### *Av Apti Dzhamurzaev* Linär regression inom AI-projekt för att prediktera priser på hus.

### För att kunna skapa effektiv modell för att kunna förutse huspriser krävs det omfattande och relevant datainsamling av information som är viktigt såsom antalet rum, lokalisation, storlek och ålder på huset. Denna data kan samlas från offentliga databaser, fastighetsregister och fatagihetsformedlares webbsidor sedan kan information lagras lokalt genom att använda databashanteringssystem såsom MySQL eller i molnbaserade tjänster för att senare ha användning i modellering. När man har samlat in data då ska man göra den redo för analys vilket sker genom att man hanterar bristfälliga värden, normaliserar numeriska variabler och konverterar kategoriska data till numerisk form. Detta sker effektivt via Pandas och NumPy i Python programmet. Sedan man säkerställer att modellen har de bästa möjligheter att prestera så bra som möjligt genom korrekt databehandling. Man vet att viktig del av processen är att undersöka och visualisera datta. Detta leder till att man kan hitta mönster och se sambandet i datasetet som är inte uppenbar endast via siffror och tabeller. Python bibliotekets Matplotlib är användbar för att skapa grafiska representationer av variabler och dess koppling till målet i detta fall är huspriser. Genom att använda visualiseringsmetoder såsom histogram och scatter plots kan man hitta tender och relationer som kan vara anvgörande hur bra själva modellen presterar. Tränning med modellen är det optimalt att använda metoden linjär regression som gör det möjligt att förutsäga kontinuerliga variaber och i detta fall passer tekniken att förutse huspriser. Linär regression kan användas med mål att hitta bästa anpassade linjen som visar sambandet mellan en eller flera oberoende variabel till exempel det kan vara antal rum, storleken på rum, storlek på själva huset och den beroende variabel i detta fall som är huspriset. Man kan genom att förminska skillnader mellan de observerande värden exempelvis de faktiska huspriserna i datasetet och förutsagda värden vilket genereras av själva modellen, justeras linjen mellan de vilket leder till att det passar datamängden så bättre och det resulterar i mer exakt förutsägelser.

### För att uppnå detta måste man dela upp datan i tränings- och testuppsättningar. Man använder tränningsdata för att bygga upp modellen medan testadata man använder för att utvärdera modellens prestanda. Detta är tillvägagångsätt som säkerställer att modellen kan generalisera till nya data istället endast bara prestera bra på den data den har tränats på. När man har färdig tränat modellen är det viktigt att utvärdera modellens prestanda och det görs via Mean Squared Error alltså det mäter genomsnittet av de kvadrerade skillnaderna mellan de förutsagda och de faktiska värderna. Ju lägre detta värde är, desto bättre är modellen på att förutsäga huspriser. Om det behövs så kan modellen justeras ytterligare för att få bättre prestande vilket exempelvis sker genom att man justerar parametrar genom att återgå till förra steget. Efter att modellen har slutförts och utvärdats är det möjlighet att implementera den för användning i praktiken genom att använda ramverk som Flask för att skapa webbapplikationer online då blir själva modellen tillgänglig online. Med hjälp av molnteknik för drift och dristibution kan man använda leverantörer exempelvis Google Cloud att kunna köra systemet och möta behovet av skalbarhet alltså att systemet kan hantera det växande mängden av data och större användarbas utan att man påverkar dess prestanda negativt. Genom man följer noggrann plan från insamling och förberedelse av data till träning och utvärdering av modellen så har linjär regression modell förmåga att skapa en framgångsrikt prediktering av huspriser. För att detta ska funka så är varje steg i processen avgöran för att få modellens framgång alltså säkerställa att datan är korrekt behandlad till implementera modellen på ett skalbart sätt. Man ska utföra arbetet systematisk då kan man skapa en modell som leverar pålitliga och goda prognoser som kommer göra den användbar i framtiden.

### Referenser

1. Pant, A. (2019, januari 11). **Workflow of a Machine Learning Project**. Towards Data Science. Hämtad från: <https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94>
2. Mayo, M. (2018, december 7). **The Machine Learning Project Checklist**. KDnuggets. Hämtad från: <https://www.kdnuggets.com/2018/12/machine-learning-project-checklist.html>
3. Linjär regression. (2024, september 8). *Miljöstatistik* . Hämtad från: <https://www.miljostatistik.se/lregression.html>
4. Python - Introduction to Matplotlib. (2024, augusti 21). *GeeksforGeeks*. Hämtad från: <https://www.geeksforgeeks.org/python-introduction-matplotlib/>
5. Matplotlib tutorial. (2024, september 5). *GeeksforGeeks*. Hämtad från: <https://www.geeksforgeeks.org/matplotlib-tutorial/>
6. NumPy tutorial*.* (2024, augusti 14). *GeeksforGeeks*. Hämtad från: <https://www.geeksforgeeks.org/numpy-tutorial/?ref=outindfooter>
7. Pandas tutorial*.* (2023, september 8). *GeeksforGeeks*. Hämtad från: <https://www.geeksforgeeks.org/pandas-tutorial/?ref=outindfooter>
8. Python – Mean Squared Error*.* (2019, juni 30). *GeeksforGeeks*. Hämtad från: <https://www.geeksforgeeks.org/python-mean-squared-error/>
9. Lewenhagen, K. (2017, november 10). Kom igång med Flask*.* *Dbwebb*. Hämtad från: <https://dbwebb.se/kunskap/kom-igang-med-flask>
10. Roos, M. (2018, mars 20). Kom igång med databasen MySQL - MariaDB och dess klienter*.* *Dbwebb*. Hämtad från: <https://dbwebb.se/kunskap/kom-igang-med-databasen-mysql-och-dess-klienter>
11. Knox, K. (2024, may 2). What is Google Cloud Platform (GCP)? Pluralsight. Hämtad från: <https://www.pluralsight.com/resources/blog/cloud/what-is-google-cloud-platform-gcp>