data som går flera år tillbaka och träna en modell för att försöka få fram riskgrupper och hitta trender. När sedan en ny kund vill teckna ett avtal kan modellen använda allt den lärt sig för att förutse hur riskabel den nya kunden är, och därmed vad försäkringen borde kosta.

**RMSE** står för root mean square error, och beskriver hur långt från snittet som den genomsnittliga datapunkten befinner sig. Ju lägre tal, desto närmare är datapunkterna snittet, och desto mer precis kan man räkna med att modellen är. Detta eftersom när man ber en regressionsmodell att göra en uppskattning från viss data, så kommer den ge tillbaka ett värde som ligger på snittet. Om datapunkterna befinner sig långt ifrån snittet (om variansen är stor), så kommer de predikterade värdena (som ligger på snittlinjen) att visa stor variation jämfört med träningsdatan, vilken gör modellen mindre precis.

En **confusion matrix** är ett sätt att visualisera resultatet från klassificeringar. Den simplaste varianten är om man har ett binärt klassificeringsproblem. Man kan då göra ett “rutnät” som representerar de olika valmöjligheterna. På y-axeln har man de faktiska värdena, och på x-axeln har man de predikterade värdena. Eftersom det är ett binärt problem, så kommer antalet rutor bli 4, eftersom det finns två faktiska värden, och två möjliga prediktioner för varje. I varje ruta kommer antalet prediktioner för just den kombinationen stå. Exempelvis så kan det faktiska värdet vara Sant, och om prediktionen också var Sant, så kommer det resultatet hamna i rutan för där både svaret och prediktionen var Sant. Om 345 observationer var i kategorin Sant, och klassades som sanna av modellen, kommer motsvarande ruta innehålla siffran 345. Om 17 av observationerna som var sanna, klassades som falska av modellen, så kommer rutan för detta (Sant, klassat som falskt) innehålla siffran 17. En bra modell kommer ha höga tal i rutorna för kombinationer där modellen predikterat korrekt, och låga tal i rutorna för kombinationer där modellen predikterat fel.

Så här kan en confusion matrix se ut:



Om vi börjar längst upp till vänster så ser vi att av de observationer som hade “true label” 0, så predikterade modellen 975 st rätt. Men fortsätter vi lodrätt åt höger kan vi se att en nolla klassades som en tvåa av modellen. Går vi nedåt istället kan vi se att 3 st tvåor klassades som nollor av modellen. Den största svårigheten modellen verkar ha är att skilja på fyror och nior; Den har klassat 9 st nior som fyror, och 10 st fyror som nior.